Towards Open-Vocabulary ALS Point Clouds Semantic Segmentation: An Empirical Study

Yanghong Lin[1,2,3], Tianyu Li[1,2,4], Shudong Zhou[1,2], Jingru Zhang[1,2], Li Fang[1,2], Wei Yao[1,2]

[1] Spatial Intelligence and Urban Computing, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China - lfang@iue.ac.cn, wyao@iue.ac.cn
[2] State Key Lab for Ecological Security of Regions and Cities, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China [3] University of Chinese Academy of Sciences, Beijing, China - linyanghong21@mails.ucas.ac.cn
[4] School of Resource and Environmental Sciences, Whuhan University, Whuhan, China

Keywords: Visual Foundation Model, ALS Point Cloud, Semantic Segmentation, Open-vocabulary, Zero-shot.

Abstract

While deep learning has advanced ALS point cloud semantic segmentation and achieved impressive results, most methods rely on predefined label sets and lack ability to recognize arbitrary categories. Recently, the visual foundation models (VFMs) has garnered significant attention, due to remarkable zero-shot generalization capabilities by leveraging open-set knowledge. However, adapting these models to large-scale ALS point clouds remains largely unexplored and highly challenging. In addition, the frequent absence of well-aligned synchronously acquired images further hinders the application of 2D VFMs in ALS point clouds. To bridge these gaps, we developed a zero-shot, open-vocabulary semantic segmentation framework for ALS point clouds based on 2D-3D transfer, utilizing three types of VFMs. We employed a combination of VFMs, including source models pre-trained on natural imagery and models fine-tuned on remote sensing data, to investigate the generalization capabilities of VFMs in inherent domain gap between natural and aerial imagery. Besides, we further introduce an adaptive global view projection module that derives optimal virtual camera poses and field-of-view (FOV) from scene extents, effectively enabling the application of 2D VFMs even in the absence of original imagery. Quantitative evaluations on the Vaihingen dataset indicate that methods trained solely on natural images achieve segmentation accuracy scores of 72% (roof) and 59% (tree) for common classes but struggle with rare categories such as powerline. GSNET improves performance across most categories, highlighting importance of domain adaptation. Evaluation on the SUM dataset reveals that our approach effectively identifies large-scale urban elements (exceeding 60% precision for buildings) without high-quality, well-aligned imagery.

1. Introduction

Understanding 3D scenes has recently gained importance due to its pivotal role in various applications, including forest survey (Polewski et al., 2015), autonomous driving (Jia et al., 2022) and urban planning (Huang et al., 2020, Shao et al., 2025). In photogrammetry and remote sensing, the semantic segmentation of airborne laser scanning (ALS) point clouds is a critical task that provides an accurate semantic prediction to facilitate various real-world applications (Guo et al., 2023).

Recently, supervised deep learning has evolved as mainstream methods for semantic segmentation of ALS point clouds. However, a significant and persistent bottleneck still exists in achieving real-world deployment robustness. The primary limitation of these supervised models is their innate dependency on massive and extensively annotated datasets (Wang and Yao, 2022, Wang et al., 2025b). For large-scale outdoor environments, the annotation procedure is exceptionally costly and labor-intensive due to the vast data volume, irregular point density, and the requirement to distinguish fine-grained object categories. The excessive cost of data annotation limits the applicability of these methods. In addition, most existing segmentation methods cannot handle arbitrary semantic categories, as they are trained on predefined label sets (Ye et al., 2025). This leads to a severe generalization problem when encountering novel scenes or previously unseen objects in dynamic, open-world settings.

In parallel with the advancement in remote sensing, the development of visual foundation models (VFMs) has garnered significant attention, such as CLIP (Radford et al., 2021), Grounding DINO (Liu et al., 2024), SAM (Kirillov et al., 2023). Pre-trained on large-scale 2D natural image–text datasets, these models have demonstrated remarkable zero-shot generalization capabilities by leveraging comprehensive, open-set knowledge. However, adapting these models to large-scale ALS point clouds remains largely unexplored and highly challenging. Although recent efforts have begun shifting towards open-vocabulary semantic segmentation by transferring 2D recognition capabilities to 3D space, most implementations focus on controlled, dense indoor scenes. Unlike indoor data, aerial urban point clouds present unique issues, including significant class imbalance, varying levels of point density, and vast spatial scales. Consequently, it is non-trivial to directly project indoor-focused pipelines into the challenging outdoor ALS domain. In addition, in many public ALS point cloud datasets, the original images are discarded due to storage constraints, making it challenging to perform 2D-3D transfer.

Motivated by these research gaps, we develop a zero-shot, open-vocabulary ALS point cloud semantic segmentation framework based on three types of VFMs. The framework can adaptively generate informative global views, enabling seamless 2D-3D transfer without the need for synchronously acquired imagery. We further conduct an empirical study to explore the potential of adapting 2D VFM for zero-shot semantic segmentation of 3D ALS point cloud. Overall, our main contributions are as follows:

- To address the absence of original imagery in public ALS datasets, we design an adaptive global view projection module that derives optimal virtual camera poses and field-of-view (FOV) from scene extents, enabling the seamless application of 2D VFMs to 3D data.

- We develop three VFMs-based zero-shot, open-vocabulary semantic segmentation methods for ALS point clouds and conduct empirical study to explore the potential of adapting 2D VFMs for zero-shot semantic segmentation of ALS point cloud.

2. Related Works

2.1 Fully Supervised Point Cloud Segmentation

The interpretation of 3D point clouds has evolved from basic geometric grouping to complex semantic scene parsing. Early strategies for point cloud segmentation mainly leveraged handcrafted spatial descriptors and clustering heuristics—such as the Random Forest classifier (Ni et al., 2017) or the RANSAC algorithm (Grilli and Remondino, 2020) to group the point into meaningful clusters. Although effective for simple geometries, these heuristic-based methods struggle with the inherent noise and varied densities in point cloud. The field has since moved toward deep neural architectures. PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) learn feature representations directly from raw coordinates by using symmetric functions to aggregate global features and address permutation invariance. Kpconv (Thomas et al., 2019) and MinkowskiNet (Choy et al., 2019) subsequently introduce a convolution-based approach to reduce computation. Recent breakthroughs have seen the rise of transformer-based backbones, including the Point Transformer series (Wu et al., 2024) , which employ self-attention to capture long-range dependencies to achieve state-of-the-art results. Despite high accuracy, these methods rely heavily on fixed, pre-labeled classes from training set, which limits their adaptability and generalization capabilities when encountering novel or untrained object classes.

2.2 Open Vocabulary Segmentation in Images

The open-vocabulary segmentation paradigm aims to overcome the limitations of closed-set labels by leveraging large-scale vision–language models (VLMs) that align visual features with natural language embeddings. CLIP establishes a joint latent space for images and text via dual-stream encoders, achieving a remarkable level of robustness in zero-shot 2D scene understanding (Radford et al., 2021). However, learned embeddings are primarily image-level representations and thus are not applicable for dense prediction tasks. Recent research has introduced modular adapters to bridge the gap between CLIP’s image-level pre-training and the requirements of dense pixel-level prediction, such as OPMapper (Wang et al., 2025c), LSeg (Li et al., 2022) and OpenSeg (Ghiasi et al., 2022). As SAM (Kirillov et al., 2023) facilitates the generation of masks independently of class, some mask-first approaches are coupled with linguistic priors to identify novel categories in unstructured visual data, such as Mask DINO (Li et al., 2023), MaskClip (Dong et al., 2023), ODISE (Xu et al., 2023). This process enables high-quality, multi-scale segmentation.

2.3 Open Vocabulary Segmentation in Point Cloud

Driven by the success of 2D vision–language frameworks, recent efforts have focused on lifting open-vocabulary semantics into the 3D domain. Existing methods typically distill knowledge from 2D VLMs into 3D backbones or align class-agnostic 3D proposals with multi-view image features. OpenScene (Peng et al., 2023) learns a dense representation for points that are aligned with text and image pixels in the CLIP feature space. By generating captions for multi-view images from 3D scenes, PLA(Ding et al., 2023) transfers knowledge from pre-trained vision–language foundation models to build explicit associations between 3D representations and semantically rich textual descriptions. Alternatively, distillation-based approaches like SEAL (Liu et al., 2023) leverage 2D–3D correspondences to transfer knowledge from 2D vision–language foundation models into 3D networks. Although there have been a number of studies on zero-shot learning for 3D point clouds, their application to raw large-scale urban point clouds is still recent and scarce. CitySeg (Xu et al., 2025) proposes a city-scale point cloud foundation model that incorporates a textual modality to enable open-vocabulary semantic segmentation and zero-shot generalization across heterogeneous UAV datasets. OpenUrban3D (Wang et al., 2025a) leverages multi-view rendered point clouds and vision–language feature distillation to achieve zero-shot semantic understanding in large-scale urban scenes.

3. Methodology

Following the methodological foundations established by (Zhang et al., 2024) and (Ye et al., 2025), we provide 3 zero-shot, open-vocabulary semantic segmentation solutions based on VFM tailored for ALS point clouds, namely Grounding DINO+SAM, SAM+CLIP, GSNET, shown in Fig. 1. Grounding DINO+SAM and SAM+CLIP directly employ general VFM that were pre-trained solely on natural images. GSNET integrates CLIP and Grouding DINO, which have been fine-tuned on remote sensing imagery to alleviate the inherent domain gap.

Grounding DINO+SAM: Grounding DINO, an open-set object detector, excels at localizing objects based on open-vocabulary text prompts, thereby acting as a robust automatic prompt generator for SAM. Specifically, Grounding DINO takes image and text prompts as input to identify object locations and labels. The Grounding DINO decoder generates precise bounding boxes corresponding to the textual description of the interest class, which serves as input for the SAM. SAM utilizes these box prompts and text prompts to narrow down the areas where SAM makes predictions and filters the predictions, thereby generating accurate semantic segmentation masks. The combination of Grounding DINO and CLIP achieves strong zero-shot semantics segmentation, mitigating the dependency on extensive labeled datasets.

SAM+CLIP: SAM takes grid points as input prompt to produce a set of instance masks. Subsequently, CLIP imbues these instance masks with semantic meaning by comparing the visual embeddings of each generated mask with the textual embeddings of predefined semantic categories. An embedding similarity threshold is set to filter out unreliable masks, thus generating the semantic prediction results for each specified category. This approach transforms the SAM output from a collection of arbitrary instances into a meaningful semantic segmentation map.

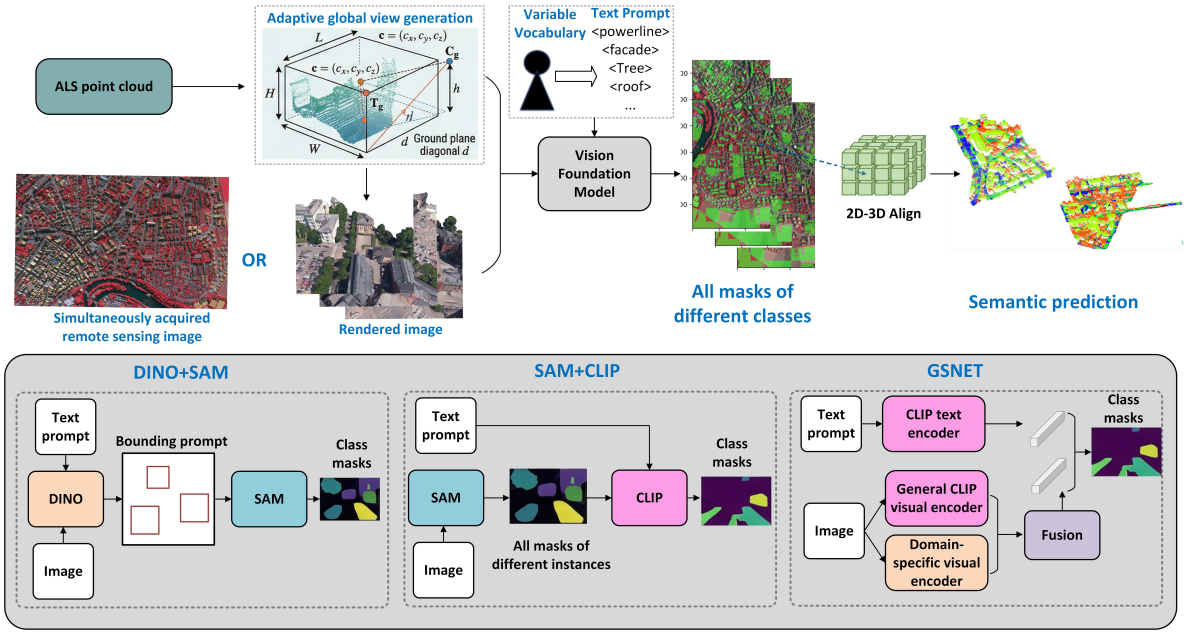

Figure 1. Workflow of zero-shot semantic segmentation for ALS point clouds.

GSNET: GSNET integrates a domain-specific remote sensing image encoder and a general CLIP-based image-text aligned encoder. The specialized RSI backbone is pre-trained on remote sensing imagery using a self-supervised learning paradigm, whereas the CLIP backbone is trained on large-scale image–text pairs through contrastive learning. The dual-stream features produced by the two encoders interact and complement each other under the guidance of variable vocabularies, and are subsequently fused to yield accurate segmentation masks.

With semantic prediction from the methods above, we transform point cloud coordination from the object coordinate system P to the image coordinate system. Given rotation R, projection centers $P_0 = (X_0, Y_0, Z_0)^T$ and the focal length f from the calibration information, the transformation from the object coordinate system $P = (X, Y, Z)$ to camera coordinate (x_c, y_c) can be formally expressed as Eq.1.

$$\begin{aligned} P_c = (X_c, Y_c, Z_c)^T &= R^T \cdot (P - P_0); \\ x_c &= -f \cdot \frac{X_c}{Z_c}; \\ y_c &= -f \cdot \frac{Y_c}{Z_c}; \end{aligned} \tag{1}$$

Given the pixel size of the camera Δ and the principal point coordinate (row_p, col_p), the camera coordinate systems could be converted to the image coordinate system via Eq.2.

$$\begin{aligned} col &= col_p + \frac{x_c}{\Delta}; \\ row &= row_p + \frac{y_c}{\Delta}; \end{aligned} \tag{2}$$

As the positions of points and pixels are well-aligned, we perform spatial projection from 2D image to 3D point cloud, transferring 2D pixel semantic prediction to 3D point, obtaining zero-shot open-vocabulary ALS point cloud semantic segmentation results.

For most of public datasets, the original synchronously captured images have been discarded due to storage constraints, limiting the application of VFMs. To address the challenge of missing imagery, we design a global view projection module to generate information-rich rendered images (see Algorithm 1).

Algorithm 1: Global View Generation for Point Cloud Rendering

Input: Point cloud $\mathcal{P}$; bounding box parameters (x_0, y_0, z_0, L, W, H); safety factor k; maximum FOV $\mathrm{FOV}_{\max}$

Output: Global camera parameters $\{\mathbf{C}_g, \mathbf{T}_g, \mathrm{FOV}_g, \mathbf{E}_g\}$

Step 1: Compute Scene Center

$\mathbf{c} \leftarrow (c_x, c_y, c_z) = \left(x_0 + \frac{W}{2},\ y_0 + \frac{L}{2},\ z_0 + \frac{H}{2}\right)$

Step 2: Define Global Camera Position

$z_g \leftarrow z_0 + H + \sqrt{L \cdot W}$

$\mathbf{C}_g \leftarrow (c_x, c_y, z_g)$

Step 3: Define Target Point

$\mathbf{T}_g \leftarrow \left(c_x, c_y, \frac{H+\sqrt{L \cdot W}}{2}\right)$

Step 4: Compute Camera Extrinsics

$\mathbf{E}_g \leftarrow \mathrm{LookAt}(\mathbf{C}_g, \mathbf{T}_g)$

Step 5: Compute Field-of-View (FOV)

$d \leftarrow \sqrt{L^2 + W^2}$; `// Diagonal of ground plane`

$h \leftarrow z_g - c_z$; `// Camera height above scene center`

$\mathrm{FOV}_g \leftarrow 2 \arctan\left(\frac{d}{2h}\right)$

Step 6: Apply Safety Margin

$\mathrm{FOV}_g \leftarrow \min\left(k \cdot \mathrm{FOV}_g,\ \mathrm{FOV}_{\max}\right)$

return $\{\mathbf{C}_g, \mathbf{T}_g, \mathrm{FOV}_g, \mathbf{E}_g\}$

Before generating the virtual camera pose, we first obtain the bounding box of the point cloud scene, defined by its length L, width W, and height H. Given the origin of coordinate systems at $(x_0, y_0, z_0$, the geometric center of the scene is defined as $\mathbf{c} = (c_x, c_y, c_z) = \left(x_0 + \frac{W}{2}, ; y_0 + \frac{L}{2}, ; z_0 + \frac{H}{2}\right)$.

Global Views Generation: We define global viewpoints of the virtual camera to capture a large-scale rendered image. The global virtual camera is located above the center point of the

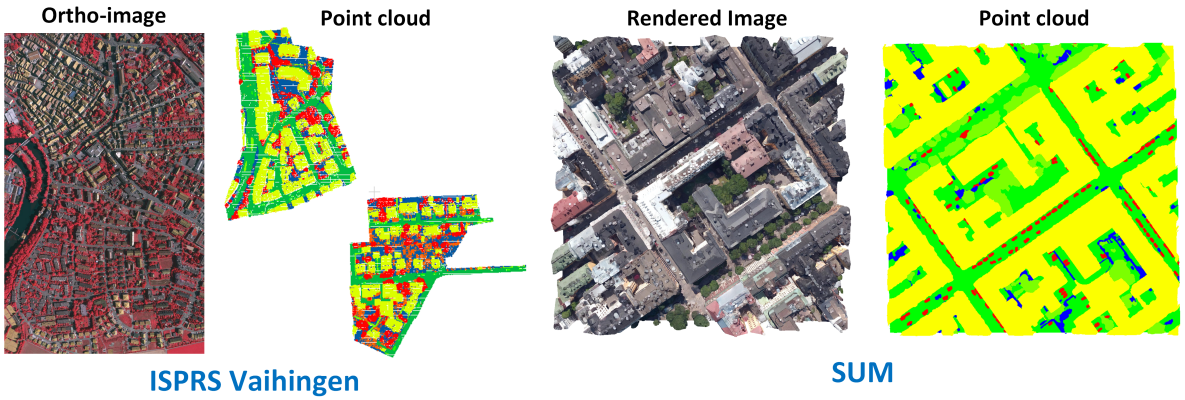

Figure 2. Example visualization from the Vaihingen and the SUM dataset, including simultaneously acquired ortho-image for Vaihingen and rendered images for SUM (left), and the ground-truth annotated point cloud (right).

Methods	powerline	low veg	imp surface	car	fence/hedge	roof	facade	shrub	tree
DINO+SAM	0%	36%	1%	21%	2%	54%	4%	12%	31%
SAM+CLIP	6%	34%	11%	1%	2%	79%	6%	14%	52%
GSNET	8%	94%	33%	4%	11%	70%	17%	17%	55%

Table 1. Zero-shot semantic segmentation precision on ISPRS Vaihingen dataset.

Methods	powerline	low veg	imp surface	car	fence/hedge	roof	facade	shrub	tree
DINO+SAM	0%	38%	1%	17%	31%	26%	27%	51%	75%
SAM+CLIP	7%	16%	4%	5%	9%	26%	20%	19%	24%
GSNET	11%	1%	20%	54%	37%	87%	1%	6%	89%

Table 2. Zero-shot semantic segmentation recall on ISPRS Vaihingen dataset.

bounding box, with its world coordinates $\mathbf{C}_g$.

$$\mathbf{C}_g = \begin{cases} x = x_0 + W/2, \\ y = y_0 + H/2, \\ z = H + \sqrt{L \cdot W}, \end{cases} \tag{3}$$

The virtual camera is at the anchor point and always oriented towards the target point $\mathbf{T}_g$, forming a consistent global observation.

$$\mathbf{T}_g = \left(c_x, c_y, \frac{H + \sqrt{L \cdot W}}{2} \right). \tag{4}$$

After determining its world coordinate $\mathbf{C}_g$ and target point $\mathbf{T}_g$, the extrinsic value of the global virtual camera is calculated using a standard lookat function.

To guarantee that the entire scene is visible within each global view, the field-of-view (FOV) is adaptively computed based on the extent of the scene. Specifically, the diagonal d of the ground plane and the vertical distance h from the camera to the center of the scene are given below.

$$\begin{aligned} d &= \sqrt{L^2 + W^2}, \\ h &= z_g - c_z. \end{aligned} \tag{5}$$

Then the FOV is set to cover exactly the entire scene bounding box with length L, width W, defined as:

$$\mathrm{FOV}_g = 2 \arctan\left(\frac{d}{2h} \right). \tag{6}$$

In practice, a safety margin is applied to ensure complete coverage:

$$\mathrm{FOV}_g = \min\left(k \cdot \mathrm{FOV}_g, \mathrm{FOV}_{\max} \right), \tag{7}$$

where k is a constant safety factor, $(\mathrm{FOV}_{\max})$ is a predefined upper bound to avoid excessive distortion.

4. Experiments

4.1 Experiment Setup

Datasets: ISPRS Vaihingen dataset (Rottensteiner et al., 2012) and SUM dataset (Gao et al., 2021) are used to evaluate the effectiveness of the three methods. The ISPRS Vaihingen dataset is an open access dataset for 2D and 3D semantic labeling contest. The orthoimages were captured by aerial cameras with three bands (Near Infrared, Red, and Green). The ALS point clouds were captured by a Leica ALS50 system, with each point containing x–y–z coordinates, intensity and return count. Each point was labeled as one of the following nine classes (powerline, low vegetation, impervious surface, car, fence/hedge, roof, facade, shrub, and tree). The SUM dataset is a benchmark dataset of semantic urban meshes that covers about $4km^2$ in Helsinki (Finland), with six classes (Terrain, Vegetation, Building, Water, Vehicle, and Boat). The dataset provides two point cloud versions sampled from the mesh model without imagery, with densities of 30 points/m^2 and 300 points/m^2. In our experiments, we utilize the version at 300 points/m^2 with image rendered by our global views generation module. The dataset examples are illustrated in Fig. 2.

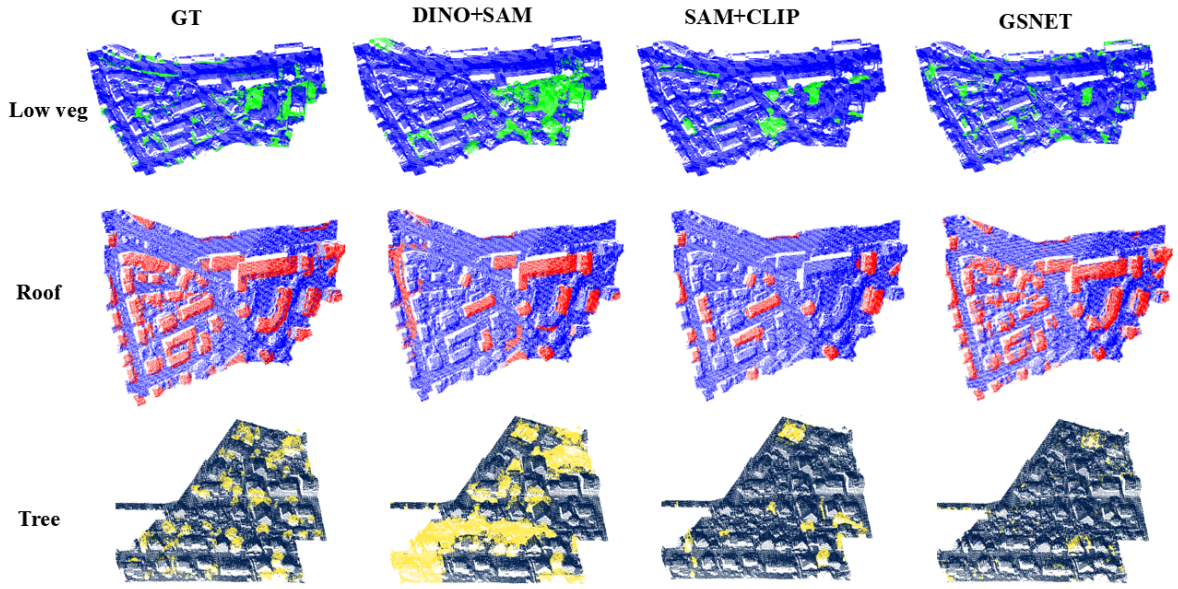

Figure 3. Result visualization from the Vaihingen dataset, including low vegetation (green), roof (red) and tree (yellow). The results of the different methods are shown from left to right.

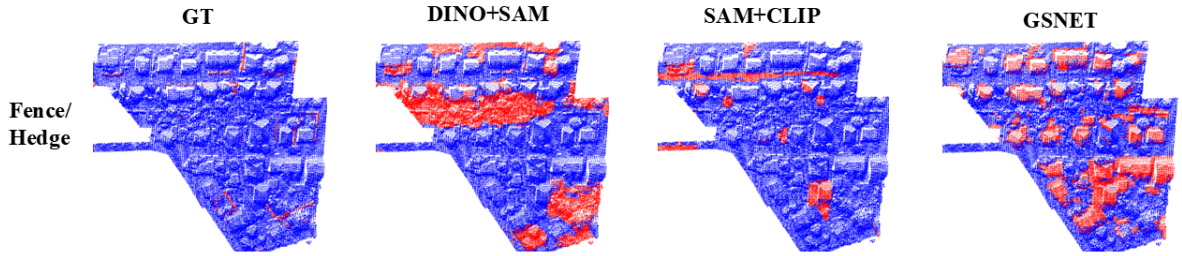

Figure 4. Result visualization from the Vaihingen dataset, including fence/hedge (red). The results of the different methods are shown from left to right.

Metrics: To comprehensively evaluate the performance of point cloud semantic segmentation, we employ precision and recall as class-wise evaluation metrics. For each semantic class, predictions are compared with ground truth labels to compute the number of true positives TP_c, false positives FP_c, and false negatives FN_c. Precision reflects the proportion of correctly predicted points among all points assigned to class (c), defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}. \tag{8}$$

Recall indicates the proportion of ground truth points of class (c) that are successfully retrieved, defined as:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{9}$$

The joint consideration of these two metrics provides a comprehensive assessment of segmentation quality, particularly in scenarios with class imbalance.

Implementation details: The methods are implemented based on Python, Pytorch and Open3d frameworks, and evaluated on a Ubuntu 18.04 workstation equipped with two NVIDIA RTX 6000 GPUs. For global view generation, we set the FOV safety factor k to 1.2, and the resulting rendered image has a resolution of 2000×2000. For the model architecture, we set the box threshold to 0.35 and the text threshold to 0.25 for Grounding DINO, and then set the similarity threshold to 0.05 for CLIP.

4.2 Experimental Results and Discussion

Tab. 1 and Tab. 2 present the quantitative results of the three methods on ISPRS Vaihingen dataset (precision% / recall%). The results demonstrate that all three approaches exhibit certain zero-shot open-vocabulary segmentation capabilities. The two methods that directly utilize VFM pretrained on natural images, namely DINO+SAM and SAM+CLIP, achieve satisfactory performance on common remote sensing categories such as roof and tree, but their segmentation accuracy degrades considerably for rare classes like powerline. In contrast, GSNET, which is finetuned on remote sensing imagery, yields performance gains across most categories, indicating that the domain discrepancy between natural and remote sensing images remains a key barrier to effectively deploying VFM in the remote sensing domain. However, even GSNET falls short of expectations when segmenting rare categories such as power line and car. Enhancing the segmentation capability for rare classes in zero-shot open-vocabulary settings therefore remains a significant and unresolved challenge.

Tab.3 presents the quantitative results of the methods on the SUM dataset (precision% / recall%). Since the original

synchronized images are not available, all methods rely on rendered images generated by the global virtual camera for zero-shot inference with VFM. The results indicate that using images rendered from high-density point clouds with RGB as input for VFM also yields satisfactory semantic segmentation results. Overall, DINO+SAM demonstrates a clear advantage in recall across most categories, particularly for large-scale structures such as terrain and building, indicating its strong capability in capturing complete object regions. In contrast, SAM+CLIP tends to produce higher precision in certain classes (e.g., terrain and water), but suffers from significantly lower recall, suggesting that it detects only limited portions of the target regions. In challenging small-object categories such as car and boat, both methods exhibit limited performance, with extremely low recall values. This can be attributed to the poor rendering quality of global views and the inherent difficulty of detecting small-scale objects in zero-shot settings.

Fig. 3 and Fig. 4 further visualize the results of different methods on the Vaihingen dataset. Compared to general-purpose VFMs (DINO+SAM, SAM+CLIP), GSNET gets better results on common large-scale categories when faced with the unique spectral and spatial characteristics of remote sensing imagery. For example, it achieves recall rate improvements of 60% and 14% for the “roof” and “tree” categories in the Vaihingen dataset, respectively. This highlights the importance of domain adaptation and demonstrates that encoders that incorporate domain-specific fine-tuning can effectively mitigate the differences between the natural domain and the remote sensing domain. However, the fine-tuning strategy can also lead to a decrease in the model’s generalization performance when faced with the same class with distinct geometric appearance, for example, weaker performance in ”low vegetation” compared to general-purpose VFMs. In addition, the most persistent challenge across all methods is the segmentation of rare and small-scale categories. In the Vaihingen dataset, the detection of powerlines and fences / hedges remains largely unresolved, with precision ranging between 0% and 8% for powerlines. This is likely due to the thin, weak geometry of small-scale objects, which provide insufficient visual features for 2D VFMs to differentiate from noise or ground.

Methods	DINO+SAM	SAM+CLIP
terrain	23.2% / 82.8 %	29.0% /11.0%
vegetation	23.6% / 50.2%	19.4% / 4.0%
building	62.5% / 84.3%	64.8% / 6.7%
water	14.5% / 42.6%	37.9% / 18.5%
car	28.6% / 0.6 %	1.4% / 7.2%
boat	0.3% / 0.2 %	0.2% / 0.7%

Table 3. Zero-shot semantic segmentation result on SUM dataset (precision% / recall%).

5. Conclusion

In this work, we develop a zero-shot, open-vocabulary ALS point cloud semantic segmentation framework based on three types of VFMs. The framework can adaptively generate informative global views, enabling the seamless application of 2D VFMs to 3D data. The results indicate that directly applying VFMs pretrained on natural images remains significant potential on ALS point cloud semantic segmentation. The domain gap and rare small-scale categories remain major obstacles to employ general-purpose VFMs in remote sensing applications. Commonly used fine-tuning strategies may lead to a degradation in the generalization ability of VFMs such as CLIP when faced with the same class exhibiting different geometric appearances. Achieving effective test-time adaptation and unsupervised domain adaptation for VFM therefore constitutes an important direction for future research.

References

Choy, C., Gwak, J., Savarese, S., 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X., 2023. Pla: Language-driven open-vocabulary 3d scene understanding. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7010–7019.

Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D. et al., 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10995–11005.

Gao, W., Nan, L., Boom, B., Ledoux, H., 2021. SUM: A Benchmark Dataset of Semantic Urban Meshes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179, 108-120.

Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y., 2022. Scaling open-vocabulary image segmentation with image-level labels. *European conference on computer vision*, Springer, 540–557.

Grilli, E., Remondino, F., 2020. Machine learning generalisation across different 3D architectural heritage. *ISPRS International Journal of Geo-Information*, 9(6), 379.

Guo, Z., Xu, R., Feng, C.-C., Zeng, Z., 2023. PIF-Net: A deep point-image fusion network for multimodality semantic segmentation of very high-resolution imagery and aerial point cloud. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–15.

Huang, R., Xu, Y., Hong, D., Yao, W., Ghamisi, P., Stilla, U., 2020. Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163, 62–81.

Jia, S., Pei, X., Yao, W., Wong, S. C., 2022. Self-supervised depth estimation leveraging global perception and geometric smoothness. *IEEE Transactions on Intelligent Transportation Systems*, 24(2), 1502–1517.

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.

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., Ranftl, R., 2022. Language-driven semantic segmentation. *International Conference on Learning Representations*.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L. M., Shum, H.-Y., 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3041–3050.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H. et al., 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European conference on computer vision*, Springer, 38–55.

Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z., 2023. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36, 37193–37229.

Ni, H., Lin, X., Zhang, J., 2017. Classification of ALS point cloud with improved point cloud segmentation and random forests. *Remote Sensing*, 9(3), 288.

Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T. et al., 2023. Openscene: 3d scene understanding with open vocabularies. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 815–824.

Polewski, P., Yao, W., Heurich, M., Krzystek, P., Stilla, U., 2015. Detection of fallen trees in ALS point clouds using a Normalized Cut approach trained by simulation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 252–271.

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.

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*, PmLR, 8748–8763.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction.

Shao, J., Yao, W., Luo, L., Zeng, L., He, Z., Wang, P., Fu, X., Qi, J., Guo, H., 2025. Geospatial big data-driven roof greening priority and benefit assessment in Hong Kong. *Remote Sensing of Environment*, 331, 115004.

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, C., Jing, K., Zhu, J., Wang, D., 2025a. OpenUrban3D: Annotation-Free Open-Vocabulary Semantic Segmentation of Large-Scale Urban Point Clouds. *arXiv preprint arXiv:2509.10842*.

Wang, P., Yao, W., 2022. A new weakly supervised approach for ALS point cloud semantic segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188, 237–254.

Wang, P., Yao, W., Shao, J., He, Z., 2025b. Test-time adaptation for geospatial point cloud semantic segmentation with distinct domain shifts. *ISPRS Journal of Photogrammetry and Remote Sensing*, 229, 422-435.

Wang, X., Si, C., Yang, X., Zhao, Y., Wang, W., Yang, X., Shen, W., 2025c. Opmapper: Enhancing open-vocabulary semantic segmentation with multi-guidance information. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S., 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2955–2966.

Xu, J., Wei, Z., You, W., Li, L., Sun, W., 2025. CitySeg: A 3D Open Vocabulary Semantic Segmentation Foundation Model in City-scale Scenarios. *arXiv preprint arXiv:2508.09470*.

Ye, C., Zhuge, Y., Zhang, P., 2025. Towards open-vocabulary remote sensing image semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39number 9, 9436–9444.

Zhang, J., Zhou, Z., Mai, G., Hu, M., Guan, Z., Li, S., Mu, L., 2024. Text2seg: Zero-shot remote sensing image semantic segmentation via text-guided visual foundation models. *Proceedings of the 7th ACM SIGSPATIAL International workshop on AI for geographic knowledge discovery*, 63–66.