Towards annotation-less semantic segmentation of aerial point clouds

Ashkan Alami [1,2], Fabio Remondino [1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy - Email: <aalami><remondino>@fbk.eu
[2] Department of Information Engineering and Computer Science, University of Trento, Italy

Keywords: point cloud, semantic segmentation, language models, deep learning, photogrammetry

Abstract:
The ability to automatically recognize a wide variety of objects in complex 3D urban environments without relying on predefined categories or annotated training data is becoming increasingly important for end-users of large-scale geospatial 3D datasets. Given that objects in urban scenes noticeably vary across locations, users and applications, flexible annotation-free methods for 3D semantic segmentation are getting desirable. In this work, we present and compare two approaches for classifying aerial photogrammetric point clouds. The first employs conventional supervised 3D neural networks trained on annotated datasets and predefined object classes. The second adopts a training-free, open-vocabulary strategy that detects objects directly in images and subsequently projects and refines them within 3D space. Approaches are evaluated through quantitative metrics and qualitative analysis, providing insights into their respective capabilities and limitations over 3D urban areas.

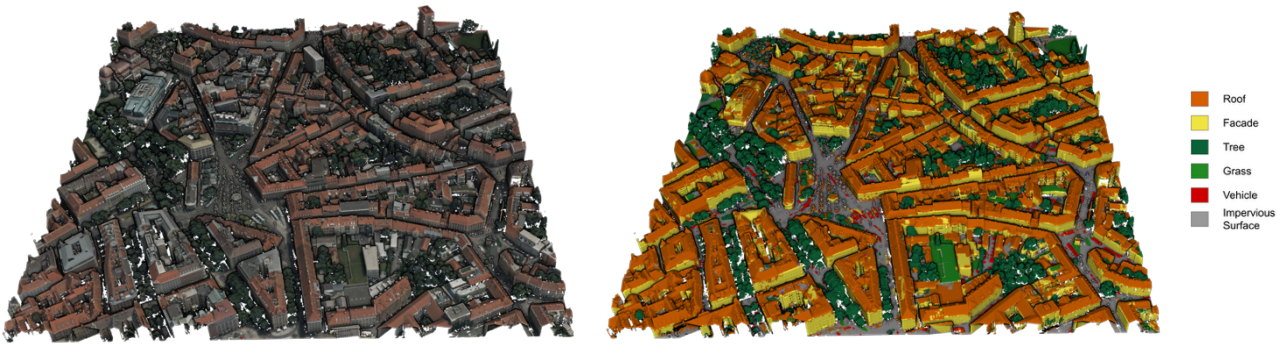

Figure 1: Unsupervised annotation-free semantic segmentation results (right) of an urban point cloud (left).

1. Introduction

3D scene understanding is becoming increasingly important in this era, with applications ranging from autonomous driving to urban planning and heritage preservation (Grilli and Remondino, 2020; Özdemir et al., 2021; Mao et al., 2023). Segmenting objects in a 3D scene is still a major challenge in the vision and geospatial communities. While many deep learning models have been developed in recent years and achieved promising methods (Guo et al., 2021; Zeng et al., 2022; Sun et al., 2024), significant bottlenecks remain: imbalanced classes, accuracy, misclassification, generalization, upscaling, etc. Solutions to address these challenges include data augmentation (Zhu et al., 2024), neuro-symbolic logic rules (Grilli et al., 2023), class weighting (Griffiths and Boehm, 2019), oversampling/undersampling techniques (Ren and Xia, 2023) or a multi-level multi-resolution approach (Bayrak et al., 2023).
Supervised 3D deep learning models are data-hungry, require extensive annotations and are not class agnostic. Acquiring and annotating the necessary 3D data is expensive. Consequently, recent developments are moving towards unsupervised or open-vocabulary (OV) segmentation methods (Chen et al., 2023a; Gelis, et al., 2023; Boudjoghra et al., 2024; Liu et al., 2024a; Nguyen et al., 2024), which aims to remove the limitation of fixed class labels. Generally, these methods utilize self-supervised learning (SSL) strategies or foundation 2D models, removing the requirement for 3D training data altogether. Application in indoor or small-scale scenarios are available whereas urban, forestry and large-scale scenarios are still of limited investigation (Takhtkeshha et al., 2024; Bieri et al., 2025; Ruoppa et al., 2025).

This work investigates how training-free open-vocabulary (OV) methods could pair or complement conventional supervised 3D neural network for urban point cloud classification (Figure 1). Experiments on urban datasets shows the capabilities and limitations of OV methods, highlighting their prospective to improve 3D segmentation performance across a wider range of urban classes and datasets.

2. Related works

2.1 Terminology jungle

The convergence of multiple communities and technologies has led from one side to open and commercial solutions able to automatically process large point cloud datasets but, from the other side to a set of new terms with a general lack of clear meanings for the geospatial sector. In the following we report some of the hyping terms and a definition to explain the meaning in particular for geospatial data:

- Transformers: Deep learning models based on self-attention mechanisms to process entire sequences of data in parallel, leading to a better understanding of context and long-term relationships in text, vision, and other types of data.
- Large Language Models (LLM): generative pre-trained transformers that can perform a variety of natural language processing (NLP) and analysis tasks, including translating, generating text, answering questions and identifying data patterns (also for images or other data, often called Visual or Multimodal Language Models).
- Vision-language model: an AI model that combines the capabilities of large language models (LLMs) with computer vision methods.

- Open-vocabulary: a deep learning object detection or segmentation method for images based on free-text prompts, without limitations on fixed set of categories. They combine vision-language models with object detection or segmentation architectures.
- Zero-shot: a deep learning method where a model is trained to recognize and classify new objects without explicitly being trained on those objects' examples.
- Few-shot learning: a deep learning method that uses a small number of examples for training a model.
- Foundation model: a 2D or 3D machine/deep learning model trained on vast datasets so it can be applied across a wide range of use cases (e.g. SAM, DINO, etc.).
- Generalization: the ability of a deep learning model to make accurate predictions on multiple scenarios and for input data it has never seen before.

2.2 3D deep learning segmentation methods

PointNet (Qi et al., 2017a) represented a significant pioneering breakthrough in deep learning models for point cloud segmentation. It analyzes point clouds directly by employing specialized operations such as max-pooling to maintain consistent understanding regardless of the point order. Similarly, PointNet++ (Qi et al., 2017b) introduces a hierarchical structure that covers local geometric details across multiple scales. KPConv (Thomas et al., 2019) performs convolution directly on point clouds by locating learnable kernel points in a local 3D neighbourhood. Each kernel point gives weights to close input points based on their spatial distance, allowing for flexible and geometry-aware feature learning. The flexible version enhances adaptability to complex shapes and varying point densities. Alternative approaches employ different architectures, such as the Graph Convolutional Neural Network (DGCNN) (Zhang et al., 2019) which utilize dynamically generated graphs in feature space to capture local neighbourhood information using edge convolution operations. Transformer architectures for 3D point cloud semantic segmentation, including Point Transformer (Zhoa et al., 2021), Superpoint transformer (Robert et al., 2023), Mask3D (Schult et al., 2023), developed rapidly and demonstrated successful results. Voxel-based – e.g. VoxelNeXt (Chen et al., 2023b), or graph-based – e.g. GTNet (Zhou et al., 2024) representations are also proposed as efficient solutions for semantic segmentation of 3D data.

2.3 Zero-shot and open-vocabulary segmentation methods

One of the main limitations of supervised deep learning methods is their high dependency on annotated data for training. For accuracy and high performance, these methods require training operations on large amounts of data, which is costly in terms of annotations, especially for point clouds. Moreover, despite the training cost, these methods are limited to the labels they are trained on and cannot perform well for new classes outside the scope of the training data. These limitations led to the development of zero-shot methods. In the image domain, researchers enrich zero-shot ability by training models on language and images together. CLIP (Radford et al., 2021) trained two encoders for images and text to align the text and vision embeddings (i.e. vision-language model). Currently, there are many vision-language methods (Minderer et al., 2022; Zhai et al., 2023) that connect the language domain with vision. Most of these models for object detection - i.e. GLIP (Li et al., 2022), OV-DETR (Zang et al., 2022) or Grounding DINO (Liu et al., 2024b) or segmentation - i.e. Mask DINO (Li et al., 2023), OV3D (Jiang et al., 2024) or Sa2VA (Yuan et al., 2025), provide open-vocabulary (OV) capability, which means they can perform detection/segmentation operations beyond their training data, simply by receiving the desired class name as text input.
In the 3D domain, recent classification methods are also moving toward OV-based segmentation. These methods highly rely on the power of previously mentioned 2D models. OpenScene (Peng et al., 2023) leverage on the OV 2D segmentation model OpenSeg (Ghiasi et al., 2022) for 3D scene segmentation. Similarly, ConceptFusion (Jatavallabhula et al., 2023) combine SAM - Segment Anything Model (Kirillov et al., 2023) with CLIP (Radford et al., 2021) to segment 3D scenes. Some methods use a class-agnostic 3D instance segmentor to generate masks. They subsequently apply OV 2D models. OpenMask3D (Takmaz et al., 2023) is an example of such a method, which uses Mask3D (Schult et al., 2023) to extract 3D masks. These masks are then projected onto the images from the scene, and 2D models are used to extract CLIP embeddings for each 3D mask. This allows them to later query the 3D scene and segment the desired objects. Search3D (Takmaz et al., 2025) builds a hierarchical OV 3D scene representation, enabling the search for entities at varying levels of granularity: fine-grained object parts, entire objects, or regions described by attributes like materials. Alami and Remondino (2024) presented a training-free and flexible method for indoor 3D point cloud segmentation using 2D OV models and geometric features. The method detects queried objects in images using 2D detectors such as YOLO-World (Cheng et al., 2024) and Grounding DINO, projects the masks to 3D and refines them with XGBoost-guided region growing. It does not use dataset-specific training and operates directly on the surveyed scene.

3. Semantic segmentation methods

3.1 3D Deep learning segmentation method

Due to their outstanding performance in various works (Chen et al., 2022; Bayrak et al., 2024), three 3D DL methods are used:

- KPConv[1]: the architecture is configured with 15 kernel points and an input radius of 18.0 m, and the initial subsampling distance is set to 0.3 m and the convolution radius to 2.5 m. The batch size is set to 3, Cross Entropy Loss function and Stochastic Gradient Optimizer are used with an initial learning rate of 10-3, and a momentum of 0.98. The learning rate is set to decrease exponentially, with a chosen exponential decay that guarantees a division by 10 every 100 epochs during a training of 250 epochs.
- PointNet++[2] (PN++): scenes are tiled into 6×6 m to 10×10 m sections, with 4096 points per tile. Coordinates are normalized (x, y to unit square; z shifted to the tile minimum). Classes are weighted by point count, normalized to mean 1. The model is trained for 100 epochs, but the epoch with the highest IoU is chosen for testing. Adam optimizer is used with a cyclic learning rate between 1×10^{-6} and 1×10^{-3}, step size 1000 and cycle momentum disabled.
- Superpoint Transformer[3] (SPT) (Robert et al, 2023, Robert et al 2024): scenes are partitioned into superpoints over tiles of about 300 × 400 m². Training is run for 2000 epochs with a batch size of 2. Optimization uses AdamW with an initial learning rate of $\sim5\times10^{-3}$. The best model is selected based on validation IoU.

[1] https://github.com/HuguesTHOMAS/KPConv
[2] https://github.com/yanx27/Pointnet_Pointnet2_pytorch
[3] https://github.com/drprojects/superpoint_transformer

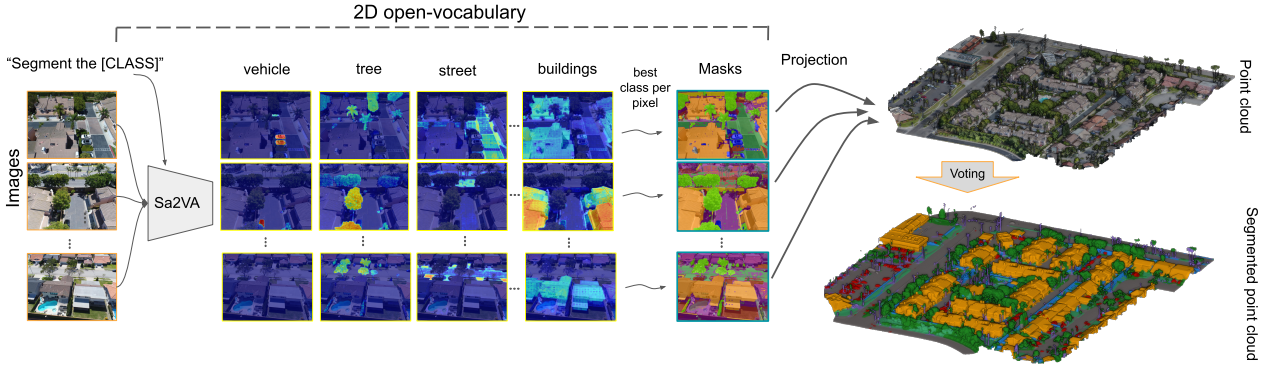

Figure 2: The proposed segmentation pipeline. Scene images are queried for a desired class using a 2D open-vocabulary (OV) model. For each image, the best class is selected for each pixel. The resulting segmented masks are then projected onto the point cloud and each point is labelled with the highest score through a voting process.

3.2 Open-vocabulary 3D segmentation method

We build upon the methodology presented by Alami and Remondino (2024), customizing the approach for aerial point clouds over urban areas and 2D OV models on large nadir and oblique aerial images (Figure 2). Their original projection method first voxelizes the point cloud, then uses ray casting to connect mask values from segmented images to the voxels and finally assigns those values to the points in each voxel. Here, after ray casting, all points visible are isolated in the segmented image and then, using camera parameters, projected back onto the segmented image to directly assign to each point its mask label. Due to the pipeline's modular design (Figure 2), different 2D open-vocabulary models are experimented. In particular, we adopted the Sa2VA model, which integrates LLaVA (Liu et al., 2023) with SAM (Kirillov et al., 2023). Unlike open-vocabulary detectors like Grounding DINO, which perform best with short, unambiguous class names, Sa2VA leverages a strong vision-language model capable of understanding natural language descriptions and segmenting objects from full-sentence prompts. This capability allows to identify classes described in more detail, rather than relying solely on a concise prompt. Therefore we further adapted Sa2VA by removing its built-in thresholding step, enabling us to obtain a per-pixel score for each queried class. For each class, we computed the score distribution and set a threshold equal to the mean score (although this can also be manually defined). Applying a threshold allows points below the threshold to be labelled as unknown/other; without it, the entire scene would be forced into one of the queried classes. For each image, the outputs from all classes are merged by assigning each pixel to the class with the highest score.

Due to the nature of urban aerial images and their cm-level GSD, some objects can be very small with respect to the original, extremely large image sizes. Moreover, passing a full aerial image to a 2D OV model is often computationally infeasible and small objects may be missed due to resizing. To address this, images are divided into smaller tiles, processed individually and then results are accumulated. For aerial images, tiling is applied in all cases, while for smaller images it is used only to detect small objects. Choosing the right tile size is crucial: tiles that are too small can fragment large objects, causing misdetection (e.g., a building wall might be classified as street if only part of it appears in the tile). In our experiments, tile sizes between 250-500 pixels worked well for the used data.

The segmented images are then projected back onto the point cloud. Since each 3D point can appear in multiple images, its final label is determined via a voting scheme, selecting the class with the highest cumulative score across all its projections.

Since the used point clouds have full image coverage, every point already has a label, making the original (Alami and Remondino, 2024) incomplete-coverage refinement method unnecessary. Moreover, geometric features per point at multiple radii are not used due to their computationally and memory intensive needs in dense aerial urban datasets. Instead, a lightweight noise-reduction step is used: for each point, we compared its label with those of its nearest neighbours and examined a very small set of geometric features (i.e. linearity, verticality and surface change). If a point's label differed from that of nearby points with similar geometric properties, its label is reassigned; otherwise, it remained unchanged.

4. Datasets and metrics

For the testing and validation procedures, the following datasets are used:

- STPLS3D (Chen et al., 2022), specifically the real-world USC scene
- Hessigheim 3D (Kölle et al., 2021)
- Graz (Farella et al., 2025).

Datasets	STPLS3D - USC	Hessigheim 3D	Graz
Source	Photogr.	Photogr.	Photogr.
Platform	Drone	Drone	Aircraft
# Classes	8 (9)	7 (11)	7
Size (km^2)	6	0.1	1.6
# Points (mil)	29	82	107
# Images	ca 4,500	ca 1,000	ca 50
Image type	Oblique	Nadir	Nadir + Oblique
Image size (px)	4,864 x 3,648 px	14,204 x 10,652 px	14,144 x 10,560 px
Avg GSD	1-2 cm	2-3 cm	5 cm (nadir)

Table 1: Employed datasets. In parenthesis the original number of classes wrt the used ones.

In the STPLS3D-USC scene, predictions are generated for the classes Vehicle, Pole, Tree, Building, Impervious Surface (road), Fence and Grass/Dirt (Dirt, which appears only rarely, is merged with Grass). For open-vocabulary methods, the unlabelled points are grouped as Others, while for deep learning models, the Clutter class is used as Others. As Sa2VA could not detect Clutter, no results are reported in Table 2.

For the Hessigheim 3D dataset, predictions are generated for Low Vegetation, Impervious Surface (road), Vehicle, Soil/Gravel.

Facade, Roof and Chimney are merged into the class Building; Shrub and Tree are merged; Vertical Surface, Urban Furniture and Unknown points are merged into the class Others. This is due to the fact that the available images are nadir, therefore vertical structures (like facade and chimney) are not clearly visible in the images, hence badly detectable by an OV method.
Fo the Graz dataset, predictions are generated for Façade, Roof, Tree, Grass, Vehicle and Impervious Surface (street/pavements). In order to report a complementary and robust evaluation, Intersection-over-Union (IoU) and F-1 Score metrics are calculated to demonstrate the spatial overlap between predictions and annotations as well as the relation between precision and recall, respectively.

5. Results

Results for the STPLS3D-USC dataset are reported in Table 2 and Figure 3. Grounding DINO queried terms such as vehicle, pole, tree, building, road, dirt, grass and fence. After detecting these objects in the images, SAM is used to extract object masks and project them onto the point cloud. Refinement is skipped because the available 4,500 images provided near-complete coverage of all points, making refinement unnecessary. On the other hand, for the Sa2VA-based method, more detailed descriptions/prompts are used requiring no additional tuning. Sa2VA outperformed the initial Grounding DINO + SAM setup, particularly for classes with large, continuous surfaces (e.g., streets, grass). While conventional 2D models detect these classes using bounding boxes that are then refined by SAM, Sa2VA bypasses the bounding box stage and produces per-pixel segmentations directly from text prompts. This difference is particularly relevant for classes with large or continuous extents, as it changes how segmentation boundaries are defined. Due to the large image size and the relatively small size of pole and fence objects, instead of downscaling the images, a tiling (1,024x1,024 px with some 256 px of overlap) is applied to help OV models detecting them more effectively. When 2D models internally resize the images, pole can vanish entirely, while fence often blend into nearby classes such as building, impervious surface or grass, therefore, tiling is only applied for these two classes. Other classes, such as building, tree and vehicle are already sufficiently visible in the full images, and tiling in these cases often led to misclassifications. For example, small image patches containing only a portion of a building are sometimes misinterpreted as impervious surface or other classes.

STPLS3D - USC						
Class	**3D DL: KPConv**		**OV: Grounding Dino + SAM**		**OV: Sa2VA**	
	IoU	F-1 Score	IoU	F-1 Score	IoU	F-1 Score
Building	93.75	96.78	77.17	87.11	82.43	90.37
Tree	86.23	92.6	65.78	79.36	70.55	82.73
Vehicle	47.44	64.36	37.47	54.51	28.46	44.31
Pole	50.62	67.21	18.98	31.91	6.16	11.6
Fence	26.41	41.78	1.98	3.89	6.61	12.39
Imp. surface	69.52	82.02	12.66	22.47	53.08	69.35
Grass & Dirt (*)	31.32	47.71	29.41	45.45	35.62	52.52
Others (**)	26.65	42.09	1.14	2.25	-	-

Table 2: Metric results on STPLS3D - USC dataset. (*) merged; (**): clutter and unknown.

As shown by the reported results, the performance of the open-vocabulary approach is highly dependent on the chosen 2D model. Accuracy can vary based on both the size of the queried object and the wording of the prompt. Class definitions also influence results. For example, when querying the OV model with road to detect impervious surface, it may not recognize pavements. Conversely, using pavement may miss actual road areas. Moreover, querying the exact impervious surface can be ambiguous for the model and often leads to reduced performance. In other cases, the OV model's interpretation of an object in the image does not match the annotation definition in the point cloud. For instance, for fence and pole classes, the OV models sometimes detect visually plausible areas, such as protective walls on rooftops or balconies (as fence) or the trunks of palm trees (as pole), that are respectively annotated as building or tree in the ground truth (Figure 3).

Hessigheim 3D				
Class	**3D DL: PN++**		**OV: Sa2VA**	
	IoU	F-1 Score	IoU	F-1 Score
Low Vegetation	77.23	87.15	47.11	64.05
Impervious Surface	78.76	88.12	46.67	63.64
Vehicle	33.93	50.76	13.7	24.09
Building (*)	83.67	91.11	76.19	86.49
Tree & Shrub	58.33	73.68	28.28	44.09
Soil / Gravel	24.50	39.36	7.13	13.32
Unknown / Others (**)	26.33	41.69	6.57	12.33

Table 3: Metric results on Hessigheim 3D dataset. (*) merged facade, roof and chimney. (**): unlabelled points, vertical surface and urban furniture.

For Hessigheim 3D, results are consistent with those on STPLS3D (Figure 4 and Table 3), with the 3D DL model achieving generally higher performance than the proposed OV method in all classes. For this dataset, the images are downscaled to 1,362x1,024 pixels. The original images are indeed too large in size and tiling has been avoided to preserve context. Indeed, unlike STPLS3D, where downscaling can make some classes harder to detect due a more complex scene, in Hessigheim 3D the classes remain visible and easier to detect after downscaling, so tiling is unnecessary. Moreover, some labels are inherently difficult for 2D models to interpret from the available images. For example, detecting building façades from nadir imagery is challenging and abstract classes such as urban furniture can encompass a wide range of objects. Prompting for every possible item in such a category is both time-consuming and expensive, while using a general prompt like "urban furniture" results in low detection rates and frequent failures. In general, for such cluttered environment, the most visually distinct and semantically simple classes are well detected by the 2D model.

Graz				
Class	**3D DL: SPT**		**OV: Sa2VA**	
	IoU	F-1 Score	IoU	F-1 Score
Roof	81.95	90.08	75.49	86.03
Façade	73.73	84.88	68.22	81.10
Grass	64.09	78.11	58.41	73.74
Tree	82.15	90.20	75.60	86.11
Vehicle	49.90	66.58	44.61	61.70
Imp. Surface	67.35	80.49	59.96	74.97
Others	14.41	25.19	-	-

Table 4: Metric results for the Graz dataset.

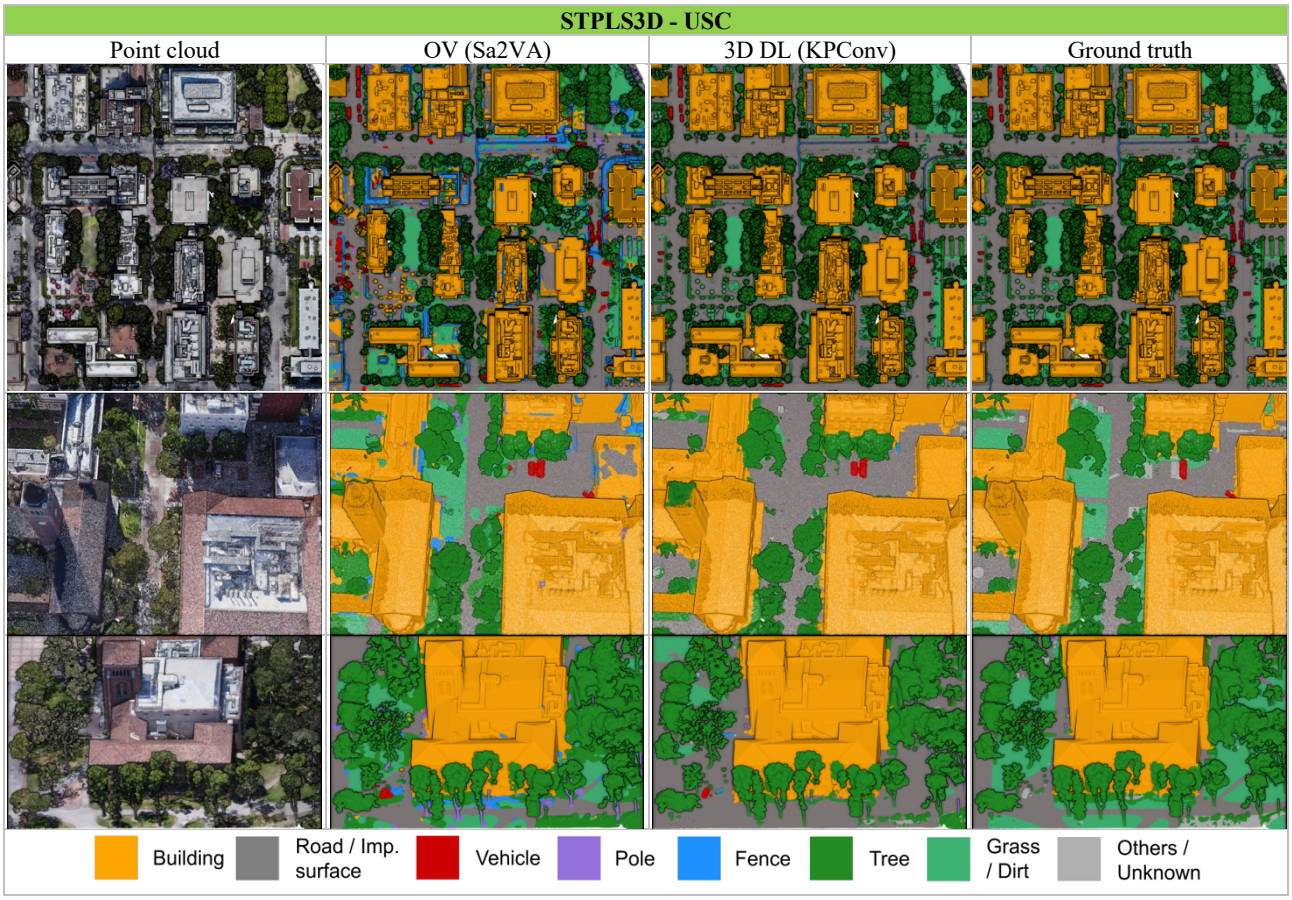

Figure 3: OV-based qualitative results for the STPLS3D scene with two close-up comparisons between open-vocabulary (OV) and deep learning (DL) predictions. The DL model produces sharper class boundaries by leveraging geometric information but shows reduced accuracy for radiometric classes such as grass (see also Table 2). The OV approach occasionally misinterprets class definitions, assigning labels such as pole to tree trunks or fence to small bushes and walls. OV segments classes with clear and unambiguous definitions more accurate. Please note the missing car in the upper-right GT figure, correctly detected by OV and DL methods.

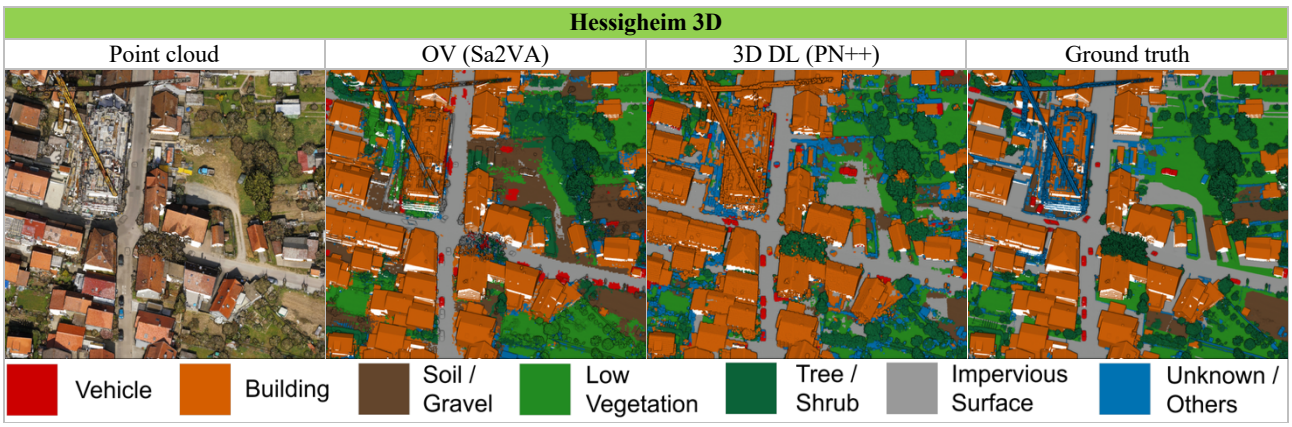

Figure 4: Visual comparison OV and DL approach on Hessigheim 3D. The OV approach struggles when images and point clouds are inconsistent (e.g., cars visible in images but absent in the point cloud due to asynchronous acquisitions). It can also miss objects due to visual ambiguities, for example, trees surrounded by low vegetation being mislabelled as low vegetation. Metrics in Table 3.

Results for the Graz dataset are reported in Figure 5-6 and Table 4. Similar to other experiments, OV approach is ideal for clear, unambiguous classifications. 3D DL produces clean and precise predictions for most of the classes.

6. Conclusions

The paper investigated how open-vocabulary (OV) methods could be used as a powerful semantic segmentation tool for urban 3D point cloud classification when training data are unavailable or when annotation efforts need to be minimized. With respect to conventional / supervised 3D deep learning (DL) methods and for the employed datasets, open-vocabulary segmentation methods tend to perform well for visually clear and unambiguous classes but not for highly specific classes or tiny objects not well recognizable in aerial images (e.g. powerline cables, poles, etc.). For clattered environments, conventional 3D deep learning trained on labelled data still provide superior performances.

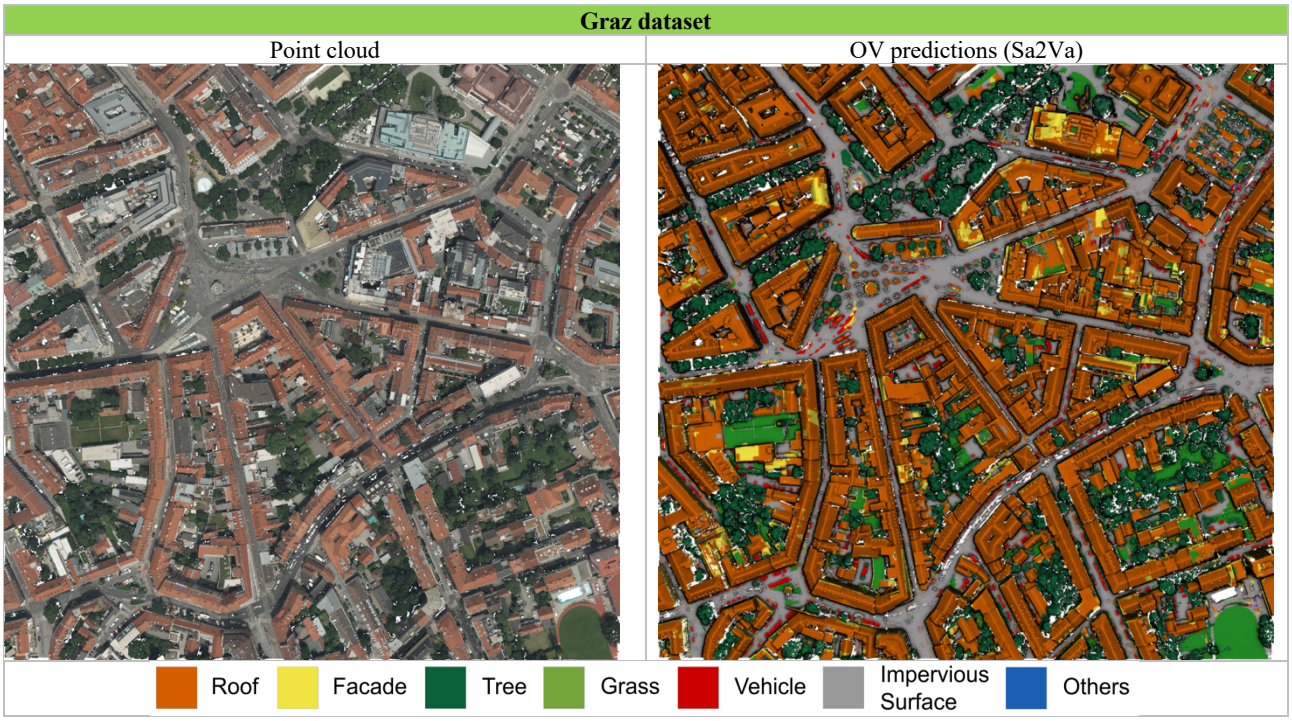

Figure 5: Qualitative results on the Graz dataset and the considered classes. Open-vocabulary methods can be particularly useful in scenarios where annotations or training data are unavailable, enabling scene segmentation without the need for model training.

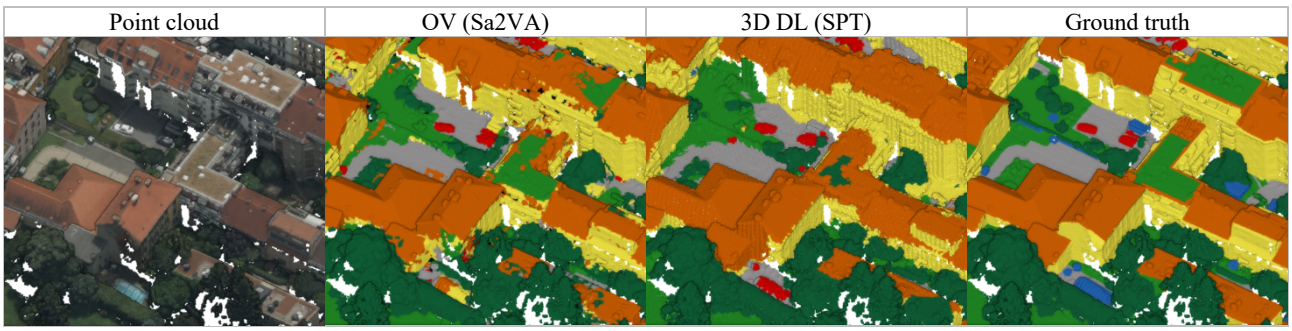

Figure 6: Visual comparison between OV and DL approaches on Graz (classes as in Figure 5). As in other datasets, the DL model provides sharper and more precise boundaries between objects. However, when classes are less ambiguous and visually simpler, the performance of both approaches becomes more comparable. Metrics in Table 4.

The proposed OV-based method achieved satisfactory results on most of the datasets and classes, revealing the potential of querying any object/class if clearly visible in the aerial images.

An important consideration is the temporal and spatial consistency between the 3D scene and the 2D images. Moving objects, such as vehicle, may be visible in the images but absent from the 3D reconstruction, creating prediction errors. A similar issue occurs with trees: in the photogrammetric 3D reconstruction, tree leaves may be missing due to smoothing or incorrect dense image matching, causing points - such as impervious surface or building - to be incorrectly predicted as tree when detections from the images are projected onto the point cloud.

Despite these issues, a key advantage of the open-vocabulary approach is that it requires no training or manual annotations and classes can be detected on the fly. It is worth noting that multimodal and vision-language models are evolving rapidly. Hence, in the near future, it is plausible that 2D open-vocabulary models used for 3D data classification could match or even outperform fully trained deep learning approaches in certain settings. The roadmap suggests combining the strengths of both paradigms. For example, open-vocabulary methods could be used to annotate part of a dataset automatically, using these labels to train a supervised 3D deep learning model (pseudo-labelling). Leveraging the complementary advantages of 2D vision-language models and 3D deep models could lead to improved segmentation performance across a wider range of urban classes and datasets.

Acknowledgments

Authors are thankful to colleagues Onur Bayak, Gabriele Mazzacca and Bartlomiej Besuch for their valuable support on supervised deep learning methods.

References

Alami, A. and Remondino, F., 2024. Querying 3D point clouds exploiting open-vocabulary semantic segmentation of images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLVIII-2/W8-2024, pp. 1-7.

Bayrak, O.C., Remondino, F., Uzar, M., 2023. A new dataset and methodology for urban-scale 3D point cloud classification. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 48, pp.1-8.

Bayrak, O. Ma, Z., Farella, E.M., Remondino, F., Uzar, M., 2024. ESTATE: A Large Dataset of Under-Represented Urban Objects for 3D Point Cloud Classification. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLVIII-2-2024, pp. 25-32.

Bieri, V., Zamboni, M., Blumer, N.S., Chen, Q., Engelmann, F., 2025. OpenCity3D: 3D Urban Scene Understanding with Vision-Language Models. Proc. *WACV*.

Boudjoghra, M., Dai, A., Lahoud, J., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.S., 2024. Open-YOLO 3D: Towards Fast and Accurate Open-Vocabulary 3D Instance Segmentation. Proc. 13th *International Conference on Learning Representations*.

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.

Chen, Z., Xu, H., Chen, W., Zhou, Z., Xiao, H., Sun, B., Xie, X., Kang, W., 2023a. PointDC: Unsupervised Semantic Segmentation of 3D Point Clouds via Cross-modal Distillation and Super-Voxel Clustering. Proc. *ICCV*.

Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J., 2023b. VoxelNext: Fully sparse VoxelNet for 3D object detection and tracking. Proc. *CVPR*, pp. 21674-21683.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y., 2024. YOLO-World: Real-time open-vocabulary object detection. Proc. *CVPR*, pp. 16901-16911.

de Gelis, I., Saha, S., Shahzad, M., Corpetti, T., Lefevre, S., Zhu, X., 2023. Deep unsupervised learning for 3D ALS point clouds change detection. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, Vol. 9, 100044.

Farella E.M., Morelli L., Remondino F., Qin R., Schachinger B. and Legat K., 2025. Investigating the new Ultracam Dragon hybrid aerial mapping system. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*

Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y., 2022. Scaling open-vocabulary image segmentation with image-level labels. Proc. *ECCV*, pp. 540-557.

Griffiths, D. and Boehm, J., 2019. Weighted point cloud augmentation for neural network training data class-imbalance. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-2/W13, 981-987.

Grilli, E., Daniele, A., Bassier, M., Remondino, F., Serafini, L., 2023. Knowledge enhanced neural networks for point cloud semantic segmentation. *Remote Sensing*, 15(10):2590.

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

Guo, Y., Wang, H., Hu., Q., Liu, H., Liu, L., Bennamoun, M., 2021. Deep learning for 3D point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.-H., 2025. Sa2va: Marrying SAM2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001.*

Jatavallabhula, K.M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Li, S., et al., 2023. ConceptFusion: Open-set multimodal 3D mapping. Proc. *RSS*, 2023.

Jiang, L., Shi, S., Schiele, B., 2024. Open-Vocabulary 3D Semantic Segmentation with Foundation Models. Proc. *CVPR*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P., Girshick, R., 2023. Segment Anything. Proc. *ICCV*, pp. 4015-4026.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J.D., Ledoux, H., 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open J. Photogramm. Remote Sens.,* 1, 100001.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.W., Hao, J., 2022. Grounded Language-Image Pre-training. Proc. *CVPR*.

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. Proc. *CVPR*.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.*, 36, 34892-34916.

Liu, J., Yu, Z., Breckon, T., Shum, H.P.H., 2024a. U3DS3: Unsupervised 3D Semantic Scene Segmentation. Proc. *WACV*.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2024b. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. Proc. *ECCV*, pp. 38-55.

Mao, J., Shi, S., Wang, X., Li. H., 2023. 3D object detection for autonomous driving: A comprehensive survey. *IJCV*.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., et al., 2022. Simple open-vocabulary object detection. Proc. *ECCV*, pp. 728-755.

Nguyen, P., Ngo, T.D., Kalogerakis, E., Gan, C., Tran, A., Pham, C., Nguyen, K., 2024. Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance. Proc. *CVPR*.

Özdemir, E., Remondino, F., Golkar, A., 2021. An Efficient and General Framework for Aerial Point Cloud Classification in Urban Scenarios. *Remote Sensing*, Vol.13, 1985

Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., 2023. OpenScene: 3D scene understanding with open vocabularies. Proc. *CVPR*, pp. 815-824.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. Proc. *CVPR*, pp. 652-660.

Qi, C.R., Yi, L., Su, H. and Guibas, L.J., 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Proc. *NeurIPS*, 30.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al., 2021. Learning transferable visual models from natural language supervision. Proc. *ICML*, pp. 8748-8763.

Ren, P., Xia, Q., 2023. Classification method for imbalanced LiDAR point cloud based on stack autoencoder. *Electron. Res. Arch.*, 31, 3453-3470.

Robert, D., Raguet, H., Landrieu, L., 2023. Efficient 3D Semantic Segmentation with Superpoint Transformer. Proc. *ICCV*.

Robert, D., Raguet, H., Landrieu, L., 2024. Scalable 3D Panoptic Segmentation as Superpoint Graph Clustering. Proc. *3DV*.

Ruoppa, L., Oinonen, O., Taher, J., Lehtomaki, M., Takhtkeshha, N., Kukko, A., Kaartinen, H., Hyyppa, J., 2025. Unsupervised deep learning for semantic segmentation of multispectral LiDAR forest point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing,* Vol. 228, pp. 694-722.

Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S. and Leibe, B., 2023. Mask3D: Mask transformer for 3D semantic instance segmentation. Proc. *ICRA*, pp. 8216–8223.

Sun, Y., Zhang, X., Miao, Y., 2024. A review of point cloud segmentation for understanding 3D indoor scenes. *Vis. Intell*., 2, 14.

Takhtkeshha, N., Bayrak, O.C., Mandlburger, G., Remondino, F., Kukko, A., Hyyppä, J., 2024. Automatic Annotation of 3D Multispectral LiDAR Data for Land Cover Classification, Proc. *IGARSS*, pp. 8645-8649.

Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F. and Engelmann, F., 2023. OpenMask3D: Open-vocabulary 3D instance segmentation. Proc. *NeurIPS*, 2023.

Takmaz, A., Delitzas, A., Summer, R.W., Engelmann, F., Wald, J., Tombari, F., 2025. Search3D: Hierarchical Open-Vocabulary 3D Segmentation. *IEEE Robotics and Automation Letters*.

Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F. and Guibas, L.J., 2019. KPConv: Flexible and deformable convolution for point clouds. Proc. *ICCV*, pp. 6411-6420.

Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C., 2022. Open-Vocabulary DETR with Conditional Matching. Proc. *ECCV*.

Zeng, J., Wang, D., Chen, P., 2022. A Survey on Transformers for Point Cloud Processing: An Updated Overview. *IEEE Access*, Vol. 10, pp. 86510-86527.

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training. Proc. *ICCV*, pp. 11975-11986.

Zhang, S., Tong, H., Xu, J., Maciejewski, R., 2019. Graph convolutional networks: A comprehensive review. *Comput. Soc. Netw.*, 6(1), pp. 1-23.

Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V., 2021. Point transformer. Proc. *ICCV*.

Zhou, W., Wang, Q., Jin, Q., Shi, Z, He, Y., 2024. Graph Transformer for 3D point clouds classification and semantic segmentation. *Computer and Graphics*, Vol. 124, 104050.

Zhu, Q., Fan, L., Weng, N., 2024. Advancements in Point Cloud Data Augmentation for Deep Learning: A Survey. *Patter Recongition*, 110532.