Research Article

Using 2D Vision-Language Models to resolve unbalanced training of 3D deep learning of street MMS data

Siyuan Meng[1], Longyong Wu[1], Fan Xue[1,*]

Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong, China
Correspondence: xuef@hku.hk

Abstract

The road digital twin (RDT) is important for various applications, from infrastructure planning to maintenance. Existing literature explored street object recognition. Whereas the segmentation accuracy of non-pole-like objects, such as utility boxes, fire hydrants, and trolleys, is relatively low because of the unbalanced training dataset and occlusion. This study, therefore, proposes a two-step method to automate the recognition of street objects based on mobile mapping systems. First, deep learning-based semantic segmentation is applied to point cloud data. Second, the segmented instances are aligned with street view imagery, and their semantic information is enriched using Vision-Language Models. This study takes streets in Central, Hong Kong, as a preliminary example to validate the proposed method. This study contributes an automated method for semantically enriching street objects, bridges pretrain large-scale human knowledge with 3D semantic segmentation training results, and provides novel perspectives for fields such as urban planning, engineering, built environment renovation, and transportation management. Future research directions include the development of automated digital twin applications for street infrastructure maintenance and management.

Keywords: Road digital twin, Street, Mobile mapping system, CIM, VLM, Deep learning,

Highlights

- Recognize street objects based on mobile mapping system using 3D deep learning and VLM.
- Average precision of 0.66 is achieved at the instance level.

1 Introduction

Road digital twin (RDT) is one of the important tools to assess, monitor, and maintain road infrastructure (Talaghat et al. 2024). Typical applications mainly focus on pavement maintenance, real-time traffic data streams simulation, and urban road planning.etc (Kušić, Schumann, and Ivanjko 2023; Talaghat et al. 2024). The geometry dimension of road digital twin refers to the 3D objects and their relationships on the road (Drobnyi et al. 2023), such as road surfaces, sidewalks, and street furniture.

Automated recognition of street objects is essential in RDT for lots of smart city application. Narrow segments, slopes, and barriers on the sidewalks may block the routines of people with limited mobility, especially in high-density cities(Meng et al. 2025). Identifying type of the detected barriers, such as traffic lights and trash bins, is helpful for accurately assessing wheelability and adjusting the position of barriers. Moreover, the enriched semantic information is needed for large-scale sidewalk maintenance simulation and optimization (Hou and Ai 2020).

The current automated recognition of street objects is limited in distinguishing the of sidewalk objects. Plenty of research discussed the automated geometry digital twin of road based on Scan-to-BIM workflow using deep learning (Justo et al. 2021; Pan et al. 2024). Problems such as occlusion and noise will affect the segmentation accuracy of the mobile laser scanning point cloud. Besides, 3D semantic segmentation requires a large-scale dataset, whereas none-pole-like objects, such as fire hydrants, utility boxes, litter bins, and trolleys, usually show up on the street less frequently, resulting in a low segmentation accuracy.

Mobile mapping system (MMS) scanning streets with both point cloud and street view imagery (SVI) has emerged as a promising solution to address the above limitations. The 3D-2D modal combination from point cloud and SVI frames helps recognize semantic information of street objects (Liu et al. 2025; Zhou et al. 2022). Deep learning, such as PointContrast, and vision language models (VLMs), such as ChatGPT and dinov2, can be utilized to recognize street objects from coarse to fine (Oquab et al. 2024; Xie et al. 2020).

Therefore, this study presents an automated street object recognition method using deep learning and VLM based on MMS to semantic enrich RDT. First, deep learning is used to segment point clouds; then, VLM is used to align and recognize objects based on SVI and point clouds.

The rest of this paper is organized as follows: Section 2 provides a review of relevant literature. Section 3 details the methods adopted in this paper. Section 4 reports the experimental findings. Sections 5 and 6 are dedicated to the discussion and conclusion, respectively.

2 Literature review

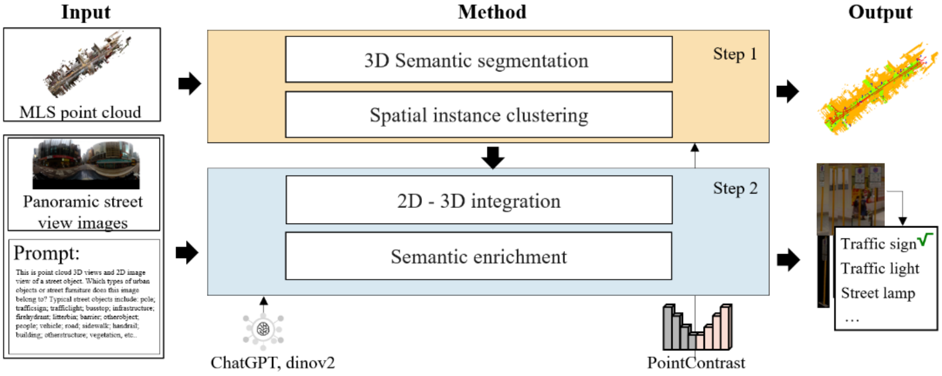
Street elements detection is conducted based on 2D, 3D, and 2D + 3D data sources. Initially, street elements quantification was achieved through semantic segmentation and object detection using SVI (Dai et al. 2024). 3D data source, such as mobile laser scanning point clouds and smartphone point clouds, was introduced to measure the 3D properties of street elements and further reconstruct RDT (Hou and Ai 2020; Meng et al. 2025; Pan et al. 2024). MMS, which equipped with laser scanning devices and high-resolution cameras, was utilized to transfer the 2D segmentation result over the 3D point cloud, using 2D deep learning, nerf, or VLM (Cao et al. 2025; Liu et al. 2025; Zhou et al. 2022). However, relying solely on the segmentation results of a single data source to detect street objects may lead to problems such as missed detection and false detection. Specifically, detection based on 2D data sources may result in false or missed detections due to factors such as insufficient resolution at long distances and the absence of a comprehensive training dataset. Besides, the accuracy of point cloud segmentation may be affected by factors such as noise and object occlusion. Combining the cross-modal features to recognize instance-level semantics can make up for the above issues.

The instance detection of street objects includes rule, learning, and VLM-based methods. The rule-based method is mainly helpful in detecting surfaces with simple structure, but fails to adapt to changing terrain and detect street objects (Babahajiani et al. 2017). The learning-based method such as PointNet and PointContrast showed better segmentation results and requires a carefully labeled dataset (Hou and Ai 2020; Meng et al. 2025; Pan et al. 2024). The VLM-based method has shown a certain degree of training-free detection of street objects using Groundingdino (Liu et al. 2025). However, due to the low pixel density of 2D images at long distances, it is easy to miss detections, for example, traffic signs across the street, and cannot be fully relied upon. The combination of the 3D-learning-based and VLM-based methods can be a feasible method to align multi-modal segmentation results from both 2D and 3D.

In summary, automatic street object recognition is needed for RDT reconstruction, road management, improving sidewalk conditions, and promoting equitable access to sidewalks for pedestrians. The current research gap lies in recognizing street objects such as non-pole-like objects.

3 Methods

Figure 1 shows an overview of the proposed two-stage method in this paper. Three input datasets are 3D point clouds and SVIs collected from MMS, a prompt for VLM semantic enrichment. First, instance-level street objects are segmented based on 3D point cloud using PointContrast and density-based spatial clustering of applications with noise (DBSCAN); then, images of 3D point clouds and 2D SVIs of clustered objects are extracted and semantically enriched by VLM.



This is point cloud 3D views and 2D image view of a street object. Which types of urban objects or street furniture does this image belong to? Typical street objects include: pole; trafficsign; trafficlight; busstop; infrastructure; firehydrant; litterbin; barrier; otherobject; people; vehicle; road; sidewalk; handrail; building; otherstructure; vegetation, etc..

Figure 1Proposed workflow

3.1 Research area

We chose urban Hong Kong as the research area in this paper. Hong Kong is a city famous for its high density of buildings and crowded streets. Narrow width, unique, and crowded street furniture make it hard to automatically reconstruct RDT. 1.2km of street in Central, Hong Kong, was collected as the train (totalling 1.0km) and validation (totalling 0.2km) dataset. The dataset was manually labelled and pre-processed using CloudCompare (ver 2.14). Ten classes were manually annotated for each point:

pole-like, non-pole-like, people, vehicle, road, sidewalk, handrail, building, vegetation, and outlier. Class pole-like includes objects such as traffic signs, poles, and traffic lights. Class non-pole-like includes all objects that are standing on the ground, such as fire hydrants, trash bins, and trolleys.

3.2 Step 1: Semantic segmentation

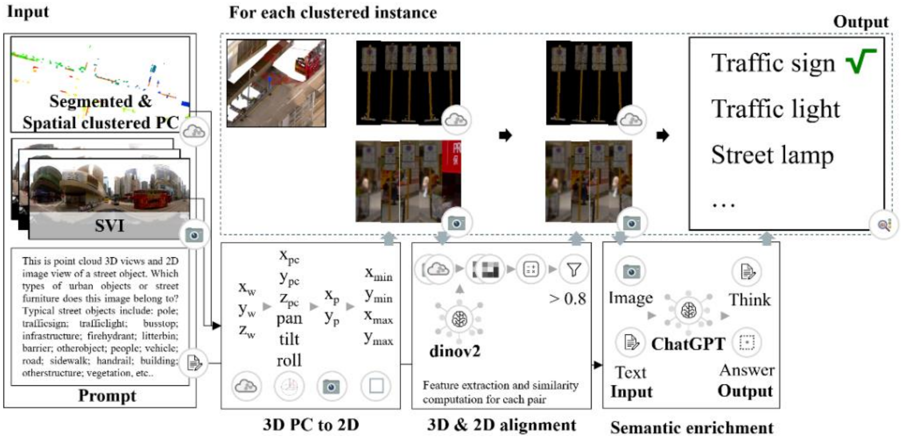
In this step, we take PointContrast as the semantic segmentation model due to its high accuracy and efficiency in outdoor semantic results (Meng et al., 2025; Xie et al., 2020). We follow the workflow of PointContrast and take Kitti as the pre-training dataset to perform contrastive learning (Meng et al., 2025). Based on the pre-trained model, we utilize the dataset in the research area to fine-tune the training process. Mean Intersection over Union (mIoU) was utilized to evaluate the segmentation results. To compare the proposed method with 3D semantic segmentation-based method, we train another model based on the refined-label dataset, e.g., class pole-like with pole, traffic sign, and bus stop; class non-pole-like with infrastructure and other-objects.

Then, DBSCAN is used to spatially cluster street objects of classes that can be represented as points (e.g., pole, other object, people, and vehicle). The input of DBSCAN is x, y coordinates to avoid point cloud discontinuity caused by occlusion and incomplete segmentation. The final output of this stage is the instance-level street objects.

3.3 Step 2: Semantic enrichment

This step enriches the spatially clustered instances. For each clustered instance, we extract the movable/immovable attributes using the following process, as shown in Figure 2.

First, project the instance point cloud into 2D based on the point cloud coordinates and the camera pose of a series of SVIs. Based on the projection results, the 2D image from SVI is extracted based on the corresponding instance bounding box. Then, dinov2 is employed to extract the dense features from each pair of images, since multiple SVIs may correspond to one instance (Oquab et al., 2023). The feature similarity score for each pair is computed, and pairs with a similarity score exceeding 0.8 are selected for further processing. All selected pairs of one instance are concatenated as an image prompt. The image prompt and text prompt, as shown in Figure 2, are set as input to ChatGPT. ChatGPT then returns the specific class of the input image prompt. The instance-level results are validated using metrics of precision, recall, and F1-score. The value TP denotes correctly identified instances, FP represents instances incorrectly identified, and FN indicates the instances that are not recognized successfully. Instances with IoU > 0.25 will be considered as correct.



This is point cloud 3D views and 2D image view of a street object. Which types of urban objects or street furniture does this image belong to? Typical street objects include: pole; trafficsign; trafficlight; busstop; infrastructure; firehydrant; litterbin; barrier; otherobject; people; vehicle; road; sidewalk; handrail; building; otherstructure; vegetation, etc..

Figure 2 Step 2

4 Results

4.1 Experimental settings

The experimental environment of semantic segmentation was set up in a high-performance computing cluster. A node on the cluster had dual Intel Xeon 6226R (16-core), 384GB RAM, and Nvidia V100 (32GB) GPU. We completed training based on PyTorch (Version 1.9) and Python (Version 3.8). The movable/immovable semantic information extraction was tested in single-threading mode on a desktop computer with Intel Core i7-13700K (3.40 GHz), 128 GB RAM, and Nvidia RTX A4000 GPU.

4.2 Semantic segmentation accuracy

Table 1 presents the point-level semantic segmentation results of two segmentation datasets, demonstrating satisfactory performance achieved on the datasets. The mIoU of both datasets achieved high semantic segmentation results, of 71.83 and 76.47, respectively. Classes of building, road, and vehicle achieved the highest semantic segmentation results in dataset, with IoUs equal to 98.13%, 93.01%, and 90.11%, respectively, which may be because they have relatively continuous surfaces and highly similar spatial structures. The pole-like class also achieved a high IoU score (85.80%), while the class object-like, due to their diverse types, only achieved an IoU score of 35.29%. In the refined dataset, pole-like classes, such as pole, traffic sign, and busstop, all achieved high segmentation results. However, infrastructure and other-object classes, only have 46.35% and 27.21% IoU.

Table 1 IoU result of semantic segmentation of two datasets.

Refined dataset		Dataset	
Class	**IoU**	**Class**	**IoU**
pole	80.72	pole-like	85.80
traffic sign	82.90		
bus stop	89.89		
infrastructure	46.35	non-pole-like	35.29
other-object	27.21		
people	78.64	people	77.52
vehicle	86.45	vehicle	90.11
road	89.98	road	93.01
sidewalk	66.47	sidewalk	71.89
handrail	69.50	handrail	65.72
building	97.19	building	98.13
other-structure	57.06		
vegetation	64.99	vegetation	69.81
outlier	68.29	outlier	77.42
mIoU	71.83		76.47

4.3 Semantic enrichment results

Table 2 demonstrates the precision, recall, and F1-score results of detected instances. There are 18 poles, 55 traffic signs, 4 bus stops, 10 infrastructure, and 15 other objects in the validation dataset. The 3D + VLM-based method performed better with classes of bus stop, traffic sign, infrastructure, and other-object, with precision equal to 1, 0.83, 0.5, and 0.5, respectively. However, the recalls of the 3D + VLM-based method are relatively low, except for the pole and infrastructure. The F1-score of non-pole-like objects is better than 3D-based result, contrary to pole-like objects. For infrastructure and otherobect with diverse geometry shapes, the 3D + VLM based method shows its ability in enriching semantic information.

The following reasons led to the lower result of pole-like objects compared to the 3D-based method: Firstly, the number of pole-like objects in the training dataset is enough for achieving high semantic segmentation accuracy. Whereas the 2D SVI will have lower pixel values for those objects far from the SVI camera, resulting in VLM's wrong classification. Besides, DBSCAN cannot separate objects that are very close to each other (e.g., surface distance < 0.1), which also results in low instance IoU values and fails to be recognized as TP.

Table 2 Precision, recall, and F1-score of detected street object instances.

Method / *Classes*	*3D*			*3D + VLM*		
	Precision	Recall	F1-score	Precision	Recall	F1-score
pole	0.59	0.72	0.65	0.45	0.83	0.59
busstop	0.8	1	0.89	1	0.25	0.4
trafficsign	0.74	0.57	0.65	0.83	0.43	0.56
infrastructure	0.25	0.56	0.34	0.5	0.78	0.61
otherobject	0.22	0.43	0.29	0.5	0.21	0.3
Average	0.52	0.66	0.56	0.66	0.5	0.49

5 Discussion

Street object information in RDT is essential for large-scale sidewalk maintenance simulation and optimization. The combination of multi-modal features and recognizing street objects on the streets enriches the semantic information in the RDT. Authorities are able to remove or adjust those objects that may influence walkability and wheelability based on those recognized object attributes.

There are several limitations and future work of the proposed method. First, the result of semantic segmentation of non-pole-like class exhibited lower performance. 2D VLM model, such as segment-anything, can subdivide imagery and further improve instance-level segmentation results (Kirillov et al. 2023). Furthermore, fine-tuned image-text models like CLIP can be used to achieve semantic classification with fewer samples, thereby enhancing the semantic enrichment capabilities of ChatGPT (Radford et al. 2021). Finally, the proposed method can be applied to applications such as wheelability calculation as well as CIM reconstruction, and further empower fields such as engineering, urban planning, and transportation to improve wheelability and promote equal mobility.

6 Conclusions

Street object recognition in RDT for serving pedestrians is essential because it can be used for applications such as pedestrian navigation, sidewalk maintenance, and renovation. However, current literature is limited in recognizing street objects because of unbalanced training data and occlusion of non-pole-like objects.

This study proposed a two-step method to address the above research gap. Based on 3D semantic segmentation results from PointContrast, we utilize VLM, including dinov2 and ChatGPT, to further enrich semantic information. The average precision equal to 0.66 demonstrates the feasibility of the proposed methods.

The proposed method contributes a precise, automated method for semantic enrichment of RDT. Future research direction mainly includes using VLM (e.g., segment-anything and CLIP) to improve instance-level segmentation and classification accuracy, as well as applying the proposed method to CIM reconstruction and 3DPN extraction.

Acknowledgements

The work presented in this paper was supported by the Hong Kong Innovation and Technology Commission (ITC) (No. ITP/029/20LP), and in part by the Department of Science and Technology of Guangdong Province (GDST) (No. 2023A1515010757).

Data Availability Statement

Data will be made available on request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Babahajiani, Pouria, Lixin Fan, Joni-Kristian Kämäräinen, and Moncef Gabbouj. 2017. “Urban 3D Segmentation and Modelling from Street View Images and LiDAR Point Clouds.” *Machine Vision and Applications* 28(7):679–94. doi:10.1007/s00138-017-0845-3.

Cao, Zhen, Xiaoxin Mi, Bo Qiu, Zhipeng Cao, Chen Long, Xinrui Yan, Chao Zheng, Zhen Dong, and Bisheng Yang. 2025. “Cross-Modal Semantic Transfer for Point Cloud Semantic Segmentation.” *ISPRS Journal of Photogrammetry and Remote Sensing* 221:265–79. doi:10.1016/j.isprsjprs.2025.01.024.

Dai, Shaoqing, Yuchen Li, Alfred Stein, Shujuan Yang, and Peng Jia. 2024. “Street View Imagery-Based Built Environment Auditing Tools: A Systematic Review.” *International Journal of Geographical Information Science* 38(6):1136–57. doi:10.1080/13658816.2024.2336034.

Davletshina, Diana, Varun Kumar Reja, and Ioannis Brilakis. 2024. “Automating Construction of Road Digital Twin Geometry Using Context and Location Aware Segmentation.” *Automation in Construction* 168:105795. doi:10.1016/j.autcon.2024.105795.

Drobnyi, Viktor, Zhiqi Hu, Yasmin Fathy, and Ioannis Brilakis. 2023. “Construction and Maintenance of Building Geometric Digital Twins: State of the Art Review.” *Sensors* 23(9):4382. doi:10.3390/s23094382.

Hou, Qing, and Chengbo Ai. 2020. “A Network-Level Sidewalk Inventory Method Using Mobile LiDAR and Deep Learning.” *Transportation Research Part C: Emerging Technologies* 119:102772. doi:10.1016/j.trc.2020.102772.

Justo, Andrés, Mario Soilán, Ana Sánchez-Rodríguez, and Belén Riveiro. 2021. “Scan-to-BIM for the Infrastructure Domain: Generation of IFC-Compliant Models of Road Infrastructure Assets and Semantics Using 3D Point Cloud Data.” *Automation in Construction* 127:103703. doi:10.1016/j.autcon.2021.103703.

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. “Segment Anything.”

Kušić, Krešimir, René Schumann, and Edouard Ivanjko. 2023. “A Digital Twin in Transportation: Real-Time Synergy of Traffic Data Streams and Simulation for Virtualizing Motorway Dynamics.” *Advanced Engineering Informatics* 55:101858. doi:10.1016/j.aei.2022.101858.

Liu, Chong, Mingyu Xie, Changzheng Yuan, Fuxun Liang, Zhen Dong, and Bisheng Yang. 2025. “Training-Free Open-Set 3D Inventory of Transportation Infrastructure by Combining Point Clouds and Images.” *Automation in Construction* 178:106377. doi:10.1016/j.autcon.2025.106377.

Meng, Siyuan, Xian Su, Guibo Sun, Maosu Li, and Fan Xue. 2025. “From 3D Pedestrian Networks to Wheelable Networks: An Automatic Wheelability Assessment Method for High-Density Urban Areas Using Contrastive Deep Learning of Smartphone Point Clouds.” *Computers, Environment and Urban Systems* 117:102255. doi:10.1016/j.compenvurbsys.2025.102255.

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu

Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. “DINOv2: Learning Robust Visual Features without Supervision.”

Pan, Yuandong, Mudan Wang, Linjun Lu, Ran Wei, Stefano Cavazzi, Matt Peck, and Ioannis Brilakis. 2024. “Scan-to-Graph: Automatic Generation and Representation of Highway Geometric Digital Twins from Point Cloud Data.” *Automation in Construction* 166:105654. doi:10.1016/j.autcon.2024.105654.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. “Learning Transferable Visual Models From Natural Language Supervision.”

Talaghat, Mohammad Amin, Amir Golroo, Abdelhak Kharbouch, Mehdi Rasti, Rauno Heikkilä, and Risto Jurva. 2024. “Digital Twin Technology for Road Pavement.” *Automation in Construction* 168:105826. doi:10.1016/j.autcon.2024.105826.

Xie, Saining, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. 2020. “PointContrast: Unsupervised Pre-Training for 3D Point Cloud Understanding.”

Zhou, Yuzhou, Xu Han, Mingjun Peng, Haiting Li, Bo Yang, Zhen Dong, and Bisheng Yang. 2022. “Street-View Imagery Guided Street Furniture Inventory from Mobile Laser Scanning Point Clouds.” *ISPRS Journal of Photogrammetry and Remote Sensing* 189:63–77. doi:10.1016/j.isprsjprs.2022.04.023.