UrbanBIS: a Large-scale Benchmark for Fine-grained Urban Building Instance Segmentation

Guoqing Yang
yanggq2020@gmail.com
Shenzhen University
China

Fuyou Xue
fullcyxuc@gmail.com
Shenzhen University
China

Qi Zhang
qi.zhang.opt@gmail.com
Shenzhen University
China

Ke Xie
ke.xie.siat@gmail.com
Shenzhen University
China

Chi-Wing Fu
philip.chiwing.fu@gmail.com
The Chinese University of Hong Kong
China

Hui Huang*
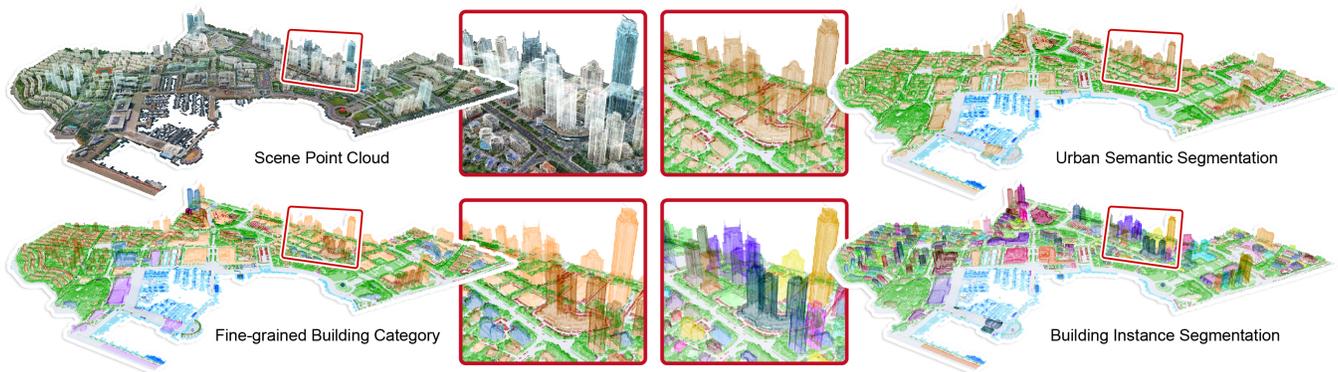hhzhiyan@gmail.com
Shenzhen University
China

Figure 1: UrbanBIS provides 2.5 billion 3D point samples over six scenes, covering a vast area of 10.78 km^2. Particularly, this large 3D dataset is annotated not only in the urban semantic level (top right) but also in the building instance level (bottom right) with fine-grained building categories (bottom left).

ABSTRACT

We present the *UrbanBIS* benchmark for large-scale 3D urban understanding, supporting practical urban-level semantic and building-level instance segmentation. UrbanBIS comprises six real urban scenes, with 2.5 billion points, covering a vast area of 10.78 km^2 and 3,370 buildings, captured by 113,346 views of aerial photogrammetry. Particularly, UrbanBIS provides not only semantic-level annotations on a rich set of urban objects, including buildings, vehicles, vegetation, roads, and bridges, but also instance-level annotations on the buildings. Further, UrbanBIS is the first 3D dataset that introduces fine-grained building sub-categories, considering a wide variety of shapes for different building types. Besides, we propose *B-Seg*, a building instance segmentation method to establish UrbanBIS. B-Seg adopts an end-to-end framework

with a simple yet effective strategy for handling large-scale point clouds. Compared with mainstream methods, B-Seg achieves better accuracy with faster inference speed on UrbanBIS. In addition to the carefully-annotated point clouds, UrbanBIS provides high-resolution aerial-acquisition photos and high-quality large-scale 3D reconstruction models, which shall facilitate a wide range of studies such as multi-view stereo, urban LOD generation, aerial path planning, autonomous navigation, road network extraction, and so on, thus serving as an important platform for many intelligent city applications. UrbanBIS and related code can be downloaded at https://vcc.tech/UrbanBIS.

CCS CONCEPTS

• **Computing methodologies** → **Shape modeling**.

KEYWORDS

Urban scene dataset and benchmark, urban semantic segmentation, building instance segmentation, point clouds

---

*Corresponding author: Hui Huang (hhzhiyan@gmail.com)

1 INTRODUCTION

3D machine learning is an emerging research topic, drawing great attention in recent years, as it facilitates a wide range of downstream tasks and applications. However, existing works focus mainly on tasks at the object level (e.g., object recognition, parts segmentation, shape synthesis, etc.) [Mo et al. 2019; Uy et al. 2019], at the indoor scene level (e.g., semantic segmentation, instance segmentation, floor plan recognition, etc.) [Roberts et al. 2021], and in outdoor road level (e.g., object detection, autonomous driving, etc.) [Aygun et al. 2021; Geiger et al. 2013; Golovinskiy et al. 2009; Zhou et al. 2020a] or mainly focuses on buildings [Selvaraju et al. 2021]. The bottleneck comes mainly from the need for datasets and the data annotations, required to train the machine learning models. In this work, we focus on building an urban-level dataset to facilitate 3D machine learning at the urban level, aiming to support large-scale urban-level analysis and understanding [Lafarge and Mallet 2012].

So far, urban datasets are acquired mainly by LiDAR [Behley et al. 2019; Boyko and Funkhouser 2014] or UAV photography [Gao et al. 2021; Li et al. 2020]. The data collection and annotation processes involve tremendous costs, so only few real-world urban-level datasets have been released, e.g., SensatUrban [Hu et al. 2022] and STPLS3D [Chen et al. 2022a]. Especially, the recent urban 3D point-cloud datasets [Chen et al. 2022a; Hu et al. 2022] provide only one to three scenes over relatively small areas. Also, most existing datasets provide only semantic-level annotations without instance-level annotations for real large-scale urban scenes, only small-scale dataset InstanceBuilding [Chen et al. 2022b] and some datasets containing virtual scenes provide instance information [Chen et al. 2022a; Griffiths and Boehm 2019]. Furthermore, they provide only urban-level semantics without fine-grained semantic information, e.g., building categories, thereby limiting the applications of the datasets.

Beyond the prior datasets, we propose UrbanBIS (Figure 1), the *largest real 3D urban dataset* that we are aware of. Compared with the latest 3D urban datasets SensatUrban [Hu et al. 2022] and STPLS3D [Chen et al. 2022a], UrbanBIS provides *2.5 billions of annotated 3D point samples* and *six real urban scenes*, covering a total area of $10.78km^2$ in different cities; see Section 3.3 for a quantitative comparison between UrbanBIS and existing urban datasets. Very importantly, UrbanBIS provides *both* semantic-level and instance-level annotations, as well as *fine-grained building categories*, which would facilitate many high-level tasks and applications on 3D machine learning.

Building UrbanBIS is a very tedious and expensive process, involving UAV image acquisition (0.5TB of 113,346 photos), 3D urban reconstruction, 3D mesh annotations, point cloud sampling (2.5 billion points in 3D), and fine-grained building annotations (3,370 buildings), consuming about 1,600 man-hours of annotation works in total. Besides, we propose a new instance segmentation method called B-Seg for handling large-scale point-cloud scenes (see Figure 2). Compared with the existing 3D instance segmentation methods, e.g., [Chen et al. 2021; Jiang et al. 2020; Vu et al. 2022], which typically adopt a time-costly point-wise clustering, we formulate an efficient pipeline in B-Seg to exploit point feature similarity and

segment the points through a relation matrix, thus avoiding the processing of all points in the whole large-scale urban scenes. Compared with existing methods, B-Seg achieves better performance and higher inference speed on the large-scale UrbanBIS dataset.

Overall, the contribution of our work is three-fold: (i) UrbanBIS, a large-scale real-world 3D urban dataset annotated with fine-grained building categories and instance segmentation information; (ii) B-seg, a fast and accurate new 3D point cloud instance segmentation method for buildings in large-scale urban scenes and (iii) A 3D urban platform with a rich variety of data that has great potential for developing many applications such as 3D reconstruction, depth prediction, multi-view stereo, and aerial path planning.

2 RELATED WORK

2.1 Indoor datasets for segmentation

Several datasets [Armeni et al. 2017; Dai et al. 2017; Roberts et al. 2021; Rozenberszki et al. 2022; Xiao et al. 2013] have been built for indoor scene parsing and understanding. Yet, they are mainly for classification [Uy et al. 2019], part segmentation [Mo et al. 2019], and indoor-scene semantic segmentation [Roberts et al. 2021].

Due to the expensive cost and difficulties of preparing the annotations, only a few indoor datasets [Armeni et al. 2017; Dai et al. 2017] provide instance-level masks for supporting the instance segmentation task. Dai et al. [2017] prepared ScanNet, an RGB-D video dataset, containing 2.5M views in 1,513 scenes annotated with 3D camera poses for surface reconstructions and segmentations. This dataset was obtained using commercial cameras, consisting of 21 categories and facilitating both 2D and 3D semantic and instance segmentation. Another common large-scale indoor dataset is S3DIS [Armeni et al. 2017], which covers a total area of $6000m^2$, which was acquired using the Matterport cameras. Beyond the 20 semantic categories in ScanNet [Dai et al. 2017], Rozenberszki et al. [2022] recently proposed ScanNet200, a new dataset with 200 semantic categories for fine-grained 3D indoor scene understanding.

2.2 Urban datasets

Urban scenes have unique features such as large-scale and complex lighting conditions, for which LiDAR and multi-view stereo are the major technologies for acquiring urban-level scene data.

According to the acquisition platform that carries the LiDAR [Behley et al. 2019], the scanning can be roughly divided into mobile laser scanning (MLS) [Munoz et al. 2009; Roynard et al. 2018; Tan et al. 2020], terrestrial laser scanning (TLS) [Hackel et al. 2017], and airborne laser scanning (ALS) [Kölle et al. 2021; Varney et al. 2020]. Okland [Munoz et al. 2009] is a point cloud dataset collected by MLS, consisting of 44 semantic categories, in which 24 of the categories contain less than 1,000 point clouds. Semantic3D [Hackel et al. 2017] is collected via TSL. It comprises eight semantic categories and provides two different benchmarks on 2D images and 3D points. The H3D dataset [Kölle et al. 2021] adopts a multimodal data acquisition method with a LiDAR and an optical camera installed on an airborne platform. It consists of 11 semantic categories. While LiDAR-based scanning gives promising results, its drawbacks are obvious. First, the scanning angle is fixed and there is an occlusion issue. Second, the scanning density is inversely proportional to the
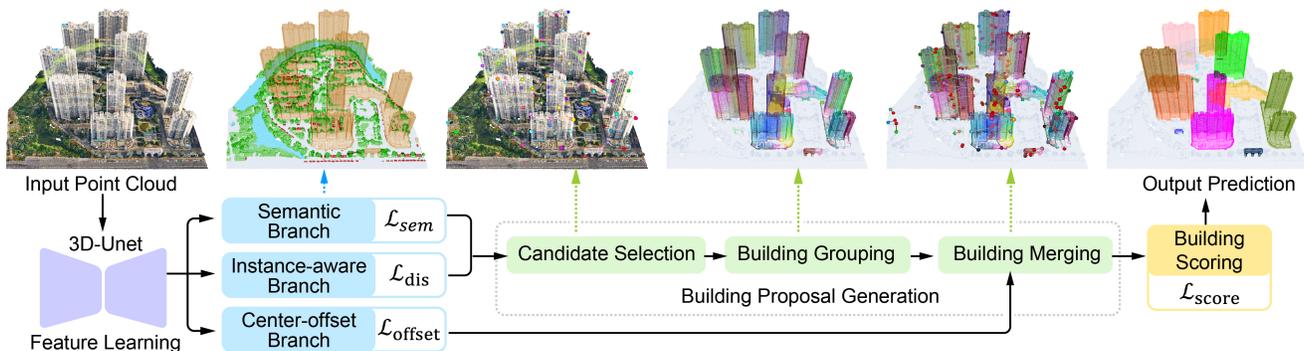
Figure 2: The pipeline of B-Seg: (i) backbone network and feature learning for point-wise feature extraction and sub-task predictions; (ii) building proposal generation to produce candidate building instances; and (iii) building scoring to filter the candidate instances and produce the final building instances.

object-sensor distance, leading to varying point patterns and sparse points for distant objects in the results.

Recently, researchers started to collect 3D urban data using multi-view stereo [Gao et al. 2021; Li et al. 2020]. Swiss3DCity [Can et al. 2021] uses an array of high-resolution cameras to capture images and provides more complete and dense point clouds for three Swiss cities using UAV photogrammetry. It also provides point clouds of different resolutions for research purposes. The SensatUrban dataset [Hu et al. 2022] was proposed to mitigate the small-scale and limited semantic annotations in the current dataset, providing three billion points that cover an area of $7.6km^2$ in three different urban scenes. It provides annotations of 13 semantic categories. The STPLS3D dataset [Chen et al. 2022a] contains instance-labeled urban virtual scenes and real scenes labeled only with semantic information. While the aforementioned datasets have contributed greatly to the development of research on urban understanding, they are annotated mainly with semantic information and lack instance- or building-level annotations (except STPLS3D and SynthCity [Griffiths and Boehm 2019], which contain instance labels only for virtual scenes). InstanceBuilding [Chen et al. 2022b] is a real urban scene dataset with building instances, but it is small in scale and has only two semantic categories.

In addition to the above urban scene datasets aimed for segmentation or detection, some urban scene datasets were proposed for other domains. Lin et al. [2022] built UrbanScene3D, a large-scale urban scene dataset for path planning and 3D reconstruction, providing ten virtual scenes and six real scenes. Besides, there are 3DCityDB [Yao et al. 2018] and the large-scale 3D geospatial data [Biljecki and Dehbi 2019] proposed for urban management and urban planning, providing visualizations and management of urban scenes.

2.3 3D Instance segmentation methods

Existing 3D instance segmentation methods are oriented mainly to indoor scenes. They can be roughly divided into proposal-based methods and proposal-free methods. Proposal-based methods are kind of a top-down approach that first generates region proposals such as bounding boxes then predicts instance masks [Hou et al. 2019; Yi et al. 2019]. 3D-BotNet [Yang et al. 2019] directly regresses

3D bounding boxes for object candidates with a multi-criteria loss. 3D-MPA [Engelmann et al. 2020] samples proposals from shifted centroids and uses a graph neural network [Wang et al. 2019b] to enhance the features. GICN [Liu et al. 2020] regards centroids of instances as a Gaussian distribution. Overall, these methods show results with good objectness, yet they lack efficiency. Also, the quality of their results highly depends on the proposals.

Proposal-free methods use a bottom-up style that directly extracts point features then clusters points into object instances [Chen et al. 2021; Han et al. 2020; Liang et al. 2021; Pham et al. 2019; Wang et al. 2019a; Zhao and Tao 2020]. SGPN [Wang et al. 2018] introduces a feature matrix to represent point-pair similarity to aid the clustering. MTML [Lahoud et al. 2019] groups instances from discriminative feature embedding with mean-shift clustering. Point-Group [Jiang et al. 2020] clusters object instances by simultaneously considering two different sets of point coordinates. SoftGroup [Vu et al. 2022] improves PointGroup [Jiang et al. 2020] with the soft semantic scores. Some other methods [He et al. 2021; Wu et al. 2022] adopt kernel-based strategies to exploit kernel features.

So far, existing 3D instance segmentation methods aim to handle mainly indoor scenes with small household-level objects. It remains unclear whether they can effectively process and segment large and complex urban-level scenes. Especially, modern buildings have diverse shapes, sizes, and appearances.

3 THE URBANBIS DATASET

3.1 UrbanBIS acquisition and annotations

3.1.1 Data acquisition. To obtain complete scene data for a large-scale urban area, we adopt aerial photogrammetry to collect images due to the flexibility of the UAV platform. To speed up the reconstruction process and to improve the quality of the reconstructed models, we adopt [Zhou et al. 2020b] to generate trajectories instead of using conventional oblique photogrammetry. The acquisition devices in our setting include the *DJI PHANTOM 4 RTK* drone[1] with the built-in camera and the *DJI M300RTK* drone[2] loaded with five

---

[1]https://www.dji.com/phantom-4-rtk
[2]https://www.dji.com/matrice-300

Table 1: Statistics of the collected aerial photos in UrbanBIS.

Scene	Area (km^2)	Image (#)	Resolution (pixels)	Size (GB)
Qingdao	2.31	98,373	$6,000 \times 4,000$	189.9
Wuhu	2.92	1,053	$14,204 \times 10,652$	140
Longhua	2.41	11,307	$8,192 \times 5,640$	162
Yuehai	1.60	1,030	$5,472 \times 3,648$	8.33
Lihu	1.46	728	$6,000 \times 4,000$	6.9
Yingrenshi	0.08	855	$5,472 \times 3,648$	6.97
Total	10.78	113,346	-	514.1

Table 2: UrbanBIS provides 2.5 billion sampled points in six scenes.

Category	Qingdao	Wuhu	Longhua	Yuehai	Lihu	Yingrenshi
Size (GB)	26.5	27.8	29.1	17.5	11.5	0.92
Building (#)	269.59M	285.28M	256.39M	117.98M	65.18M	14.97M
Ground (#)	114.22M	133.32M	158.62M	69.60M	80.54M	4.39M
Water (#)	11.46M	20.95M	0.26M	3.86M	2.46M	0
Boat (#)	4.20M	409	852	0	2,490	0
Vegetation (#)	179.50M	175.69M	225.50M	197.83M	104.09M	1.66M
Vehicle (#)	15.05M	8.24M	11.35M	1.16M	2.08M	0.85M
Bridge (#)	37,074	1.61M	1.77M	2.93M	0.78M	0.35M
Total (#)	594.06M	625.08M	653.90M	393.37M	255.12M	22.22M

HD PSDK 102S aerial cameras. Table 1 shows the basic information of the aerial photos (0.5 TB) we collected in six different scenes.

3.1.2 Mesh annotation. We follow conventional datasets [Gao et al. 2021] to define the semantic categories. Overall, UrbanBIS provides two kinds of semantic information: (i) urban-level semantics and (ii) building-level semantics. Below, we list the urban-level semantic categories: (i) *Ground*: Impervious surface, roads, parking, etc.; (ii) *Water*: Lake, river, sea, etc.; (iii) *Boat*: Cruise ships, small boats, etc.; (iv) *Vegetation*: Trees, lawn, bushes, etc.; (v) *Bridge*: Bridges across rivers or in the parks; (vi) *Vehicle*: Cars, buses, bikes, etc.; and (vii) *Building*: The instantiated categories, including various man-made buildings. Note that some datasets put grass into the ground category while others put it into vegetation. Here, we regard lawns and bushes as vegetation and regard wild grasses on the ground as ground. Figure 4 shows the semantic definition in UrbanBIS. The detailed number of mesh faces for each urban-semantic category can be found in the supplementary material.

3.1.3 Point cloud sampling. As the reconstructed 3D meshes lack water tightness, performing machine learning directly on them may lead to serious deviations in results. Hence, we generate point samples on the reconstructed meshes to represent the 3D scene. In detail, we adopt *CloudCompare*[3] to generate point samples in each 3D scene. Considering that the sampled points should be consistent with the geometry of the reconstructed 3D mesh and also uniform in distribution, we choose to sample the points based on the surface density of the mesh. Here, we set the number of sample points per square meter to 80 and represent each point by a 6D feature with 3D coordinates and RGB information. Table 2 reports the number of points for different urban semantics in each scene.

---

[3]https://www.cloudcompare.org/

Table 3: The distribution of building categories (middle) and building heights (right) in each scene in the UrbanBIS dataset. Co, Re, Of, Cu, Tr, Mu, and Te stand for Commercial, Residential, Office, Cultural, Transportation, Municipal, and Temporary, respectively. L, H, and SH stand for Low-rise, High-rise, and Super High-rise, respectively. We ignore buildings that are difficult to determine their functions during the data category annotation.

Scene	Co	Re	Of	Cu	Tr	Mu	Te	L	H	SH
Qingdao	64	238	26	8	18	106	124	554	77	27
Wuhu	24	813	32	7	0	47	117	1000	73	28
Yuehai	7	55	39	16	1	12	114	220	35	1
Lihu	1	14	26	7	4	44	211	300	21	1
Longhua	12	274	96	1	17	111	454	844	132	20
Yingrenshi	3	11	10	0	0	4	6	19	18	0
Total	111	1405	229	39	40	324	1026	2937	356	77

3.2 Fine-grained Building-level information in UrbanBIS

Concerning the building-level semantics, UrbanBIS provides not only segmentation information on building instances but also semantic information on buildings. Buildings can be divided into different categories based on their functions. According to the scheme proposed by the Department of Engineering Quality and Safety Supervision, Ministry of Housing and Urban-Rural Development [2009], buildings can be divided into 28 categories, e.g., residential, commercial, cultural, etc. Buildings of different functions usually have large variations in appearance and shape (see Figure 5). To provide a more detailed description of urban scenes, we further classify buildings into fine-grained sub-categories. By merging the building types based on the building functions, we consider the following 7 building categories: (i) *Commercial*: including commercial, catering, entertainment, hotel, financial, and exhibition buildings; (ii) *Residential*: including residential and dormitory buildings; (iii) *Office*: including office, research, medical, hygienic, and telecommunication buildings; (iv) *Cultural*: including cultural, museums, gyms, and religious buildings; (v) *Transportation*: including subway entrances and bus stops; (vi) *Municipal*: including waste disposal stations and electrical facilities; and (vii) *Temporary*: including temporary and garden buildings. Also, we consider another aspect of classifying buildings: *low-rise* (under 24m); *high-rise* (24m to 100m); and *super high-rise* (over 100m). Please refer to Table 3 for the detailed number of buildings in UrbanBIS for each category.

3.3 Analysis and comparison

As seen from Table 3 (middle) , UrbanBIS contains a rich variety of different building categories. Also, the six scenes exhibit different compositions of building categories. Interestingly, both Yuehai and Lihu contain fewer commercial or transportation buildings, but have a lot of office, cultural, and residential buildings, as they are both campuses. Besides, we can observe that the two campus scenes have a large proportion of temporary buildings, since both scenes contain a large number of prefab houses that are under construction. The Wuhu scene contains more residential buildings, while

Table 4: Comparing existing 3D urban segmentation datasets. MLS, TLS, and ALS stand for mobile laser scanning, terrestrial laser scanning, and aerial laser scanning, respectively. In the Application column, SS, IR, SC, and IS stand for semantic segmentation, image reconstruction, scene completion, and instance segmentation, respectively. Ptgy is Photogrammetry and sub-cat. is Sub-category. Only real scenes are considered in this table.

Dataset	Year	Acquisition	Data-type	Area/Length	Scene	Points/Triangles	Classes	Application	Annotation
Okland [Munoz et al. 2009]	2009	MLS	PC	$1.5km$	1	1.6 M	5	SS	Semantic
Semantic3D [Hackel et al. 2017]	2017	TLS	PC	-	3	4000M	8	SS	Semantic
Paris-Lille-3D [Roynard et al. 2018]	2018	MLS	PC	$1.94km$	2	143M	9	SS	Semantic
DublinCity [Zolanvari et al. 2019]	2019	ALS	PC/Image	$2km^2$	1	260M	13	SS/IR	Semantic
SemanticKITTI [Behley et al. 2019]	2019	MLS	PC	$39.2km$	1	4549M	25	SS/SC	Semantic
Toronto-3D [Tan et al. 2020]	2020	MLS	PC	$1.0km$	1	78.3M	8	SS	Semantic
DALES [Varney et al. 2020]	2020	ALS	PC	$10km^2$	1	505.3M	8	SS	Semantic
Campus3D [Li et al. 2020]	2020	UAV Ptgy	PC/Image	$1.58km^2$	1	937.1M	14	SS/IS	Hierarchical
Hessigheim 3D [Kölle et al. 2021]	2021	UAV LiDAR/Camera	PC/Mesh	$0.19km^2$	1	125.7M/36.76M	11	SS	Semantic
SUM [Gao et al. 2021]	2021	Airplane Camera	Mesh	$4km^2$	1	19M	6	SS	Semantic
Swiss3DCities [Can et al. 2021]	2021	UAV Ptgy	PC	$2.7km^2$	3	226M	5	SS	Semantic
SensatUrban [Hu et al. 2022]	2022	UAV Ptgy	PC	$7.46km^2$	1	2847M	13	SS	Semantic
STPLS3D [Chen et al. 2022a]	2022	UAV Ptgy	PC/Mesh	$1.27km^2$	1	150.4M	6	SS	Semantic
InstanceBuilding [Chen et al. 2022b]	2022	UAV Ptgy	Mesh/Image	$0.434km^2$	1	7.46M	2	IS	Instance
UrbanBIS (Ours)	2023	UAV Ptgy	PC/Mesh/Image	$10.78km^2$	5	2523.8M/284.3M	7/8 (Sub-cat.)	SS/IS/IR/SC	Semantic/Instance

Table 5: The correlations among all the scenes in UrbanBIS.

	Qingdao	Wuhu	Longhua	Yuehai	Lihu	Yingrenshi
Qingdao	1	0.89	0.68	0.50	0.26	0.65
Wuhu	0.89	1	0.47	0.34	-0.05	0.66
Longhua	0.68	0.47	1	0.96	0.85	0.56
Yuehai	0.50	0.34	0.96	1	0.88	0.53
Lihu	0.26	-0.05	0.85	0.88	1	0.18
Yingrenshi	0.65	0.66	0.56	0.53	0.18	1

Qingdao has more commercial and transportation buildings, showing that Qingdao, as a larger city, is at a higher city development level. Current urban datasets often ignore differences across scenes. This may lead to models trained on specific scenes being difficult to generalize. UrbanBIS includes a wider range of scenes and we analyze the similarities between scenes based on the number of buildings in different semantic categories. Table 5 reports the correlation coefficients between scenes, revealing significant differences in their semantic categories, except for Yuehai and Lihu, which are both campus scenes. Researchers can choose to employ scenes that are more similar to the intended use in their studies.

UrbanBIS is a real-world dataset, providing various semantic categories. Therefore, when designing segmentation models based on this data, the 'long-tail' problem that the data may bring needs to be considered. Figure 3 shows the number of point clouds for each semantic category in UrbanBIS, indicating that the three larger categories account for the vast majority, and current methods still face certain problems in recognizing smaller categories.

Compared with existing 3D point cloud datasets, UrbanBIS is the largest real urban dataset that provides both semantic and instance annotations. As Table 4 shows, UrbanBIS covers an area of 10.78 km^2, larger than the latest urban datasets SensatUrban [Hu et al. 2022] and STPLS3D [Chen et al. 2022a]. Importantly, UrbanBIS is currently the only large-scale real urban dataset that provides instance annotations and fine-grained building categories. Besides, it has the richest variety of scenes (six real urban scenes), including

three large scenes in different cities (Qingdao, Wuhu, and Longhua), two campus scenes (Yuehai and Lihu), and one small residential area (Yingrenshi). Hence, UrbanBIS facilitates not only semantic segmentation and instance segmentation but also many high-level tasks and applications, e.g., image reconstruction and scene completion.

Further, from Table 3 (right), we can see that UrbanBIS contains buildings of various heights. An intriguing observation is that even though different scenes have different compositions of building categories, they have similar distributions of building heights (if the scene area is sufficiently large), except for Yingrenshi, which is a small scene. So, we can take Yingrenshi as a test scene in UrbanBIS to explore the domain gap between the train and test samples, as well as the model generalization. Note also that the supplementary material presents more analysis, e.g., building density. See our accompanying video for a vivid understanding of UrbanBIS.

4 B-SEG METHOD

Existing mainstream methods [Chen et al. 2021; Jiang et al. 2020; Liang et al. 2021; Vu et al. 2022] commonly use clustering algorithms to group foreground points into instance proposals. Our B-Seg is *clustering-free*, enabling higher computational efficiency on processing large-scale urban-level data. As depicted in Figure 2, B-Seg has three components: (i) backbone network and feature learning for point-wise feature extraction and sub-task predictions; (ii) building proposal generation to produce and form preliminary building proposals; and (iii) building scoring to evaluate the quality of building proposals and filter out the errors.

(i) Backbone network and feature learning. We use a submanifold sparse convolutional network [Graham et al. 2018] as our backbone due to its strong capability of extracting features from 3D point clouds. Then, with the extracted point-wise features, we construct three parallel branches for three different sub-tasks: the semantic branch for semantic segmentation, the center-offset branch for predicting the offset of each point to its associated center, and the instance-aware branch for enhancing the instance segmentation result. Please refer to the supplementary material for details.
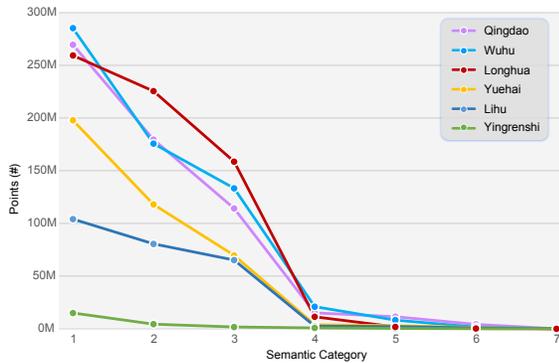
Figure 3: The long tail statistics of different scenes in UrbanBIS. The X-axis represents the 7 semantic categories defined in UrbanBIS in numeric form. They are arranged in descending order of the number of point clouds in each scene separately.

(ii) Building proposal generation. Next, we generate instance proposals for buildings with the following three modules: (1) *Candidate selection*: First, we employ furthest point sampling [Eldar et al. 1997] to randomly select foreground points as candidates for forming building proposals. In our experiments, we dynamically select one candidate point for every 3,000 foreground points, with a maximum of 100 per block. (2) *Building grouping*: We construct a relation matrix, in which the element at i-th row, j-th column denotes the learned feature distance between the i-th foreground point and j-th candidate point; the lower the feature distance, the higher the probability that the two points belong to the same building. We then use the *argmin* function to obtain the building proposal label for each foreground point. (3) *Building Merging*: Since there are still a large number of candidate points, the same building instance may be covered by multiple candidates. So, we offset each candidate point towards its predicted instance center and then merge candidate points and instance proposals for the same building.

(iii) Building scoring. The previous module may mistakenly produce some wrong proposals, which may affect the quality of the final instance predictions. So, we adopt ScoreNet [Jiang et al. 2020] to further predict a score for each proposal. In detail, we pass each proposal into a tiny 3D sparse convolution U-Net with an instance-aware pooling and a softmax layer to predict a score and filter out building proposals with scores less than 0.1.

In the end, we train B-Seg framework end-to-end from scratch with an overall optimization objective with four loss terms (see Figure 2): $\mathcal{L}_{sem}$ for semantic segmentation, $\mathcal{L}_{offset}$ for optimizing the center-offset vector prediction, $\mathcal{L}_{dis}$ for learning the instance-aware features, and $\mathcal{L}_{score}$ for learning the instance-proposal scores.

5 EXPERIMENTAL RESULTS

5.1 Experiment settings

5.1.1 Implementation details. We use one Quadro RTX 6000 GPU for model training with a batch size of 2 for a single scene and 4 for joint scenes, and the Adam optimizer [Kingma and Ba 2015] with

an initial learning rate of 0.001 for 400 epochs. For a fair comparison, we follow the original experimental settings of each method and modify only the hyperparameters to suit our dataset, e.g., the clustering grouping radius for [Chen et al. 2021; Jiang et al. 2020; Vu et al. 2022]. For the 3D sparse convolution, we set the voxel size as $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} m^3$. In the training, to balance the GPU memory limit and data block size, we set the maximum number of points as 500,000 and randomly adjust the input size by cropping a block if its size exceeds the maximum, similar to [Jiang et al. 2020]. At the inference stage, we directly input the whole block into the network.

5.1.2 Evaluations metrics. We employ Average Precision (AP, AP_{25}, AP_{50}) to explore the building instance segmentation. AP_{25} and AP_{50} are the AP scores with an Intersection-over-Union (IoU) threshold of 25% and 50%, respectively. AP is the mean Average Precision score from the IoU threshold of 25% to 95% with a step of 5%. Besides, we use mean Intersection-over-Union (mIoU) to evaluate the instance segmentation performance on the fine-grained building categories.

5.2 Experimental results

We evaluate our UrbanBIS dataset and the B-Seg method using some mainstream and representative 3D instance segmentation methods. From Sec. 3.3, we can see the large variations across different scenes in the data, so we set up different training settings to explore the method's effectiveness for different situations.

5.2.1 Building instance segmentation performance. First, both the train and test sets come from the same scene. As reported in Table 6, we can see that B-Seg achieves the best performance over the SOTA 3D point cloud segmentation methods *consistently for all the metrics in all the scenes*. Also, B-Seg has a faster processing speed than others; see the T columns. The other methods [Chen et al. 2021; He et al. 2021; Jiang et al. 2020; Vu et al. 2022] are clustering-based, thus requiring a time-consuming point-wise grouping process. For the very recent method DKNet [Wu et al. 2022], though it is clustering-free, it needs to encode every instance proposal into a kernel feature for the dynamic convolution. While DyCo3D [He et al. 2021] also adopts a dynamic-convolution-based model, it still requires a point-wise grouping first. On the contrary, B-Seg adopts a simple but effective strategy by aggregating points with a relation matrix and merging points only through a small set of candidates. The visualization result can be seen in Figure 6.

5.2.2 Fine-grained building instance segmentation performance. We further evaluate the segmentation performance of all methods on the building sub-categories. As Table 7 shows, B-Seg performs favorably over other methods for most sub-categories, except that it performs slightly worse than PointGroup [Jiang et al. 2020] on three categories in the Campus. These categories are overall harder to distinguish and their features are more similar in Campus, so requiring a stronger feature extractor. Note also that we do not have results for some categories in Wuhu and Longhua (see "-" in the table), as the number of associated buildings in these cases is either too small or not existent.

From the experiments, we can see the strength of B-Seg in analyzing buildings in 3D urban scenes and its potential for supporting building-aware 3D machine learning tasks.

Table 6: Benchmark results. Comparing the 3D point cloud instance segmentation performance of our method with state-of-the-art methods on the train and test sets of UrbanBIS. All methods are trained and tested under the same scene. Campus combines Yuehai and Lihu. T stands for the time cost for inference.

Method	Qingdao				Wuhu				Longhua				Campus			
	AP	AP50	AP25	T (s)	AP	AP50	AP25	T (s)	AP	AP50	AP25	T (s)	AP	AP50	AP25	T (s)
PointGroup [Jiang et al. 2020]	0.364	0.512	0.578	9.80	0.502	0.662	0.748	5.90	0.318	0.443	0.556	5.73	0.117	0.235	0.455	3.65
HAIS [Chen et al. 2021]	0.320	0.465	0.506	7.11	0.383	0.616	0.711	3.62	0.159	0.249	0.350	3.17	0.002	0.012	0.146	3.26
SoftGroup [Vu et al. 2022]	0.383	0.446	0.487	6.55	0.536	0.649	0.721	3.61	0.151	0.199	0.300	3.06	0.253	0.364	0.439	2.16
DyCo3D [He et al. 2021]	0.285	0.376	0.498	5.20	0.470	0.620	0.732	3.04	0.020	0.045	0.196	1.77	0.029	0.063	0.180	1.67
DKNet [Wu et al. 2022]	0.383	0.434	0.474	2.15	0.474	0.575	0.650	1.20	0.077	0.154	0.253	1.78	0.044	0.109	0.251	0.88
B-Seg (Ours)	**0.453**	**0.550**	**0.672**	**1.19**	**0.549**	**0.674**	**0.767**	**0.99**	**0.402**	**0.513**	**0.618**	**1.16**	**0.261**	**0.386**	**0.535**	**0.74**

Table 7: 3D instance segmentation performance of various methods on the fine-grained building sub-categories in UrbanBIS. Co, Re, Of, Cu, Tr, Mu, Te, L, H, and SH stand for Commercial, Residential, Office, Cultural, Transportation, Municipal, Temporary, Low-rise, High-rise, and Super High-rise, respectively.

Scene	Method	Co	Re	Of	Cu	Tr	Mu	Te	L	H	SH
Qingdao	PointGroup [Jiang et al. 2020]	0.831	0.946	0.798	0.620	0.329	0.527	0.480	0.780	0.867	0.918
	HAIS [Chen et al. 2021]	0.913	0.981	0.945	0.662	0.621	0.510	0.329	0.872	0.920	0.970
	SoftGroup [Vu et al. 2022]	0.823	0.94	0.71	0.479	0.645	0.413	0.242	0.423	0.856	0.859
	DyCo3D [He et al. 2021]	0.794	0.909	0.691	0.502	0.224	0.377	0.305	0.699	0.814	0.912
	DKNet [Wu et al. 2022]	0.933	0.973	0.884	0.700	**0.901**	0.633	**0.658**	0.879	0.945	0.962
	B-Seg (Ours)	**0.972**	**0.988**	**0.967**	**0.905**	0.735	**0.789**	0.580	**0.937**	**0.976**	**0.985**
Wuhu	PointGroup [Jiang et al. 2020]	**0.926**	0.970	0.958	0.928	-	0.746	0.655	**0.959**	0.950	0.979
	HAIS [Chen et al. 2021]	0.717	0.898	0.876	0.613	-	0.105	0.330	0.840	0.813	0.978
	SoftGroup [Vu et al. 2022]	0.843	0.944	0.929	0.862	-	0.479	0.504	0.915	0.917	0.935
	DyCo3D [He et al. 2021]	0.811	0.908	0.906	0.648	-	0.347	0.419	0.869	0.863	0.959
	DKNet [Wu et al. 2022]	0.919	0.972	**0.968**	0.894	-	0.675	0.610	0.956	**0.953**	0.985
	B-Seg (Ours)	0.882	**0.973**	0.955	**0.934**	-	**0.785**	**0.688**	0.958	0.930	**0.990**
Campus	PointGroup [Jiang et al. 2020]	0.891	0.939	0.960	-	**0.568**	**0.663**	**0.819**	0.895	0.959	0.959
	HAIS [Chen et al. 2021]	0.852	0.854	0.885	-	0.068	0.231	0.545	0.715	0.917	0.904
	SoftGroup [Vu et al. 2022]	0.813	0.816	0.841	-	0.061	0.134	0.521	0.897	0.939	0.949
	DyCo3D [He et al. 2021]	0.631	0.600	0.749	-	0.081	0.222	0.476	0.436	0.723	0.845
	DKNet [Wu et al. 2022]	0.942	0.919	0.920	-	0.418	0.599	0.805	0.890	0.942	0.896
	B-Seg (Ours)	**0.962**	**0.957**	**0.969**	-	0.498	0.588	0.794	**0.910**	**0.970**	**0.968**
Longhua	PointGroup [Jiang et al. 2020]	0.868	0.844	0.811	0.745	0.766	0.699	0.696	0.764	0.828	0.374
	HAIS [Chen et al. 2021]	0.830	0.870	0.914	0.823	0.090	0.234	0.247	0.529	0.918	**0.957**
	SoftGroup [Vu et al. 2022]	0.697	0.847	0.798	0.768	0.216	0.329	0.416	0.846	0.955	0.948
	DyCo3D [He et al. 2021]	0.227	0.815	0.791	0.663	0.142	0.362	0.253	0.375	0.798	0.829
	DKNet [Wu et al. 2022]	0.456	0.847	0.812	0.799	0.872	0.749	0.760	0.621	0.831	0.925
	B-Seg (Ours)	**0.991**	**0.939**	**0.984**	**0.951**	**0.902**	**0.789**	**0.848**	**0.922**	**0.979**	0.924

Table 8: 3D instance segmentation performance of various methods for different cross-scene training/testing settings.

Method	Train: Qingdao + Wuhu; Test: Longhua			Train: Campus; Test: Qingdao + Wuhu			Train: Longhua; Test: Yingrenshi		
	AP	AP50	AP25	AP	AP50	AP25	AP	AP50	AP25
PointGroup [Jiang et al. 2020]	0.300	0.482	0.618	0.243	0.374	0.514	0.558	0.660	0.722
Hais [Chen et al. 2021]	0.158	0.265	0.380	0.367	0.493	0.568	0.427	0.530	0.671
SoftGroup [Vu et al. 2022]	0.121	0.198	0.310	**0.391**	0.472	0.540	0.439	0.535	0.566
DyCo3D [He et al. 2021]	0.037	0.082	0.356	0.009	0.035	0.244	0.019	0.140	0.411
DKNet [Wu et al. 2022]	0.139	0.208	0.286	0.075	0.136	0.212	0.297	0.389	0.389
B-Seg (Ours)	**0.323**	**0.486**	**0.622**	0.353	**0.507**	**0.615**	**0.621**	**0.700**	**0.739**

5.2.3 Joint-training on all scenes. We train the methods collectively on samples of all scenes. As Table 9 shows, B-Seg still achieves satisfactory results in average performance with the fastest speed.

5.2.4 Cross-scene building instance segmentation performance. Next, we explore the model generalization capability of various methods by training each method on some scenes and testing its trained model on others. As Table 8 shows, B-Seg is able to achieve better cross-scene 3D instance segmentation performance than the SOTA methods for the different combinations of train and test scenes. Overall, the experiment demonstrates the strong generalization capability of B-Seg. Such a capability is particularly important for handling *unseen* urban scenes.

Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang

Table 9: Joint training on all scenes in UrbanBIS. We combine all the training and validation samples to train all the methods below, and employ all the test samples in the evaluation. The performance for fine-grained buildings can be found in supplementary material.

Method	AP	AP50	AP25	Time (s)
PointGroup [Jiang et al. 2020]	0.377	0.549	0.664	6.918
HAIS [Chen et al. 2021]	0.373	0.515	0.587	4.500
SoftGroup [Vu et al. 2022]	**0.402**	0.490	0.560	4.389
DyCo3D [He et al. 2021]	0.129	0.246	0.487	2.386
DKNet [Wu et al. 2022]	0.271	0.348	0.408	1.059
B-Seg (Ours)	0.401	**0.551**	**0.671**	**0.969**

Table 10: Experimental results on the InstanceBuilding dataset. Note that B-Seg does not need images and uses only 3D data without color.

Scene	[Chen et al. 2022b] (RBG)			[Chen et al. 2022b] (RGBH)			B-Seg (Ours)		
	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
Scene1	0.648	0.832	0.665	0.713	0.891	0.773	0.561	0.725	0.594
Scene2	0.571	0.748	0.597	0.665	0.840	0.681	0.764	0.930	0.782
Scene3	0.636	0.888	0.693	0.667	0.929	0.717	0.599	0.820	0.625
Scene4	0.653	0.857	0.699	0.673	0.876	0.711	0.519	0.739	0.491

5.2.5 Test on InstanceBuilding. To further explore B-Seg, we conducted experiments on InstanBuilding [Chen et al. 2022b], a triangle mesh dataset with building instance annotations primarily for segmenting adjacent buildings. Table 10 reports the results, in which the results of [Chen et al. 2022b] are directly copied from original paper of InstanceBuilding. In terms of AP, B-Seg does not show obvious disadvantages and even outperforms other methods on Scene 2. Yet, B-Seg does not use any relevant image data or depth information at all, showing that B-Seg does have advantages in dealing with unknown 3D scenes and can achieve good performance.

It can be seen that B-Seg has advantages in dealing with the instance segmentation in urban scenes from the above results, including faster processing speed, better segmentation performance and more robust generalization. This is because B-Seg does not rely on the training set for the clustering makes it less prone to overfitting and ensures its generalizability in different urban scenes.

6   DISCUSSION AND CONCLUSION

In this paper, we presented UrbanBIS, a large-scale 3D urban dataset, providing six real-world urban scenes of 2.5 billion annotated point samples over a vast area of $10.78km^2$. Beyond the existing datasets, it is a large-scale 3D real-world urban dataset, which is multifunctional with multiple data formats: dense semantic annotations on the point clouds and meshes, fine-grained instance and semantic segmentation on the buildings, and high-quality 3D reconstruction models, as well as high-resolution aerial-acquisition photos. UrbanBIS comprises scenes of different urban styles and building compositions, so different portions of the dataset can be employed for different purposes of study. Further, we develop B-Seg, an end-to-end framework to segment the large point clouds in UrbanBIS. Compared with mainstream methods, B-Seg demonstrates better accuracy and faster inference speed on UrbanBIS. In the future, we shall release UrbanBIS and B-Seg.

REFERENCES

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *arXiv:1702.01105* (2017).

Mehmet Aygun, Aljosa Osep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, and Laura Leal-Taixe. 2021. 4D Panoptic LiDAR Segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 5527–5537.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. Int. Conf. on Computer Vision*. 9297–9307.

F Biljecki and Y Dehbi. 2019. Raise the roof: towards generating LoD2 models without aerial surveys using machine learning. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (2019), 27–34.

Aleksey Boyko and Thomas Funkhouser. 2014. Cheaper by the Dozen: Group Annotation of 3D Data. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, 33−−42.

Gülcan Can, Dario Mantegazza, Gabriele Abbate, Sébastien Chappuis, and Alessandro Giusti. 2021. Semantic segmentation on Swiss3DCities: A benchmark study on aerial photogrammetric 3D pointcloud dataset. *Pattern Recognition Letters* 150 (2021), 108–114.

Jiazhou Chen, Yanghui Xu, Shufang Lu, Ronghua Liang, and Liangliang Nan. 2022b. 3-D Instance Segmentation of MVS Buildings. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.

Meida Chen, Qingyong Hu, Zifan Yu, Hugues THOMAS, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. 2022a. STPLS3D: A Large-Scale Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset. In *Proc. British Machine Vision Conf.*

Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. 2021. Hierarchical Aggregation for 3D Instance Segmentation. In *Proc. Int. Conf. on Computer Vision*. 15467–15476.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 5828–5839.

Department of Engineering Quality and Safety Supervision, Ministry of Housing and Urban-Rural Development. 2009. *National Technical Measures for Design of Civil Construction*. China Institute of Building Standard Design & Research Center.

Y. Eldar, M. Lindenbaum, M. Porat, and Y.Y. Zeevi. 1997. The farthest point strategy for progressive image sampling. *IEEE Trans. on Image Processing* 6, 9 (1997), 1305–1315.

Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 2020. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 9028–9037.

Weixiao Gao, Liangliang Nan, Bas Boom, and Hugo Ledoux. 2021. SUM: A benchmark dataset of Semantic Urban Meshes. *ISPRS J. Photogrammetry and Remote Sensing* 179 (2021), 108–120.

A Geiger, P Lenz, C Stiller, and R Urtasun. 2013. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Research* 32, 11 (2013), 1231–1237.

Aleksey Golovinskiy, Vladimir G. Kim, and Thomas Funkhouser. 2009. Shape-based recognition of 3D point clouds in urban environments. In *Proc. Int. Conf. on Computer Vision*. 2154–2161.

Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 9224–9232.

David Griffiths and Jan Boehm. 2019. SynthCity: A large-scale synthetic point cloud. In *arXiv:1907.04758*.

Timo Hackel, N. Savinov, L. Ladicky, Jan D. Wegner, K. Schindler, and M. Pollefeys. 2017. SEMANTIC3D.NET: A New Large-Scale Point Cloud Classification Benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 91–98.

Lei Han, Tian Zheng, Lan Xu, and Lu Fang. 2020. OccuSeg: Occupancy-Aware 3D Instance Segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 2937–2946.

Tong He, Chunhua Shen, and Anton van den Hengel. 2021. DyCo3D: Robust Instance Segmentation of 3D Point Clouds through Dynamic Convolution. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 354–363.

Ji Hou, Angela Dai, and Matthias Nießner. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 4416–4425.

Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. 2022. Sensaturban: Learning Semantics from Urban-Scale Photogrammetric Point Clouds. *Int. J. Computer Vision* 130, 2 (2022), 316–343.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2020. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 4867–4876.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: a Method for Stochastic Optimization. In *Proc. Int. Conf. on Learning Representations*.

Michael Kölle, Dominik Laupheimer, Stefan Schmohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. 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 J. Photogrammetry and Remote Sensing* 1 (2021), 100001.

Florent Lafarge and Clément Mallet. 2012. Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. *Int. J. Computer Vision* 99 (2012), 69–85.

Jean Lahoud, Bernard Ghanem, Martin R. Oswald, and Marc Pollefeys. 2019. 3D Instance Segmentation via Multi-Task Metric Learning. In *Proc. Int. Conf. on Computer Vision*. 9255–9265.

Xinke Li, Chongshou Li, Zekun Tong, Andrew Lim, Junsong Yuan, Yuwei Wu, Jing Tang, and Raymond Huang. 2020. Campus3D: A Photogrammetry Point Cloud Benchmark for Hierarchical Understanding of Outdoor Scene. In *Proc. ACM Int. Conf. on Multimedia*. 238––246.

Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. 2021. Instance Segmentation in 3D Scenes Using Semantic Superpoint Tree Networks. In *Proc. Int. Conf. on Computer Vision*. 2783–2792.

Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. 2022. Capturing, Reconstructing, and Simulating: the UrbanScene3D Dataset. In *Proc. Euro. Conf. on Computer Vision*. 93–109.

Shih-Hung Liu, Shang-Yi Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. 2020. Learning Gaussian Instance Segmentation in Point Clouds. In *arXiv:2007.09860*.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. 2019. PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 909–918.

Daniel Munoz, J. Andrew Bagnell, Nicolas Vandapel, and Martial Hebert. 2009. Contextual classification with functional Max-Margin Markov Networks. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 975–982.

Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. 2019. JSIS3D: Joint Semantic-Instance Segmentation of 3D Point Clouds With Multi-Task Pointwise Networks and Multi-Value Conditional Random Fields. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 8819–8828.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *Proc. Int. Conf. on Computer Vision*. 10912–10922.

Xavier Roynard, Jean-Emmanuel Deschaud, and François Goulette. 2018. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *Int. J. Robotics Research* 37, 6 (2018), 545–557.

Dávid Rozenberszki, Or Litany, and Angela Dai. 2022. Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In *Proc. Euro. Conf. on Computer Vision*, Vol. 13693. 125–141.

Pratheba Selvaraju, Mohamed Nabail, Marios Loizou, Maria Maslioukova, Melinos Averkiou, Andreas Andreou, Siddhartha Chaudhuri, and Evangelos Kalogerakis. 2021. BuildingNet: Learning To Label 3D Buildings. In *Proc. Int. Conf. on Computer Vision*. 10397–10407.

Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. 2020. Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition Workshops*. 797–806.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data. In *Proc. Int. Conf. on Computer Vision*. 1588–1597.

Nina Varney, Vijayan K. Asari, and Quinn Graehling. 2020. DALES: A Large-Scale Aerial LiDAR Data Set for Semantic Segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 717–726.

Thang Vu, Kookhoi Kim, Tung M. Luu, Thanh Nguyen, and Chang D. Yoo. 2022. SoftGroup for 3D Instance Segmentation on Point Clouds. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 2708–2717.

Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. 2018. SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 2569–2578.

Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. 2019a. Associatively Segmenting Instances and Semantics in Point Clouds. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 4091–4100.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2019b. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. on Graphics* 38, 5 (2019), 146:1–146:12.

Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2022. 3D Instances as 1D Kernels. In *Proc. Euro. Conf. on Computer Vision*. 235––252.

Jianxiong Xiao, Andrew Owens, and Antonio Torralba. 2013. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels. In *Proc. Int. Conf. on Computer Vision*. 1625–1632.

Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. 2019. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Proc. Conf. on Neural Information Processing Systems*. 6737–6746.

Zhihang Yao, Claus Nagel, Felix Kunde, György Hudra, Philipp Willkomm, Andreas Donaubauer, Thomas Adolphi, and Thomas H Kolbe. 2018. 3DCityDB-a 3D geodatabase solution for the management, analysis, and visualization of semantic 3D city models based on CityGML. *Open Geospatial Data, Software and Standards* 3, 1 (2018), 1–26.

Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. 2019. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*. 3947–3956.

Lin Zhao and Wenbing Tao. 2020. JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds. In *Proc. AAAI Conf. on Artificial Intelligence*. 12951–12958.

Dingfu Zhou, Jin Fang, Xibin Song, Liu Liu, Junbo Yin, Yuchao Dai, Hongdong Li, and Ruigang Yang. 2020a. Joint 3D Instance Segmentation and Object Detection for Autonomous Driving. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*.

Xiaohui Zhou, Ke Xie, Kai Huang, Yilin Liu, Yang Zhou, Minglun Gong, and Hui Huang. 2020b. Offsite Aerial Path Planning for Efficient Urban Scene Reconstruction. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 39, 6 (2020), 192:1–192:16.

Iman Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Aljosa Smolic, Rogerio Da Silva, and Morteza Rahbar. 2019. DublinCity: Annotated LiDAR Point Cloud and its Applications. In *Proc. British Machine Vision Conf.*

Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang

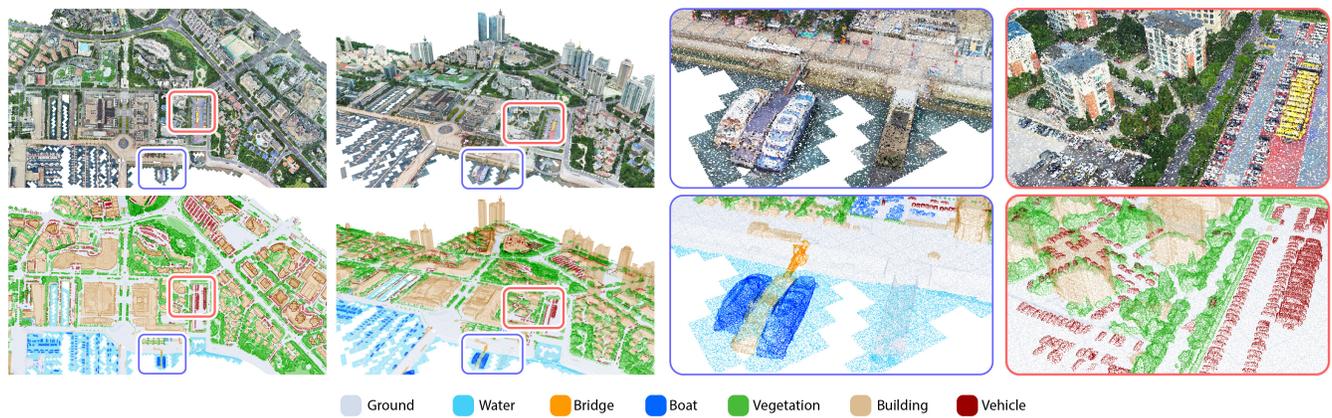Ground   Water   Bridge   Boat   Vegetation   Building   Vehicle

Figure 4: The examples of the semantic labeling of different blocks in UrbanBIS. The color legend is shown at the bottom. See more in the supplementary material.

Commercial   Residential   Office   Cultural   Transportation   Municipal   Temporary

Figure 5: UrbanBIS provides fine-grained building-level information, including segmentation information on building instances and semantics information about building categories (from left to right): Commercial, Residential, Office, Cultural, Transportation, Municipal and Temporary.
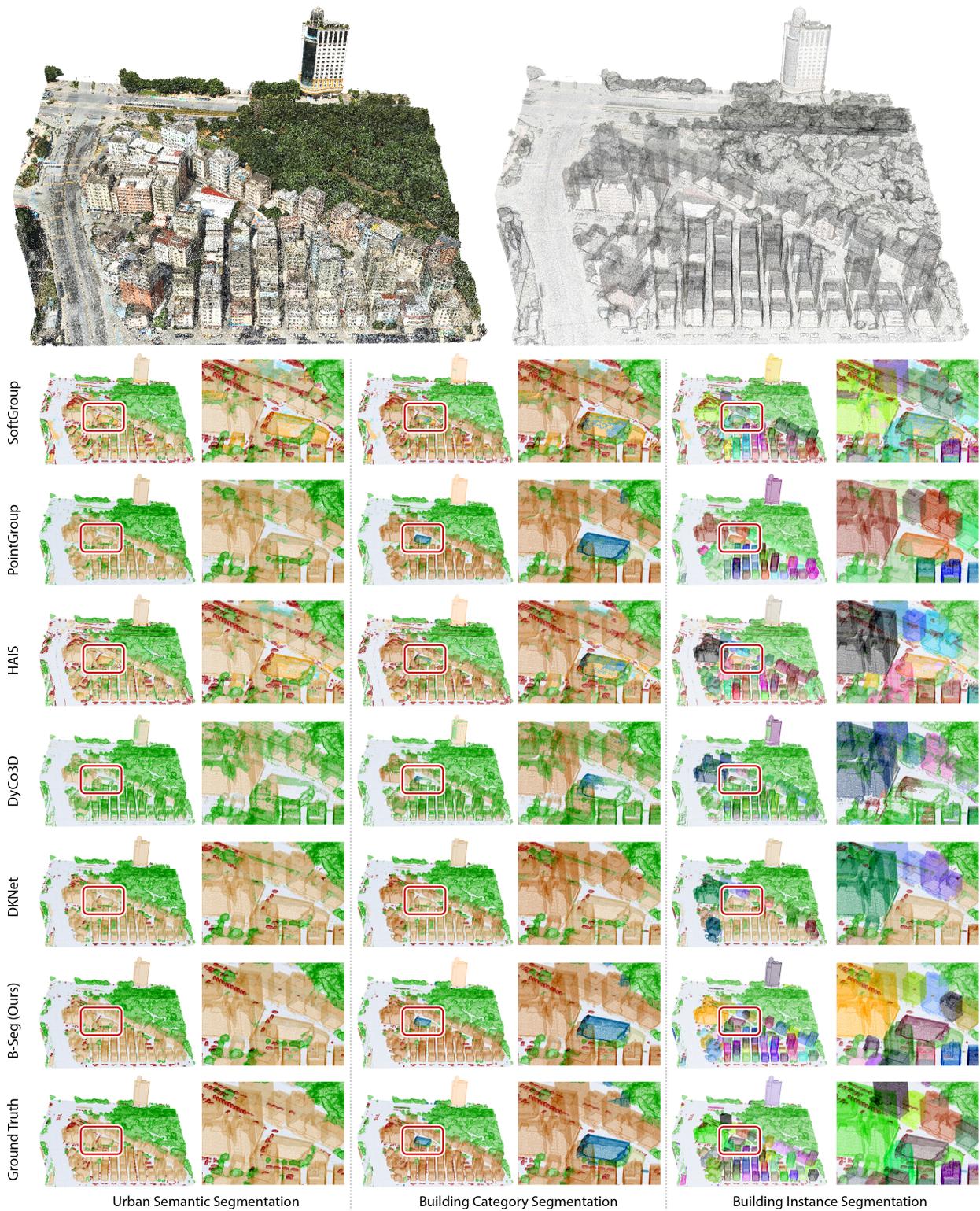
Figure 6: Qualitative results of B-Seg and comparison methods on a test block in the Longhua scene. From left to right is the urban semantic segmentation, building category segmentation, and building instance segmentation. The prediction of B-Seg shows more accurate segmentation results in this dense scene.