Towards Semantic Segmentation of Urban-Scale 3D Point Clouds:
A Dataset, Benchmarks and Challenges

Qingyong Hu[1], Bo Yang[1,2*], Sheikh Khalid[3], Wen Xiao[4], Niki Trigoni[1], Andrew Markham[1]

[1]University of Oxford, [2]The Hong Kong Polytechnic University, [3]Sensat Ltd, [4]Newcastle University

qingyong.hu@cs.ox.ac.uk, bo.yang@polyu.edu.hk, wen.xiao@ncl.ac.uk, andrew.markham@cs.ox.ac.uk

Abstract

An essential prerequisite for unleashing the potential of supervised deep learning algorithms in the area of 3D scene understanding is the availability of large-scale and richly annotated datasets. However, publicly available datasets are either in relative small spatial scales or have limited semantic annotations due to the expensive cost of data acquisition and data annotation, which severely limits the development of fine-grained semantic understanding in the context of 3D point clouds. In this paper, we present an urban-scale photogrammetric point cloud dataset with nearly three billion richly annotated points, which is three times the number of labeled points than the existing largest photogrammetric point cloud dataset. Our dataset consists of large areas from three UK cities, covering about 7.6 km^2 of the city landscape. In the dataset, each 3D point is labeled as one of 13 semantic classes. We extensively evaluate the performance of state-of-the-art algorithms on our dataset and provide a comprehensive analysis of the results. In particular, we identify several key challenges towards urban-scale point cloud understanding. The dataset is available at https://github.com/QingyongHu/SensatUrban.

1. Introduction

The three-dimensional world around us is composed of a rich variety of objects: buildings, trees, cars, and so forth, each with distinct appearance, morphology, and function. Giving machines the ability to precisely segment and label these diverse objects is of key importance to allow them to interact competently within our physical world, for applications such as object-level robotic grasping [39], scene-level robot navigation [54] and autonomous driving [16], or even large-scale urban 3D modeling, which is critical for the future of smart city planning and management [11, 4].

The ongoing revolution in data-driven deep networks has
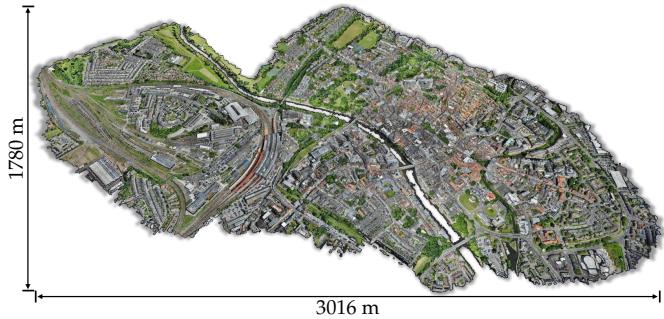
---

*Corresponding author

Figure 1: An urban-scale point cloud collected from a region on the perimeter of the city of York, UK. It covers a contiguous area of more than 3 square kilometer and represents a typical urban suburb.

led to a radical boost in the performance of 3D point cloud segmentation. A series of neural pipelines proposed to address the core problem of semantic segmentation, including: 1) 3D voxel-based methods such as SparseConvNet [19] and MinkowskiNet [10], 2) 2D projection-based approaches such as RangeNet++ [33] and SqueezeSeg [59], and 3) recent point-based architectures *e.g.* PointNet/PointNet++ [37, 38], KPConv [51] and RandLA-Net [23].

To a large degree, these techniques have been driven forward by the availability of open datasets which act as benchmarks for objective comparison of algorithms and their performance. These existing 3D repositories can be generally classified as 1) object-level 3D models such as ModelNet [60] and ShapeNet [8], 2) indoor scene-level 3D scans, *e.g.,* S3DIS [3], ScanNet [13], and SceneNN [70], and 3) outdoor roadway-level 3D point clouds including SemanticKITTI [5] and Semantic3D [21].

However, there remain a number of key open questions as to whether these techniques are capable of learning accurate semantics over urban-scale 3D point clouds. **Firstly**, unlike the existing datasets for objects, rooms or roadways which are usually less than $200m$ in scale, the urban-scale datasets are expected to be collected by aerial platforms, spanning over extremely wide areas. How to efficiently

and effectively preprocess massive points to feed into neural networks? **Secondly**, the real-world urban space is usually dominated by large-sized buildings or ground, and therefore the urban-scale datasets demonstrate extreme class imbalance - a majority of points fall into a few categories with sparse, yet important classes being under represented. How to overcome this data imbalance in neural networks? **Thirdly**, with the advancement of aerial mapping systems, the urban-scale point clouds can not only capture the depth information, but also true color for the scene appearance. (How) does color information, in addition to depth, aid in semantic segmentation of urban areas? **Lastly**, and potentially most importantly, how are the existing networks trained on one urban area able to generalize to a novel area?

To this end, we aim to establish a new paradigm for urban-scale 3D semantic segmentation, enabled by UAV photogrammetry. Our dataset, called **SensatUrban**, represents sub-sections of three large cities in the UK, *i.e.*, Birmingham, Cambridge, and York. It consists of nearly four billion 3D points covering more than 7.6 square kilometers urban area in these cities (as shown in Figure 1). The 3D point clouds are generated from high-quality aerial images captured by a professional-grade UAV mapping system. Details of data acquisition are presented in Section 3. We manually labeled each point in the Birmingham and Cambridge city as one of 13 semantic categories such as *ground*, *vegetation*, *car*, *etc.*. Compared with exiting 3D datasets, our SensatUrban is unique in two-fold.

- Unlike existing datasets for objects [60, 8], rooms [70, 3, 13] and roadways [21, 5] which are usually less than two hundred meters in scale, the SensatUrban point clouds continuously occupy kilometers in real-world urban areas, opening up new opportunities towards urban-scale applications such as smart cities, and large national infrastructure planning and management.
- Being reconstructed from high-resolution aerial images, our point clouds provide unique top-down and oblique perspectives for the entire landscape of cities. Inherently, the geometric patterns, textures, natural colours and distributions are distinct from the existing datasets.

On the basis of SensatUrban, we further identify a number of key challenges and empirically investigate them from various aspects in Section 5. In particular, we firstly study how the large-scale urban point clouds can be preprocessed, to adapt to existing approaches without losing segmentation accuracy. Secondly, we explore the necessity of colorful appearance for better semantic learning of several key categories, highlighting the advantage of photogrammetric point clouds over LiDAR-based point clouds. Thirdly, we examine the imbalance of semantic categories in the urban-scale scenarios. Lastly, the difficulty of cross-city semantic learning is analysed. Note that, this paper

does not aim to thoroughly tackle these challenges, but expose them to the community for future research.

Overall, our primary contributions are: 1) a unique urban-scale 3D dataset for semantic learning, and 2) an indepth study of generalizing existing algorithms to the large-scale urban point clouds and an outlook on future directions for 3D point cloud segmentation at massive scale and resolution. We aspire to highlight the challenges faced in the 3D semantic learning on large and dense point clouds of urban environments, sparking innovation in applications such as smart cities, digital twins, autonomous vehicles, automated asset management of large national infrastructures, and intelligent construction sites.

2. Related Work

2.1. Existing 3D Datasets

Existing 3D datasets can be broadly classified into four categories: **1) Object-level 3D models**. These include the synthetic ModelNet [60], ShapeNet [8], ShapePartNet [65], PartNet [34] and the real-world ScanObjectNN [53]. **2) Indoor scene-level 3D scans**. These datasets are usually collected by short-range depth scanners, such as NYU3D [46], SUN RGB-D [47], S3DIS [3], SceneNN [70] and ScanNet [13]. In addition, there are two synthetic datasets SceneNet [22] and SceneNet RGB-D [31], which covers large-scale complex indoor environments. **3) Outdoor roadway-level 3D point clouds.** The majority of these datesets are specifically collected for applications such as autonomous driving using a LiDAR scanner together with RGB cameras, such as the early Oakland [35], KITTI [17], Sydney Urban Objects [14] and the recent Semantic3D [21], Paris-Lille-3D [43], Argoverse [9], SemanticKITTI [5], SemanticPOSS [36], Toronto-3D [49], nuScenes [7], A2D2 [18], CSPC-Dataset [52], Lyft dataset [1] and Waymo dataset [48]. To obtain more accurate semantic labels, a number of synthetic datasets [40, 15] are generated by simulating street scenes. **4) Urban-level aerial 3D point clouds.** They are usually obtained by costly aerial LiDARs, such as the recent DublinCity [72], DALES [56], LASDU [64]. However, they are unable to capture true color information for the complex urban structures.

Being concurrent to our work, the recent Campus3D [26] also releases large-scale photogrammetric 3D point clouds generated from high-resolution aerial images. However, our SensatUrban is urban-scale and several times that of campus3d in terms of space size and labeling points.

2.2. 3D Semantic Learning

The wide availability of 3D datasets has facilitated rapid progress in semantic learning based on neural networks. In general, existing learning algorithms [20] can be divided into three pipelines, depending on how the 3D data is repre-

	#Name and Reference	#Year	#Spatial size[1]	#Classes[2]	#Points	#RGB	#Sensors
Object-level	ShapeNet [8]	2015	-	55	-	No	Synthetic
	PartNet [34]	2019	-	24	-	No	Synthetic
Indoor Scene-level	S3DIS [3]	2017	$6 \times 10^3 m^2$	13 (13)	273M	Yes	Matterport
	ScanNet [13]	2017	$1.13 \times 10^5 m^2$	20 (20)	242M	Yes	RGB-D
Outdoor Roadway-level	Paris-rue-Madame [45]	2014	$0.16 \times 10^3\ m$	17	20M	No	MLS
	IQmulus [55]	2015	$10 \times 10^3\ m$	8 (22)	300M	No	MLS
	Semantic3D [21]	2017	-	8 (9)	4000M	Yes	TLS
	Paris-Lille-3D [43]	2018	$1.94 \times 10^3\ m$	9 (50)	143M	No	MLS
	SemanticKITTI [5]	2019	$39.2 \times 10^3\ m$	25 (28)	4549M	No	MLS
	Toronto-3D [49]	2020	$1 \times 10^3\ m$	8 (9)	78.3M	Yes	MLS
Urban-level	ISPRS [42]	2012	-	9	1.2M	No	ALS
	DublinCity [72]	2019	$2 \times 10^6 m^2$	13	260M	No	ALS
	DALES [56]	2020	$10 \times 10^6 m^2$	8 (9)	505M	No	ALS
	LASDU [64]	2020	$1.02 \times 10^6 m^2$	5	3.12M	No	ALS
	Campus3D [26]	2020	$1.58 \times 10^6 m^2$	24	937.1M	Yes	UAV Photogrammetry
	SensatUrban (Ours)	2020	$7.64 \times 10^6 m^2$	13 (31)	2847M	Yes	UAV Photogrammetry

Table 1: Comparison with the representative datasets for segmentation of 3D point clouds. [1]The spatial size (Area/Length) in the dataset, m: meter, [2] The number of classes used for evaluation and the number of sub-classes annotated in brackets. MLS: Mobile Laser Scanning system, TLS: Terrestrial Laser Scanning system, ALS: Aerial Laser Scanning system.

sented: **1) Voxel-based approaches** [19, 10, 25, 29, 32, 67]. Although mature 3D CNN architectures can be easily applied, these techniques usually require significant computation and memory usage, thus not being easily scalable to urban-scale point clouds. **2) 2D projection-based methods** [33, 30, 12, 62]. Similarly, these pipelines leverage the well-developed 2D CNN frameworks to learn 3D semantics after projecting the point clouds onto 2D images. However, critical geometric information is very likely to be lost in the projection step, and therefore is not suitable for learning the relatively small object categories within urban-scale scenarios. **3) Point-based architectures** [37, 38, 27, 50, 58, 51, 23]. This class of techniques learns per-point semantics primarily based on the simple MLPs and typically achieves great results in 3D object detection [71] and instance segmentation [63]. Compared with both voxel and projection-based methods, these pipelines tend to be computationally efficient and have the potential to preserve the semantics for every single 3D point. However, most of the existing point-based methods are usually designed and tuned for small-scale point sets. It is still unclear how to effectively generalize the point-based methods to the more complex urban-scale scenarios. In this regard, we investigate a number of critical challenges in Section 5.

3. The SensatUrban Dataset

In this section we describe how we collect, process and label the dataset over three large urban areas in the UK.

3.1. Collecting Aerial Imagery

Due to the clear advantages of UAV imaging over similar mapping techniques, such as LiDAR, we use a cost

(b) Zoomed-in single flight survey
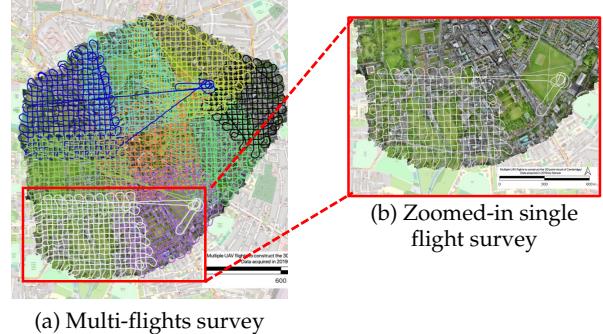
(a) Multi-flights survey

Figure 2: The survey of a region in Cambridge. All 9 flight plans (*left*) are collated together to cover the site. Lines with different colors represent different flight paths of UAVs. The circular path is the takeoff and landing pattern.

effective fixed wing drone, Ebee X[1], which is equipped with a cutting-edge SODA camera, to stably capture high-resolution aerial image sequences. In order to fully and evenly cover the survey area, all flight paths are pre-planned in a grid fashion and automated by the flight control system (e-Motion). Note that, the camera has the ability to take both oblique and nadir photographs, ensuring that vertical surfaces are captured appropriately. Since each flight lasts between 40-50 minutes due to limited battery capacity, multiple individual flights are executed in parallel to capture the whole area. These multiple aerial image sequences are then geo-referenced using a highly precise onboard Realtime Kinemtic (RTK) GNSS. Ground validation points which are measured by independent professional sur-

---

[1]https://www.sensefly.com/drone/ebee-x-fixed-wing-drone/

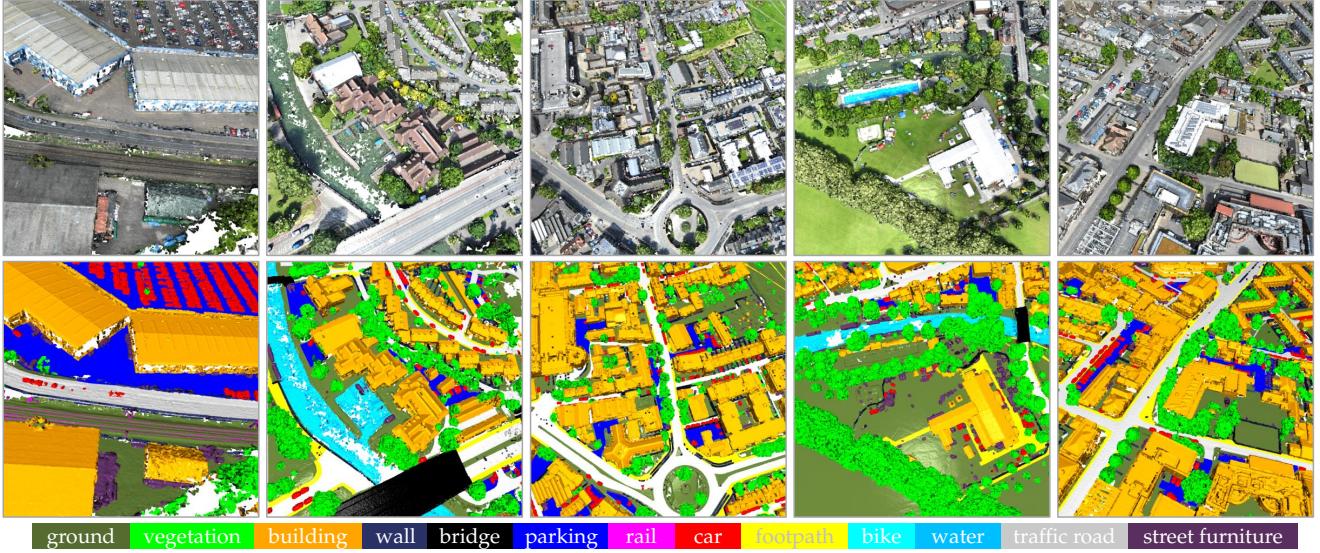| ground | vegetation | building | wall | bridge | parking | rail | car | footpath | bike | water | traffic road | street furniture |

Figure 3: Examples of our SensatUrban dataset. Different semantic classes are labeled by different colors.

veyors with high precision GNSS equipment are then used to assess the accuracy and quality of the data. For illustration, Figure 2 shows the paths of the pre-planed multiple flights to cover the selected area in the Cambridge city.

3.2. Reconstructing 3D Point Clouds

To reconstruct urban-scale 3D point clouds, we use off-the-shelf software such as Pix4D to reconstruct dense and coloured 3D point clouds from the captured aerial image sequences based on the principles of Structure from Motion (SfM) and dense image matching.

For the urban area on the periphery of **Birmingham**, we feed all the captured sequential images to Pix4D, generating 569,147,075 3D points in total, representing an area of 1.2 square kilometers. Similarly, we reconstruct 2,278,514,725 points for the urban region adjacent to the city of **Cambridge** with an area of approximately 3.2 square kilometers, and reconstruct 904,155,619 points for **York** with an area of approximately 3.2 square kilometers.

3.3. Annotating Semantic Labels

We define the semantic categories based on two criteria. 1) Each category should have a clear and unambiguous semantic meaning, and it should be of interest to social or commercial purposes, such as asset management, urban planning, and surveillance. 2) Different categories should have significant variance in terms of geometric structure or appearance. We identify the below 13 semantic classes to label all 3D points in the Birmingham and Cambridge via off-the-shelf point cloud labeling tools. The points in York are not labelled, but made available for possible pre-training in semi-supervised schemes. All labels have been manually

cross-checked, guaranteeing the consistency and high quality. It takes around 600 working hours to label the entire dataset. Figure 3 shows examples of our annotations. Table 1 compares the statistics of our SensatUrban with a number of existing 3D datasets.

1. *Ground*: including impervious surfaces, grass, terrain
2. *Vegetation*: including trees, shrubs, hedges, bushes
3. *Building*: including commercial / residential buildings
4. *Wall*: including fence, highway barriers, walls
5. *Bridge*: road bridges
6. *Parking*: parking lots
7. *Rail*: railroad tracks
8. *Traffic Road*: including main streets, highways
9. *Street Furniture*: including benches, poles, lights
10. *Car*: including cars, trucks, HGVs
11. *Footpath*: including walkway, alley
12. *Bike*: bikes / bicyclists
13. *Water*: rivers / water canals

4. Benchmarks

4.1. Statistics of Train/Val/Test Split

To setup the benchmark, we divide the point cloud of each area into similarly sized tiles similar to DALES [56], so to be suitable for training and testing on modern GPUs. In particular, the point cloud of the Birmingham urban area is divided into 14 tiles. We then select 10 tiles for training, 2 for validation and 2 for testing. Similarly, the Cambridge split has 29 tiles in total: 20 tiles for training, 5 for validation and 4 for testing. Each tile is approximately 400×400

	OA(%)	mAcc(%)	**mIoU(%)**	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet [37]	80.78	30.32	23.71	67.96	89.52	80.05	0.00	0.00	3.95	0.00	31.55	0.00	35.14	0.00	0.00	0.00
PointNet++ [38]	84.30	39.97	32.92	72.46	94.24	84.77	2.72	2.09	25.79	0.00	31.54	11.42	38.84	7.12	0.00	56.93
TagentConv [50]	76.97	43.71	33.30	71.54	91.38	75.90	35.22	0.00	45.34	0.00	26.69	19.24	67.58	0.01	0.00	0.00
SPGraph [24]	85.27	44.39	37.29	69.93	94.55	88.87	32.83	12.58	15.77	15.48	30.63	22.96	56.42	0.54	0.00	44.24
SparseConv [19]	88.66	63.28	42.66	74.10	97.90	94.20	63.30	7.50	24.20	0.00	30.10	34.00	74.40	0.00	0.00	54.80
KPConv [51]	**93.20**	63.76	**57.58**	**87.10**	**98.91**	**95.33**	**74.40**	28.69	41.38	0.00	55.99	**54.43**	**85.67**	40.39	0.00	**86.30**
RandLA-Net [23]	89.78	**69.64**	52.69	80.11	98.07	91.58	48.88	**40.75**	**51.62**	0.00	**56.67**	33.23	80.14	32.63	0.00	71.31

Table 2: Benchmark results of the baselines on our SensatUrban. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) scores are reported.
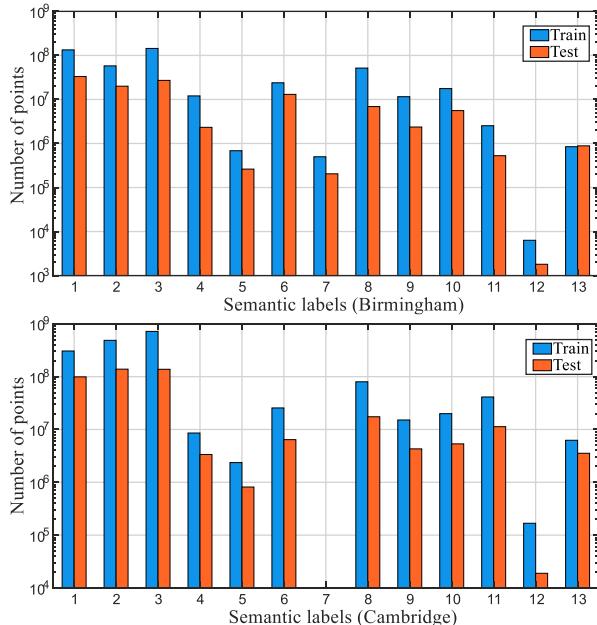

Figure 4: The distribution of different semantic categories in our SensatUrban dataset. Note that, there are no points annotated as *rail* in Cambridge. Also note the logarithmic scale for the vertical axis.

square meters. As shown in Figure 4, we show the total number of 3D points for each semantic category in the training/testing tiles in both Birmingham and Cambridge. It can be seen that the major semantic categories, *i.e.*, *ground / building / vegetation*, together comprise more than 50% of the total points, whereas the minor yet important categories (*e.g.*, *bike / rail*) only account for 0.025% of the total points. This shows that the distribution of semantic classes is extremely unbalanced, highlighting the challenges for generalizing the existing segmentation approaches.

4.2. Representative Baselines

As discussed in Section 2.1, there are three main classes of neural pipelines to learn 3D point cloud semantics. In this regard, we carefully select 7 representative methods as solid baselines to benchmark our SensatUrban dataset.

- SparseConv [19]. A solid baseline that uses submanifold sparse convolutional networks and achieves leading results on ScanNet benchmark [13].
- TagentConv [50]. It projects 3D points onto tangent planes and uses 2D convolutional networks.
- PointNet/PointNet++ [37, 38]. These are the seminal works to directly operate on orderless point clouds.
- SPGraph [24]. This is one of the first approaches capable of directly processing large-scale point clouds via the concept of superpoints.
- KPConv [51]. It introduces a flexible kernel point convolution and achieves state-of-the-art performance on the DALES dataset [56].
- RandLA-Net [23]. It is the latest work for efficient semantic segmentation of large-scale point clouds and ranks the first place on Semantic3D leaderboard [21].

4.3. Evaluation Metrics

Like the existing benchmarks [21, 5, 3], we use the Overall Accuracy (OA) and mean Intersection-over-Union (mIoU) as the principle evaluation metrics.

4.4. Benchmark Results

For fair comparison, we faithfully follow the experimental settings of each baseline in the original publication. Table 2 presents the quantitative results. PointNet [37] has the worst performance, while KPConv [51] achieves the highest mIoU scores. However, the overall segmentation performance is far from satisfactory. For example, there are still a number of key categories such as *bridge, rail, street, footpath* that are poorly segmented. Furthermore, the category *bike* is entirely unsegmented by all methods. Further note that different techniques have vastly different performances on these challenging categories, with no clear leader. Motivated by this, we then investigate the particular challenges that arise from our new urban-scale SensatUrban dataset.

	Sampling	Input sets	OA(%)	mAcc(%)	**mIoU(%)**	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet	Grid	Constant Number	**90.57**	**56.30**	**49.69**	83.55	**97.67**	90.66	22.56	**43.54**	40.35	9.29	**50.74**	**29.58**	68.24	29.27	0.00	**80.55**
PointNet	Grid	Constant Volume	88.27	49.80	42.44	80.20	96.43	87.88	8.45	35.14	32.52	0.00	43.03	19.26	54.66	18.26	0.00	75.87
PointNet	Random	Constant Number	90.34	55.17	48.49	83.47	97.51	90.89	18.55	33.31	42.82	11.85	47.95	26.83	68.37	29.12	0.00	79.71
PointNet	Random	Constant Volume	88.09	48.45	41.68	79.82	96.24	87.64	5.69	27.70	34.98	0.00	42.85	13.81	54.29	20.64	0.00	78.24
RandLA-Net	Grid	Constant Number	**91.55**	**74.87**	**58.64**	82.99	98.43	**93.41**	57.43	49.47	55.12	27.33	60.65	39.43	84.57	39.48	0.00	73.97
RandLA-Net	Grid	Constant Volume	88.11	64.91	49.18	78.18	97.92	90.87	45.02	30.89	35.82	0.00	45.73	31.96	77.78	29.90	0.00	75.30
RandLA-Net	Random	Constant Number	91.14	74.14	57.55	82.25	98.33	92.37	54.20	43.10	54.74	25.02	60.40	39.17	82.77	37.59	0.00	78.25
RandLA-Net	Random	Constant Volume	88.37	60.84	47.27	81.16	97.52	90.45	44.75	16.36	37.18	0.00	4219	26.28	76.76	30.46	0.00	71.39

Table 3: Quantitative results achieved by PointNet [23] and RandLA-Net [23] with different input preparation steps. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

5. Challenges

In this section, we identify a number of key challenges revealed by our SensatUrban, and explore the possible solutions to overcome them, eventually improving the segmentation performance for existing point-based approaches. Note that, we are not aiming to propose new algorithms in this section. Instead, we aim to generalize the existing pipelines from the perspective of dataset characteristics.

5.1. Data Partition

Due to the limited memory of modern GPUs, the first and foremost challenge is to partition the original large-scale point clouds, such that computational efficiency and segmentation accuracy can be well balanced. The early PointNet/PointNet++ techniques [37, 38] typically divide the point clouds into 1×1 meter blocks. This is however highly time consuming for such a large input tile, and causes the object geometry to be fragmented across blocks. On the other hand, if the raw point clouds are divided into extremely large blocks, the high number of points are unable to be fed into the limited GPUs. To reduce the total number of points within each block, grid or random down-sampling are applied in [51, 23]. Many other methods tend to use different sampling and partitioning tricks. Overall, there is no standard and principled preparation steps in literature to partition the large-scale point clouds.

To demonstrate the impact of different data partition schemes, we organize the data preparation into two separate steps as follows.

- Step 1. To downsample the raw point clouds at the very beginning. There are two options in literature: 1) grid downsampling [51], and 2) random downsampling [23]. Both can significantly reduce the total amount of points, but each have their relative merits.
- Step 2. To obtain individual input set of points to feed into the networks. There are two choices: 1) constant-number input sets (*i.e.*, the number of points is fixed), and 2) constant-volume input sets (*i.e.*, the volume of the point set is fixed). In particular, constant-number input sets are usually obtained by querying a fixed number of

points with regard to the set center [51, 23], while the constant-volume input sets are extracted by collecting all points of a fixed-size cube [37, 38].

By using two representative baselines PointNet [37] and RandLA-Net [23], we evaluate how the four different combinations of both Step 1 and Step 2 affect the accuracy of segmentation. In all the experiments, the grid size for downsampling is $0.2m$, the random downsampling ratio is 1/10, the size for constant-volume sets is $8 \times 8m^2$, and the constant-number sets have 4096 points.

Analysis. Table 3 shows the semantic segmentation scores of the eight groups of experiments on the testing split of SensatUrban. It can be seen that:

- Both PointNet or RandLA-Net based baselines achieve much higher scores when the input sets are number constant, compared with cases of constant volume.
- Using grid downsampling to reduce the raw 3D point clouds demonstrates marginally better results than random downsampling for both PointNet and RandLA-Net.

Overall, our experiments show that the data preparation is indeed of great importance. A simple combination of grid sampling and number-consistent block partition can bring about up to 10% improvement for mIoU scores. In this regard, we firmly believe that more studies should be conducted to further explore the effective ways for data preparation.

5.2. Geometry vs. Appearance

One of the key differences between our SensatUrban and the existing LiDAR-based datasets [56, 43, 5] is the availability of true RGB color for every 3D point. Intuitively, the colored point clouds tend to be more informative and can provide the networks with additional features for better segmentation accuracy. However, networks may overfit the appearance and fail to learn robust features from the geometry. Taking only 3D coordinates as the input, the recent ShellNet [68] achieves surprisingly good results, highlighting the power of geometry. To investigate whether and how the appearance impacts the final segmentation performance,

	OA(%)	mAcc(%)	**mIoU(%)**	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet [37] (w/o RGB)	83.50	33.52	28.85	67.35	92.66	84.72	16.02	0.00	13.65	2.68	17.09	0.33	54.54	0.00	0.00	26.04
PointNet [37] (w/ RGB)	90.57	56.30	49.69	83.55	97.67	90.66	22.56	43.54	40.35	9.29	50.74	29.58	68.24	29.27	0.00	80.55
PointNet++ [38] (w/o RGB)	90.85	56.94	50.71	79.05	98.37	94.22	66.76	39.74	37.51	0.00	51.53	38.82	81.71	5.80	0.00	65.68
PointNet++ [38] (w RGB)	93.10	64.96	58.13	86.38	98.76	94.72	65.91	50.41	50.53	0.00	58.40	46.95	82.31	38.40	0.00	**82.88**
SPGraph [24] (w/o RGB)	84.81	42.12	35.29	69.60	94.18	88.15	34.55	20.53	15.83	16.34	31.44	10.54	55.01	0.98	0.00	21.57
SPGraph [24] (w RGB)	85.27	44.39	37.29	69.93	94.55	88.87	32.83	12.58	15.77	15.48	30.63	22.96	56.42	0.54	0.00	44.24
KPConv [51] (w/o RGB)	91.47	57.43	51.79	80.43	98.82	94.93	74.17	44.53	32.11	0.00	54.32	37.83	84.88	14.48	0.00	56.79
KPConv [51] (w RGB)	**93.92**	71.44	64.50	87.04	**99.01**	96.31	77.73	58.87	49.88	**37.84**	62.74	56.60	86.55	44.86	0.00	81.01
RandLA-Net [23] (w/o RGB)	88.90	67.96	51.53	77.30	97.92	91.24	51.94	47.46	45.04	9.71	49.79	34.21	79.97	21.13	0.00	64.18
RandLA-Net [23] (w RGB)	91.24	**74.68**	58.14	82.23	98.39	92.69	56.62	49.00	**54.19**	25.10	60.98	38.69	83.42	38.74	0.00	75.80

Table 4: Quantitative results of five selected baselines on our SensatUrban dataset. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

we conduct the following ten experiments using five different baselines namely PointNet/PointNet++ [37, 38], SP-Graph [24], KPConv [51], and RandLA-Net [23]. These are either trained using only geometrical information (*i.e.*, 3D coordinates) or both 3D coordinates and RGB information.

Analysis. Table 4 presents the quantitative results of the five baselines with respect to the different types of input point clouds. It can be seen that:

- All of PointNet/PointNet++, KPConv and RandLA-Net achieve significantly better segmentation accuracy when the networks are trained given both point coordinates and RGB information. Fundamentally, this is because a number of urban classes (*e.g.*, bridge, footpath, water, *etc.*.) are virtually impossible to be discriminated between, if only supplied with 3D coordinates.
- For SPGraph, the performance depends largely on the geometrical partition which purely relies on the point coordinates, hence the inclusion of RGB does not yield a significant performance boost.

For all techniques, the presence of color information is critical to improve the accuracy of semantic segmentation in urban-scale scenarios. This highlights the advantage of our SensatUrban over the existing LiDAR based datasets such as DALES [56] and also suggests that future aerial mapping campaigns should consider including RGB.

5.3. The Impact of Imbalance Class Distribution

Regardless of whether RGB is included or not, there still remain significant performance gaps between different categories. For example, the score of *vegetation* is around 99%, while the *bike* is completely unable to be recognized. Fundamentally, urban areas are dominated by a small number of categories such as *vegetation*, and *road*, while the minor yet important classes such as *bike* account for a minute portion of points. This extremely skewed distribution is another significant challenge arising from SensatUrban.

To alleviate this problem, a typical solution is to use more sophisticated loss functions. We evaluate the effec-

tiveness of five off-the-shelf loss functions, with Point-Net and RandLA-Net as baselines. The loss functions are: cross-entropy, weighted cross-entropy with inverse frequency [12], or with inverse square root (sqrt) frequency [41], *Lovász*-softmax loss [6], and focal loss [28].

Analysis. Table 5 shows the quantitative results of the two baselines with the five different loss functions. It can be seen that the inclusion of advanced loss functions indeed improves the segmentation performance. The mIoU scores gain up to 5%. Notably, for the extremely challenging category *bike*, the baseline RandLA-Net trained with weighted cross-entropy and sqrt [41] obtains more than 20% improvement. This shows that data imbalance can be alleviated, to an extent, by using off-the-shelf loss functions. However, even this increased performance is hardly satisfactory, and we suggest that it is still an open question to explore more effective solutions to fully tackle this challenge.

5.4. Cross-City Generalization

A common issue of deep neural networks lies in their (in)ability to directly generalize to unseen scenarios. To this end, our SensatUrban includes large-scale point clouds from two different urban areas, which allows us to fully evaluate their generalization ability. We conduct experiments based on 5 baselines: PointNet/PointNet++ [37, 38], SPGraph [24], KPConv [51], and RandLA-Net [23].

- Train Birmingham/Test Birmingham: Each of the 5 baselines is only trained on the training split of Birmingham, and then tested on the testing split of the same region.
- Train Birmingham/Test Cambridge: The above well-trained 5 baseline models are directly tested on the testing split of Cambridge.

Analysis. Table 6 compares the quantitative results of our experiments. It can be seen that the segmentation performance of all baselines drops significantly when the trained models are directly applied to novel urban scenarios. The mIoU scores have up to 20% gaps for most approaches.

	OA(%)	mAcc(%)	**mIoU(%)**	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet+ce	**90.57**	56.30	49.69	**83.55**	97.67	**90.66**	22.56	43.54	40.35	9.29	**50.74**	29.58	**68.24**	29.27	0.00	**80.55**
PointNet+wce [23]	88.13	**68.05**	51.24	81.01	97.12	87.87	24.46	45.76	47.78	**34.93**	49.82	29.58	61.28	31.78	0.00	74.67
PointNet+wce+sqrt [2]	89.72	67.97	52.35	82.87	97.33	90.42	28.32	44.94	48.39	32.07	49.58	32.63	65.11	32.59	**2.60**	73.71
PointNet+lovas [6]	89.58	67.50	**52.53**	82.74	97.27	90.28	28.11	43.89	48.53	33.58	49.68	32.21	64.01	33.05	1.46	78.13
PointNet+focal [28]	89.46	67.33	52.37	82.47	97.34	90.25	28.36	**51.87**	46.40	30.50	48.62	32.43	65.00	32.23	1.21	74.10
RandLA-Net+ce	**93.10**	64.30	57.77	**85.39**	98.63	95.40	62.55	54.85	56.49	0.00	58.13	45.90	82.24	30.68	0.00	80.70
RandLA-Net+wce [23]	91.24	74.68	58.14	82.23	98.39	92.69	56.62	49.00	54.19	25.10	**60.98**	38.69	83.42	38.74	0.00	75.80
RandLA-Net+wce+sqrt [2]	92.51	**79.92**	62.80	84.94	98.47	95.07	59.01	62.18	56.76	28.96	57.36	44.47	84.67	41.67	24.31	78.49
RandLA-Net+lovas [6]	92.56	76.99	61.51	84.92	98.55	94.64	63.17	52.37	55.43	36.37	59.35	45.79	84.28	41.24	2.66	80.89
RandLA-Net+focal [28]	92.49	77.26	60.41	85.03	98.38	94.74	59.49	58.70	**57.11**	25.97	58.19	42.74	82.26	**42.00**	2.71	77.97

Table 5: Quantitative results achieved by PointNet [37] and RandLA-Net [23] with different loss functions. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

	OA(%)	mAcc(%)	**mIoU(%)**	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet [37]	87.33	54.76	48.73	80.91	94.58	87.40	33.69	0.51	66.23	16.98	49.55	36.08	74.59	1.49	0.00	91.51
PointNet++ [38]	89.85	64.24	57.39	84.34	97.11	89.74	61.56	**3.78**	68.08	41.95	54.43	51.54	84.73	14.43	0.00	**94.34**
SPGraph [24]	80.13	42.87	36.95	65.75	93.33	87.24	41.28	0.00	42.69	20.94	2.28	32.05	64.06	0.00	0.00	30.76
KPConv [51]	91.44	68.41	61.65	86.00	97.66	92.90	75.07	0.91	69.74	55.50	57.94	60.73	89.48	21.44	0.00	94.13
RandLA-Net [23]	90.77	72.11	59.72	85.14	96.89	90.77	59.45	1.52	75.83	48.88	62.58	48.65	86.31	28.82	0.00	91.51
PointNet [37]	86.06	38.56	29.70	74.94	94.57	85.38	8.62	13.42	16.47	0.00	38.64	14.27	36.96	0.09	0.00	2.75
PointNet++ [38]	89.46	44.64	36.93	77.68	97.28	91.95	54.59	0.52	15.84	0.00	42.08	29.00	67.71	0.24	0.00	3.16
SPGraph [24]	82.02	24.83	20.70	61.72	88.26	78.27	8.29	0.00	0.00	0.00	0.64	1.87	30.00	0.00	0.00	0.00
KPConv [51]	90.62	48.71	**40.51**	78.88	98.33	94.24	76.20	0.01	14.70	0.00	41.77	**39.32**	74.22	0.39	0.00	**8.61**
RandLA-Net [23]	88.92	**51.57**	40.29	78.46	97.12	89.93	46.77	**28.76**	20.03	0.00	**46.98**	18.70	65.99	**24.91**	0.00	6.15

Table 6: All baselines are trained on the Birmingham split. The top five records show the testing results on the testing split of Birmingham, while the bottom five rows show the scores on the testing split of Cambridge. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.

Interestingly, the major categories such as ground and building do not observe severe performance drops, while the classes such as rail, street and water have the worst generalization scores. From this, we hypothesize that: 1) the imbalanced semantic distribution plays a key role in preventing the model generalization, mainly because the model tends to fit with major classes and fails to learn robust features of minor categories; 2) the more variable morphology of some urban classes such as *parking* and *water* are hard to be generalized from one dataset to another. Due to a lack of realistic datasets, few studies have been conducted to investigate this critical issue of generalization. It is thus an open question of how to robustly label novel urban-scale regions.

6. Summary and Outlook

In this paper, we introduce a large and rich urban-scale dataset including two accurately labelled regions covering $4.4km^2$ and an extra unlabelled region covering $3.2km^2$ provided for the self/semi-supervised learning schemes. Through extensive benchmarking, we highlight a number of open challenges, which include how to sample and partition the large point clouds, whether to acquire RGB (color) information or not, the impact of a significantly imbalanced class distribution, and the lack of robust generalization to unseen scenarios. Other pressing challenges include instance-level and panoptic segmentation. In the near-future, we envisage that autonomous aerial vehicles will intelligently navigate through dense cities, urban, and rural areas, and as such, real-time photogrammetric reconstruction and segmentation are also of key consideration. Accurate and high resolution 3D maps of reality are also necessary ingredients for emerging cyberphysical areas such as smart cities, intelligent transport and digital twins. It is our hope that our SensatUrban dataset and benchmark will be a stepping stone towards advancing research in related areas.

References

[1] Lyft level 5 dataset. https://self-driving.lyft.com/level5/data/. 2

[2] Eren Erdal Aksoy, Saimir Baci, and Selcuk Cavdar. SalsaNet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving. IV, 2020. 8

[3] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. In CVPR, 2017. 1, 2, 3, 5

[4] Mark Austin, Parastoo Delgoshaei, Maria Coelho, and Mohammad Heidarinejad. Architecting smart city digital twins: Combined semantic model and machine learning approach. Journal of Management in Engineering, 2020. 1

[5] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In ICCV, 2019. 1, 2, 3, 5, 6

[6] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In CVPR, 2018. 7, 8

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In CVPR, 2020. 2

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. arXiv preprint arXiv:1512.03012, 2015. 1, 2, 3

[9] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3D tracking and forecasting with rich maps. In CVPR, 2019. 2

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In CVPR, 2019. 1, 3

[11] Nico Cornelis, Bastian Leibe, Kurt Cornelis, and Luc Van Gool. 3D urban scene modeling integrating recognition and reconstruction. IJCV, 2008. 1

[12] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast semantic segmentation of lidar point clouds for autonomous driving. arXiv preprint arXiv:2003.03653, 2020. 3, 7

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In CVPR, 2017. 1, 2, 3, 5

[14] Mark De Deuge, Alastair Quadros, Calvin Hung, and Bertrand Douillard. Unsupervised feature learning for classification of outdoor 3D scans. In ACRA, 2013. 2

[15] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In CVPR, 2016. 2

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. IJRR, 2013. 1

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In CVPR, 2012. 2

[18] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2D2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320, 2020. 2

[19] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In CVPR, 2018. 1, 3, 5

[20] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3D point clouds: A survey. IEEE TPAMI, 2020. 2

[21] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3D.Net: A new large-scale point cloud classification benchmark. ISPRS, 2017. 1, 2, 3, 5

[22] A Handa, V Patraucean, V Badrinarayanan, S Stent, and R Cipolla. SceneNet: understanding real world indoor scenes with synthetic data. In CVPR, 2016. 2

[23] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. CVPR, 2020. 1, 3, 5, 6, 7, 8, 17

[24] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In CVPR, 2018. 5, 7, 8

[25] Truc Le and Ye Duan. PointGrid: A deep network for 3D shape understanding. In CVPR, 2018. 3

[26] Xinke Li, Chongshou Li, Zekun Tong, Andrew Lim, Junsong Yuan, Yuwei Wu, Jing Tang, and Raymond Huang. Campus3d: A photogrammetry point cloud benchmark for hierarchical understanding of outdoor scene. In ACM MM, 2020. 2, 3

[27] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. PointCNN: Convolution on X-transformed points. In NeurIPS, 2018. 3

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV, 2017. 7, 8

[29] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3D deep learning. In NeurIPS, 2019. 3

[30] Yecheng Lyu, Xinming Huang, and Ziming Zhang. Learning to segment 3D point clouds in 2D image space. In CVPR, 2020. 3

[31] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. SceneNet RGB-D: 5m photorealistic images of synthetic indoor trajectories with ground truth. arXiv preprint arXiv:1612.05079, 2016. 2

[32] Hsien-Yu Meng, Lin Gao, Yu-Kun Lai, and Dinesh Manocha. VV-Net: Voxel vae net with group convolutions for point cloud segmentation. In ICCV, 2019. 3

[33] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. RangeNet++: Fast and accurate lidar semantic segmentation. In IROS, 2019. 1, 3

[34] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In CVPR, 2019. 2, 3

[35] Daniel Munoz, J Andrew Bagnell, Nicolas Vandapel, and Martial Hebert. Contextual classification with functional max-margin markov networks. In CVPR, 2009. 2

[36] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. arXiv preprint arXiv:2002.09147, 2020. 2

[37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In CVPR, 2017. 1, 3, 5, 6, 7, 8, 12, 13, 17

[38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In NeurIPS, 2017. 1, 3, 5, 6, 7, 8, 17

[39] Deepak Rao, Quoc V Le, Thanathorn Phoka, Morgan Quigley, Attawith Sudsang, and Andrew Y Ng. Grasping novel objects with depth segmentation. In IROS, 2010. 1

[40] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In CVPR, 2016. 2

[41] Radu Alexandru Rosu, Peer Schütt, Jan Quenzel, and Sven Behnke. Latticenet: Fast point cloud segmentation using permutohedral lattices. arXiv preprint arXiv:1912.05905, 2019. 7

[42] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS, 2012. 3

[43] Xavier Roynard, Jean-Emmanuel Deschaud, and François Goulette. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. IJRR, 2018. 2, 3, 6

[44] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In Advances in Neural Information Processing Systems, pages 12962–12972, 2019. 12, 13

[45] Andrés Serna, Beatriz Marcotegui, François Goulette, and Jean-Emmanuel Deschaud. Paris-rue-madame database: a 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods. 2014. 3

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. In ECCV, 2012. 2

[47] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In CVPR, 2015. 2

[48] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In CVPR, 2020. 2

[49] Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. Toronto-3D: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In CVPRW, 2020. 2, 3

[50] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3D. In CVPR, 2018. 3, 5

[51] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KPConv: Flexible and deformable convolution for point clouds. In ICCV, 2019. 1, 3, 5, 6, 7, 8, 17

[52] Guofeng Tong, Yong Li, Dong Chen, Qi Sun, Wei Cao, and Guiqiu Xiang. CSPC-dataset: New lidar point cloud dataset and benchmark for large-scale scene semantic segmentation. IEEE Access, 2020. 2

[53] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In ICCV, 2019. 2

[54] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In ICRA, 2017. 1

[55] Bruno Vallet, Mathieu Brédif, Andrés Serna, Beatriz Marcotegui, and Nicolas Paparoditis. TerraMobilita/iQmulus urban point cloud analysis benchmark. Computers & Graphics, 2015. 3

[56] Nina Varney, Vijayan K Asari, and Quinn Graehling. DALES: A large-scale aerial lidar data set for semantic segmentation. In CVPRW, 2020. 2, 3, 4, 5, 6, 7, 12

[57] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J Kusner. Pre-training by completing point clouds. arXiv preprint arXiv:2010.01089, 2020. 12, 13

[58] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. ACM TOG, 2019. 3, 12, 13

[59] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D lidar point cloud. In ICRA, 2018. 1

[60] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In CVPR, 2015. 1, 2, 13

[61] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas J Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. arXiv preprint arXiv:2007.10985, 2020. 12

[62] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation. ECCV, 2020. 3

[63] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3D instance segmentation on point clouds. In NeurIPS, 2019. 3

[64] Zhen Ye, Yusheng Xu, Rong Huang, Xiaohua Tong, Xin Li, Xiangfeng Liu, Kuifeng Luan, Ludwig Hoegner, and Uwe Stilla. Lasdu: A large-scale aerial lidar dataset for semantic labeling in dense urban areas. ISPRS International Journal of Geo-Information, 2020. 2, 3

[65] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3D shape collections. ACM TOG, 2016. 2

[66] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In 3DV, 2018. 12, 13

[67] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. PolarNet: An improved grid representation for online lidar point clouds semantic segmentation. In CVPR, 2020. 3

[68] Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics. In ICCV, 2019. 6

[69] Zhenchao Zhang et al. A patch-based method for the evaluation of dense image matching quality. Int. J. Appl. Earth. Obs. Geoinf, 2018. 12

[70] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In CVPR, 2017. 1, 2

[71] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3D object detection. In CVPR, 2018. 3

[72] SM Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Rogerio Eduardo da Silva, Morteza Rahbar, and Aljosa Smolic. DublinCity: Annotated lidar point cloud and its applications. In BMVC, 2019. 2, 3

Appendix

A. Details of the Data Collection

Our dataset is reconstructed from 2D aerial images using the well-established structure-from-motion technique, which recovers the camera extrinsic parameters for each image. The byproduct orthomosaics are only used for visualization purposes. The data are validated using GNSS RTK manual surveying carried out by professional operators. The final horizontal and vertical RMSEs are ±50mm and ±75mm, respectively. As a comparison, the positioning accuracy of LiDAR point clouds is around 5 to 10 cm, depending on the equipment quality, flying configuration, post-processing, etc. [69]. We use Sensefly Soda 3D to capture the aerial images. The detailed specification of the camera can be found in Table 7. The 2D aerial images are filmed from both nadir and oblique perspectives, therefore the points on vertical surfaces are well captured. The resolution of our data depends on the number of input images and 3D reconstruction settings. Normally, photogrammetric point clouds are very dense from the process of dense image matching and so need to be subsampled. In our case, all points are subsampled at 2.5 cm, which is denser than most LiDAR data such as DALES [56].

	Specification
Sensor size	1 inch
RGB Lens	F/2.8-11, 10.6 mm (35 mm equivalent: 29 mm)
RGB Resolution	5,472 x 3,648 px (3:2)
Exposure compensation	±2.0 (1/3 increments)
Shutter	Global Shutter 1/30 − 1/2000s
White balance	Auto, sunny, cloudy, shady
ISO range	125-6400
RGB FOV	Total FOV: 154°, 64° optical, 90° mechanical
GNSS	RTK/PPK

Table 7: Detailed specifications of the camera used in our survey.

B. Details of the Data Annotation

We use CloudCompare to label all the points in pure 3D. There are no unassigned points discarded in the process. To ensure the annotation quality, all annotations have been manually cross-checked. We notice that the instance annotation would be a meaningful addition to our dataset. However, due to the tremendous labeling effort of point-wise instance labels, we leave the integration of instance labels for future exploration.

We initially labelled the point cloud as highly fine-grained 31 categories, including *benches*, *bollards*, *road signs*, *traffic lights*, *etc*. Considering the scarcity of data points in certain categories, we merged some similar categories together. The initial label, merged label, and detailed mapping will be released along with the dataset.

C. Visualization of the Dataset

As mentioned in Section 4, the whole urban-scale point clouds have been divided into several non-overlap tiles similar to DALES [56]. To have an intuitive and clear understanding of the data, we visualize the tiles in Birmingham and Cambridge in Figure 5 and Figure 6, respectively. In addition, we also show some zoomed-in urban scenes from the York data in Figure 7.

D. Additional Quantitative Results

D.1. Pre-training on pretext task

Recently, a handful of works [44, 57, 61] have started to design pretext tasks to achieve network pre-training based on the self-supervised learning framework. To further verify the effects of this training strategy on our urban-scale point clouds dataset, we conducted several groups of experiments on our SensatUrban dataset. Specifically, we evaluate the performance of two pretraining schemes: occlusion completion [57] and context prediction [44], based on three baseline networks, including PointNet [37], PCN [66], and DGCNN [58]. The detailed experimental results are shown in Table 8.

From the results in Table we can see that, although the baseline networks are only pre-trained on the object-level point clouds, the fine-tuning model can still achieve a certain performance improvement on our dataset. In particular, the performance of several minority categories, such as *rail* and *bridge*, has a significant performance improvement (up to nearly 10%), primary because the pre-trained models are less prone to overfitting to the majority categories, compared to directly training from scratch. This further demonstrates the feasibility of the pre-training strategy. However, the existing pre-training paradigm [57, 44] are still limited to object-level point clouds, and it is non-trivial to be extended to large-scale point clouds. To this end, we release our unlabeled York point clouds, encouraging more studies conducted in this research area.

E. Qualitative Results

We also show the corresponding qualitative results achieved by several baselines on the test set of our Sensat-Urban in Figure 8. The detailed quantitative results can be found in Section 5.2.

F. Video Illustration

We provide a video demo illustrating our SensatUrban dataset, which can be viewed at https://www.youtube.com/watch?v=IG0tTdqB3L8&t=5s.

	OA(%)	mAcc(%)	**mIoU(%)**	ground	veg.	building	wall	bridge	parking	rail	traffic.	street.	car	footpath	bike	water
PointNet-Rand [37]	86.29	53.33	45.10	80.05	93.98	87.05	23.05	19.52	41.80	3.38	43.47	24.20	63.43	26.86	0.00	79.53
PointNet-Jigsaw [44]	87.38	**56.97**	47.90	83.36	94.72	88.48	22.87	30.19	47.43	15.62	44.49	22.91	64.14	**30.33**	0.00	77.88
PointNet-OcCo [57]	**87.87**	56.14	**48.50**	**83.76**	**94.81**	**89.24**	**23.29**	**33.38**	**48.04**	**15.84**	**45.38**	**24.99**	**65.00**	27.13	0.00	**79.58**
PCN-Rand [66]	86.79	57.66	47.91	82.61	**94.82**	89.04	**26.66**	21.96	34.96	28.39	43.32	27.13	62.97	30.87	0.00	80.06
PCN-Jigsaw [44]	**87.32**	57.01	48.44	**83.20**	94.79	**89.25**	25.89	19.69	**40.90**	**28.52**	43.46	24.78	63.08	31.74	0.00	**84.42**
PCN-OcCo [57]	86.90	**58.15**	**48.54**	81.64	94.37	88.21	25.43	**31.54**	39.39	22.02	**45.47**	**27.60**	**65.33**	**32.07**	0.00	77.99
DGCNN-Rand [58]	87.54	60.27	51.96	83.12	95.43	89.58	**31.84**	35.49	45.11	38.57	45.66	32.97	64.88	30.48	0.00	**82.34**
DGCNN-Jigsaw [44]	88.65	60.80	53.01	**83.95**	**95.92**	89.85	30.05	**43.59**	46.40	35.28	49.60	31.46	69.41	**34.38**	0.00	80.55
DGCNN-OcCo [57]	**88.67**	**61.35**	**53.31**	83.64	95.75	**89.96**	29.22	41.47	**46.89**	**40.64**	**49.72**	**33.57**	**70.11**	32.35	0.00	79.74

Table 8: Quantitative results achieved by using OcCo [57], Jigsaw [44] and Random (Rand) initialization on the SensatUrban dataset, based on PointNet [37], PCN [66] and DGCNN [58] encoders. Note that, all the initialized weights are obtained by pre-training on the ModelNet40 [60], since these techniques are mainly designed for object-level classification and segmentation. Overall Accuracy (OA, %), mean class Accuracy (mAcc, %), mean IoU (mIoU, %), and per-class IoU (%) are reported.
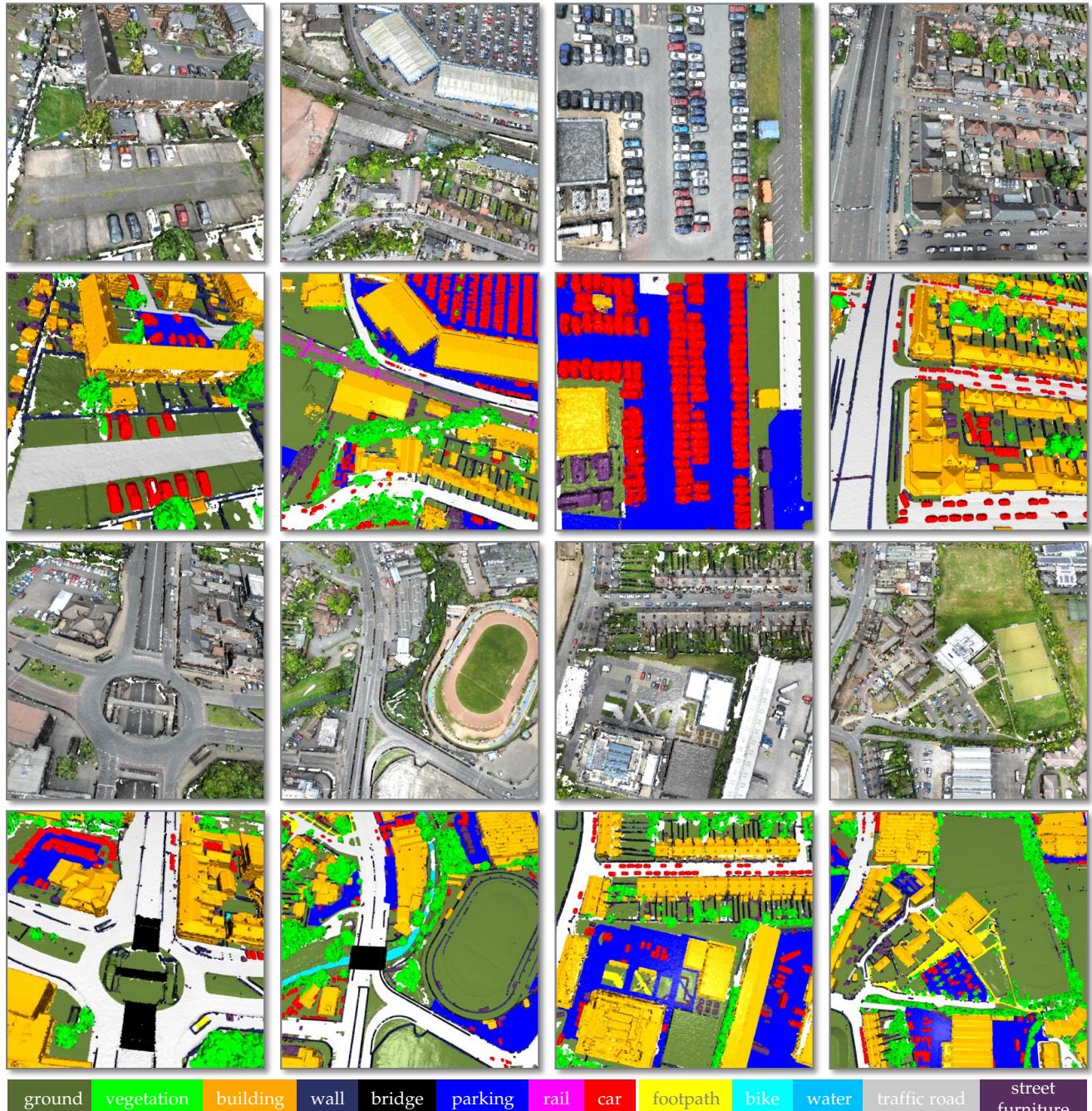
Figure 5: Birmingham split of our SensatUrban dataset. Semantic classes are labeled by different colors.

ground | vegetation | building | wall | bridge | parking | rail | car | footpath | bike | water | traffic road | street furniture
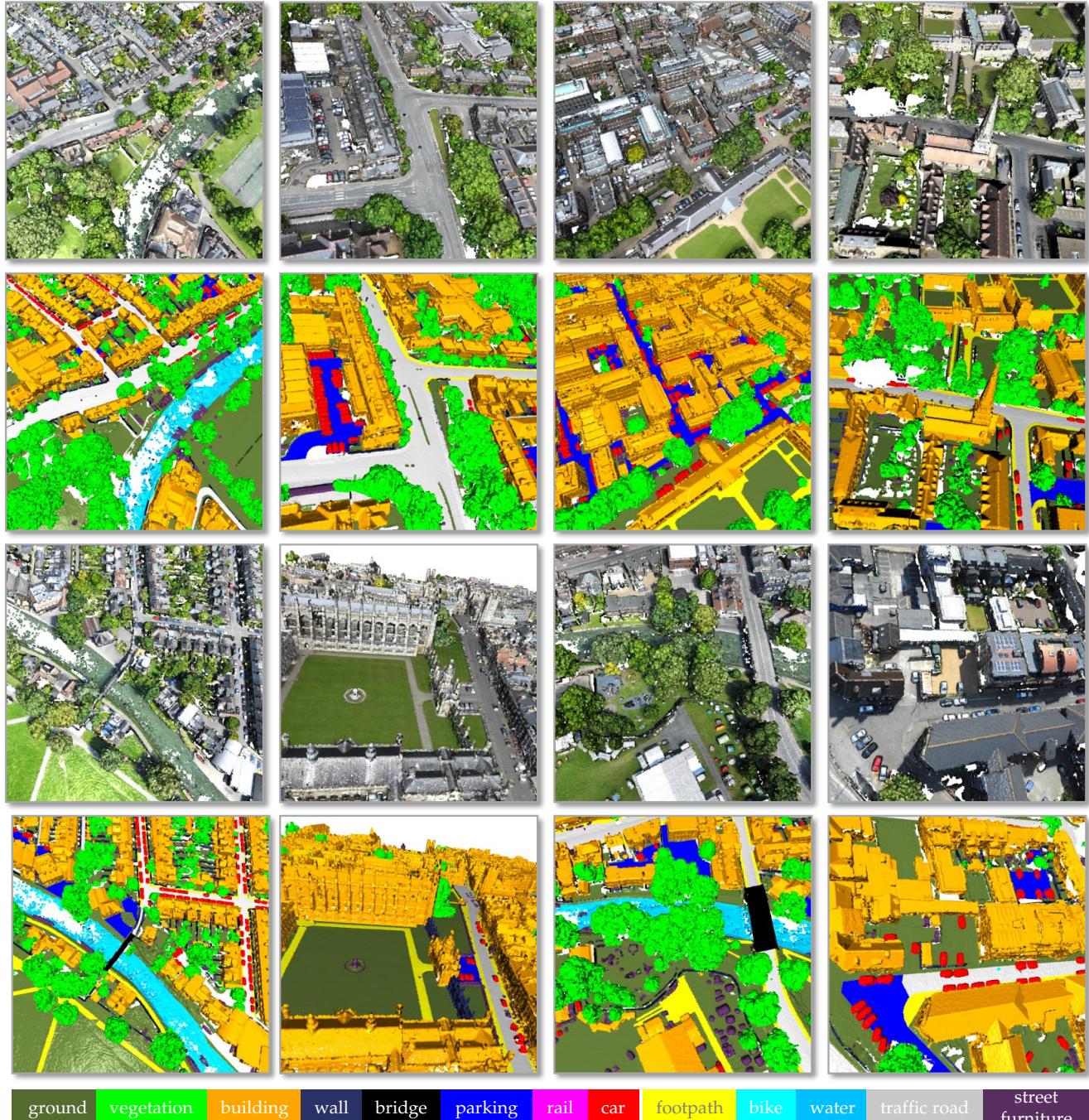
Figure 6: Cambridge split of our SensatUrban dataset. Semantic classes are labeled by different colors.

Figure 7: York split of our SensatUrban dataset. The points in York are not labeled but made available for possible pre-training in semi-supervised or self-supervised schemes. It can be seen that our urban-scale point clouds cover various elements of a real city, such as train stations, churches, stadiums, highways, etc.
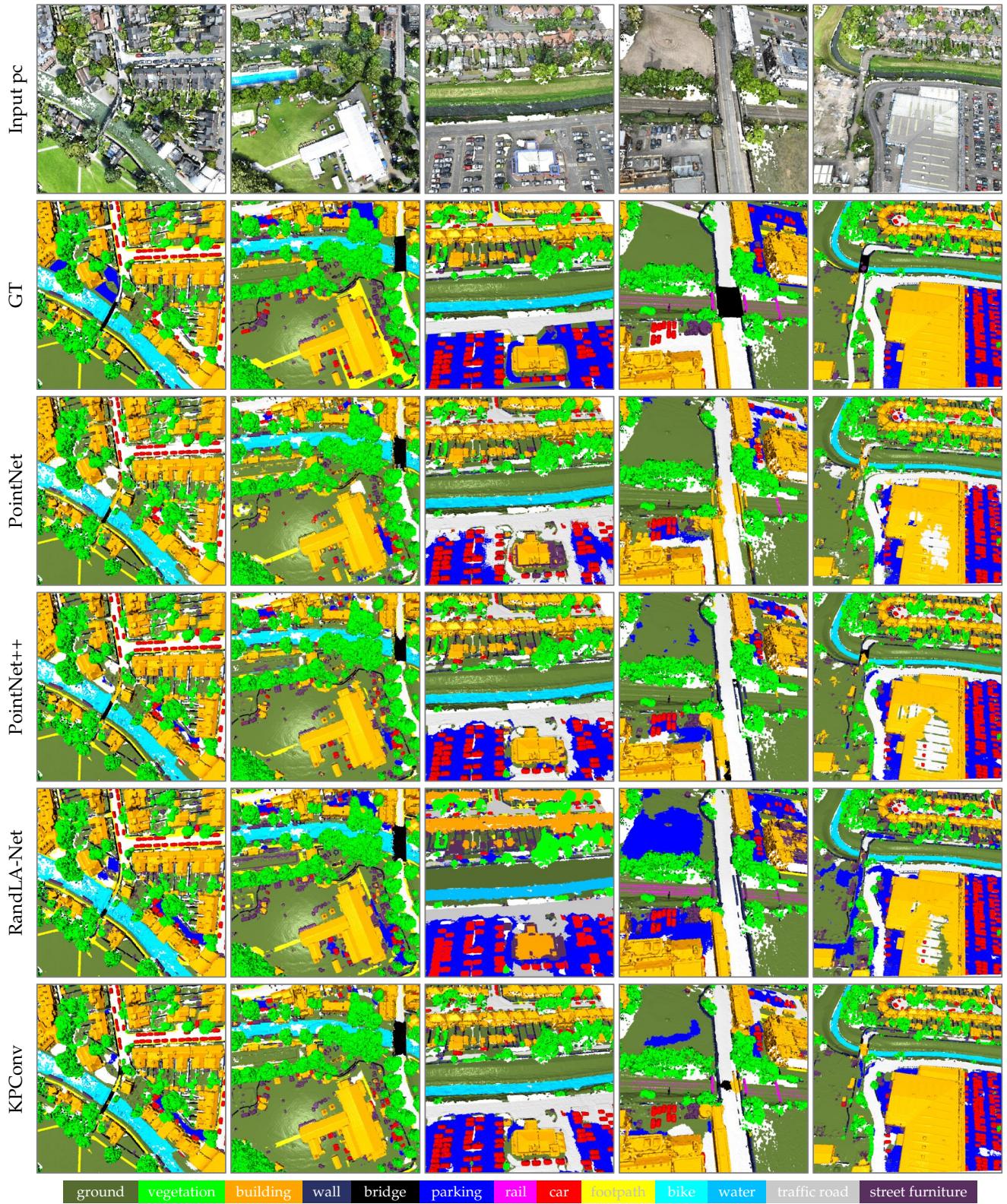
Figure 8: Qualitative results of PointNet [37], PointNet++ [38], RandLA-Net [23] and KPConv [51] on the test set of SensatUrban dataset. The black dashed box highlights the inconsistency predictions with the ground-truth label.