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 $\pm$50mm and $\pm$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 SensatUrban in Figure 8. The detailed quantitative results can be found in Section 5.2.

F. Video Illustration

We provide an anonymous video illustrating our SensatUrban dataset, which can be viewed at https://youtu.be/z84oGyEo-bs.

	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.

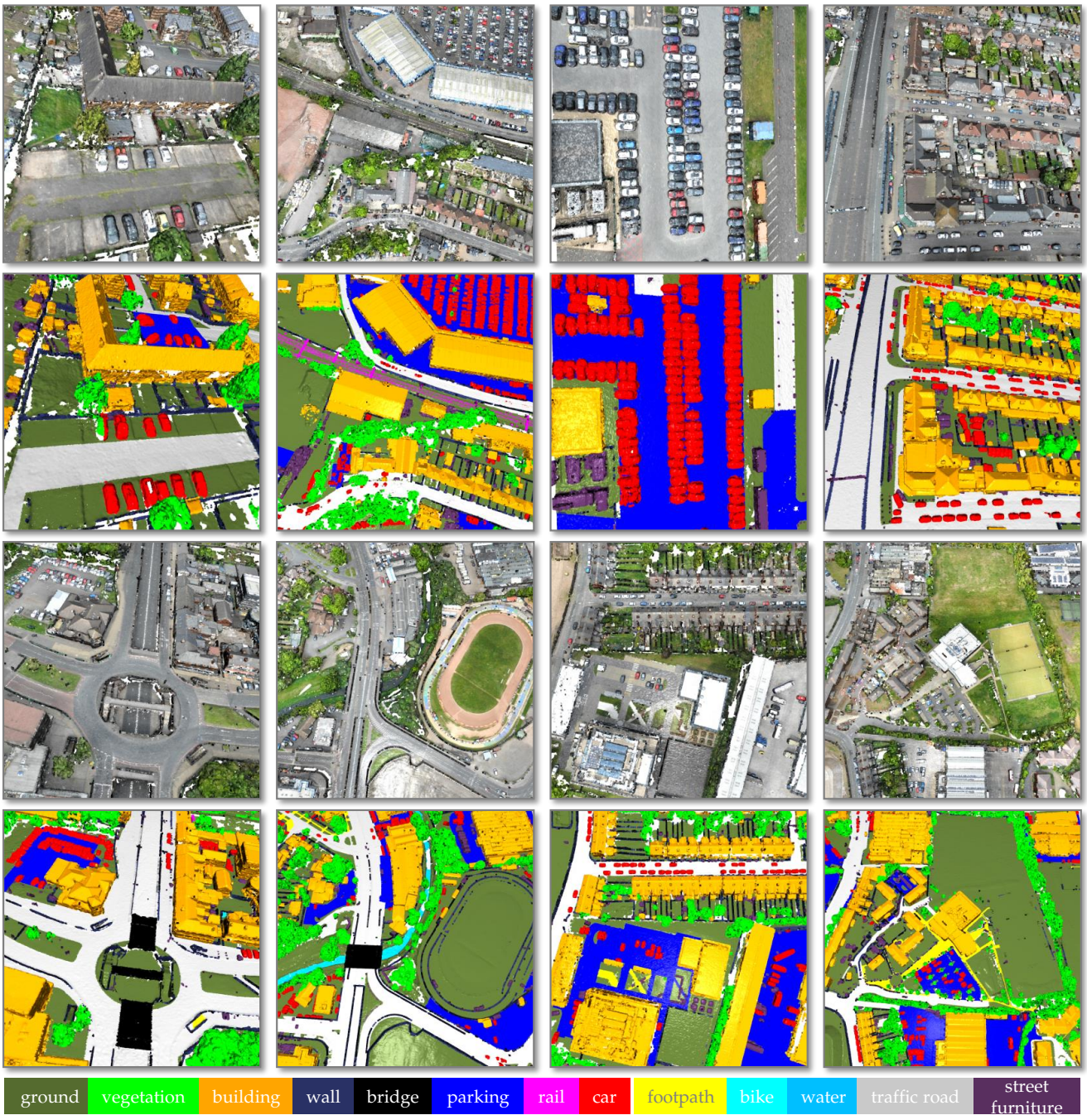

Figure 5: Birmingham split of our SensatUrban dataset. Semantic classes are labeled by different colors.

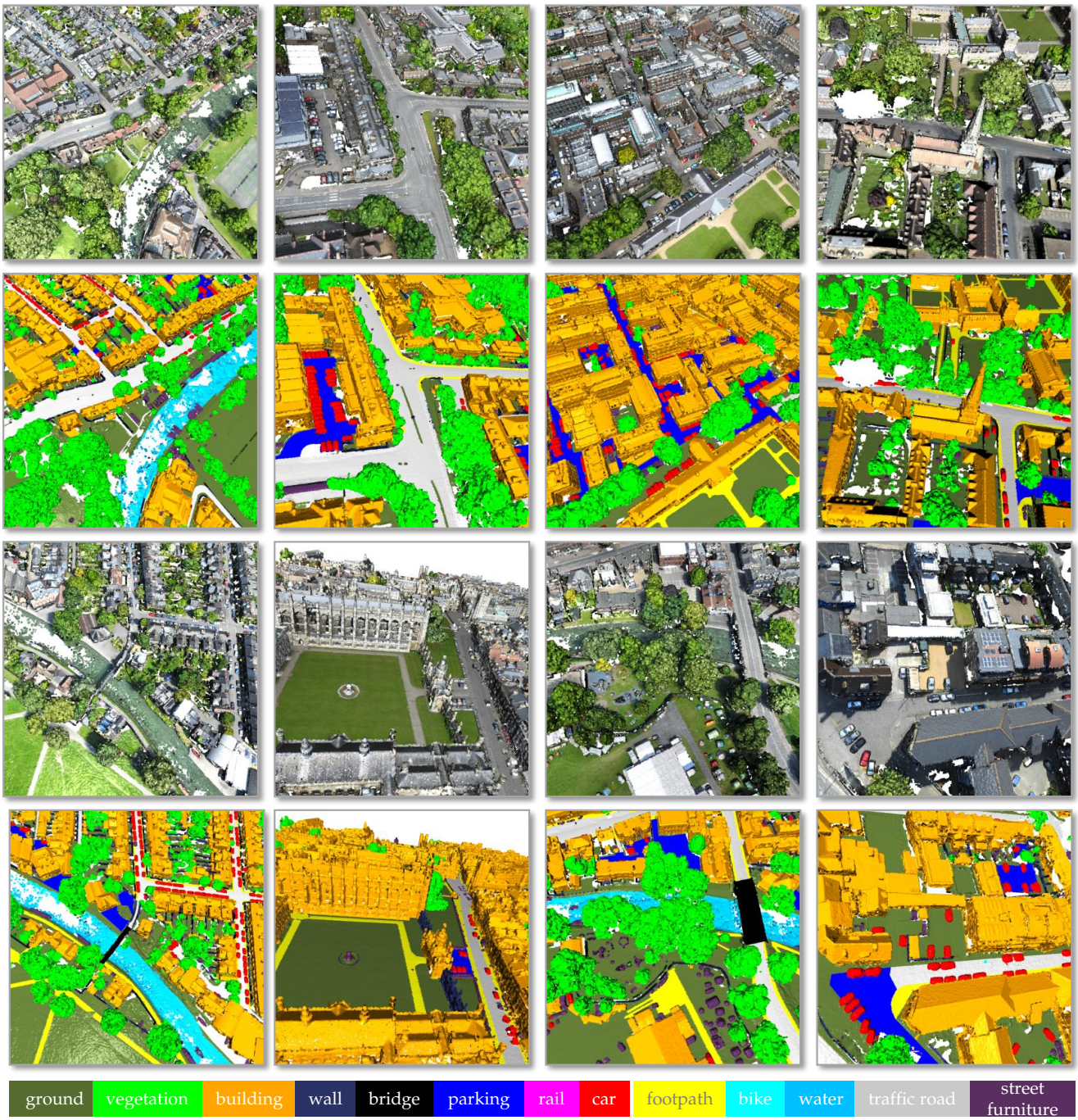

Figure 6: Cambridge split of our SensatUrban dataset. Semantic classes are labeled by different colors.

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.

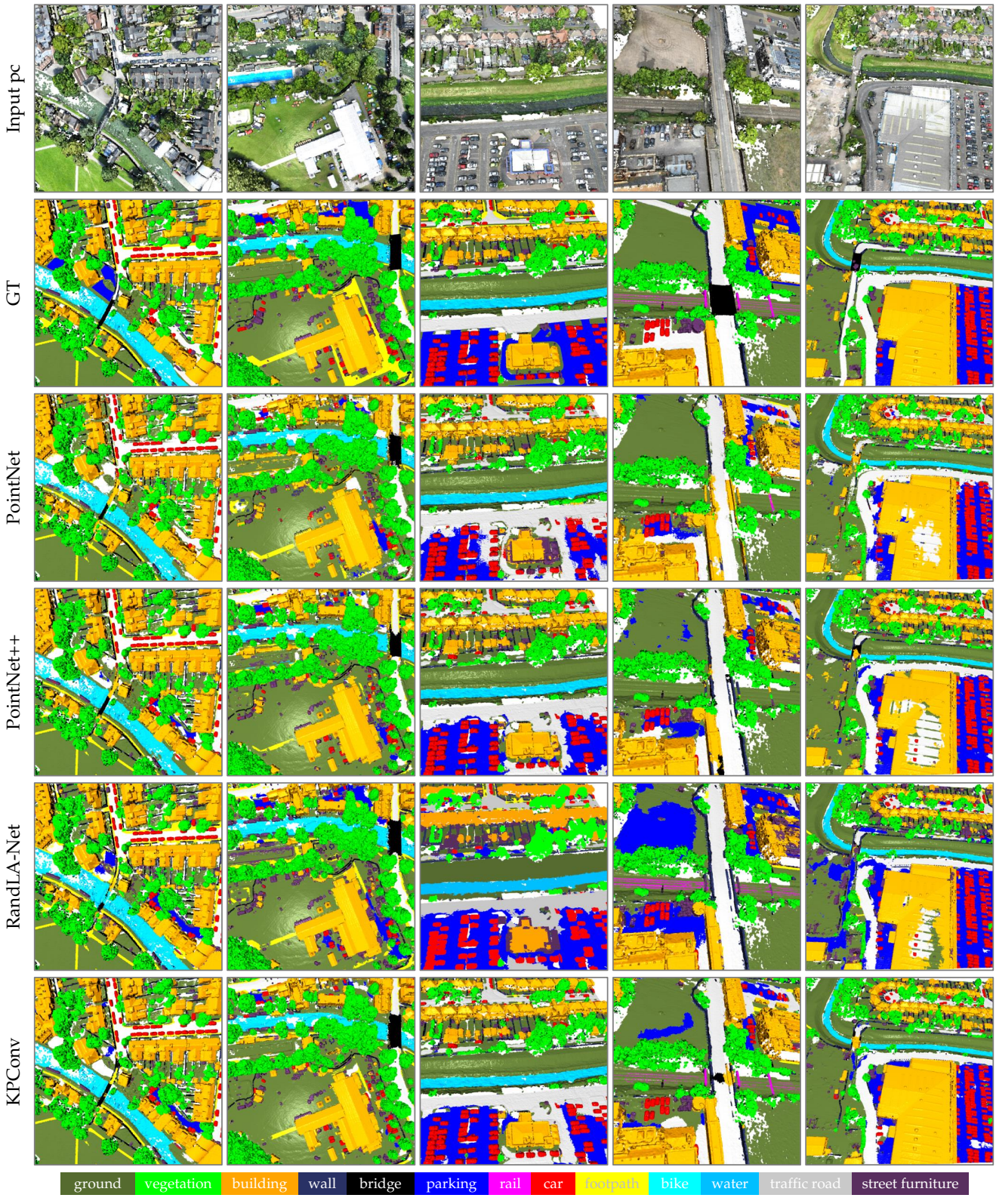

Figure 8: Qualitative results of PointNet [37], PointNet++ [38], RandLA-Net [23] and KPConv [51] on the test set of SensatUrban dataset. The black dashed box highlights the inconsistency predictions with the ground-truth label.