Campus Map: A Large-Scale Dataset to Support Multi-View VO, SLAM and BEV Estimation

James Ross[1] j.ross@surrey.ac.uk
Nimet Kaygusuz[1] n.kaygusuz@surrey.ac.uk
Oscar Mendez[1] o.mendez@surrey.ac.uk
Richard Bowden[1] r.bowden@surrey.ac.uk

***Abstract*—Significant advances in robotics and machine learning have resulted in many datasets designed to support research into autonomous vehicle technology. However, these datasets are rarely suitable for a wide variety of navigation tasks. For example, datasets that include multiple cameras often have short trajectories without loops that are unsuitable for the evaluation of longer-range SLAM or odometry systems, and datasets with a single camera often lack other sensors, making them unsuitable for sensor fusion approaches. Furthermore, alternative environmental representations such as semantic Bird's Eye View (BEV) maps are growing in popularity, but datasets often lack accurate ground truth and are not flexible enough to adapt to new research trends.**

To address this gap, we introduce Campus Map, a novel large-scale multi-camera dataset with 2M images from 6 mounted cameras that includes GPS data and 64-beam, 125k point LiDAR scans totalling 8M points (raw packets also provided). The dataset consists of 16 sequences in a large car park and 6 long-term trajectories around a university campus that provide data to support research into a variety of autonomous driving and parking tasks. Long trajectories (average 10 min) and many loops make the dataset ideal for the evaluation of SLAM, odometry and loop closure algorithms, and we provide several state-of-the-art baselines.

We also include 40k semantic BEV maps rendered from a digital twin. This novel approach to ground truth generation allows us to produce more accurate and crisp semantic maps than are currently available. We make the simulation environment available to allow researchers to adapt the dataset to their specific needs. Dataset available at: cvssp.org/data/twizy_data

I. Introduction

The unpredictability of real-world driving scenarios has led to an explosion of learning-based approaches to tackle scene understanding, odometry, navigation and mapping tasks. Each new approach requires a substantial quantity of data to train and evaluate effectively, so advances must be accompanied by datasets optimised for the application and to reflect new trends.

One such trend is the production of top-down semantic maps (BEV maps) from images. These maps provide an environmental representation that preserves the essential features and overall layout of the scene, making them very useful for downstream navigation tasks. However, training learning-based BEV predictors is challenging because of the lack of public datasets with suitable ground truth. Typically, BEV ground truth maps are generated using other data (for example, by projecting NuScenes bounding boxes onto the ground plane [1]). However, this approach is not ideal because it relies upon accurate manual annotation of 3D bounding boxes and requires significant manipulation of existing datasets for every evaluation. This is the first issue that this dataset aims to address.

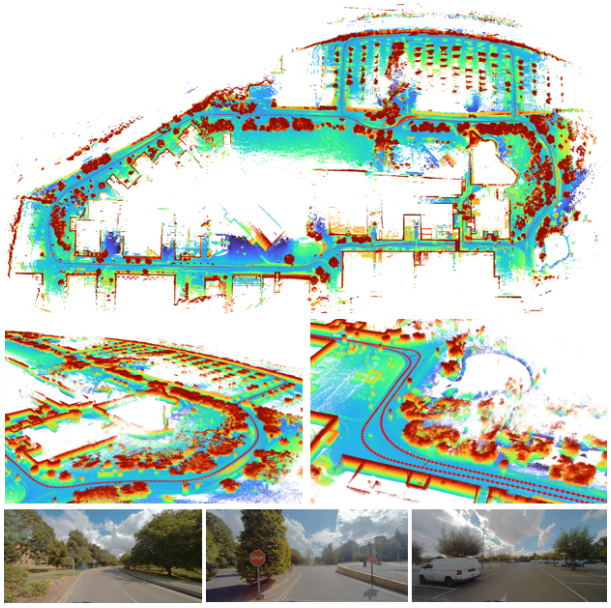

Fig. 1. Complete point cloud of university campus and car park with trajectory shown, and forward-looking sample images from the dataset.

Secondly, while there are an abundance of image datasets available for autonomous driving in general, there are few that are suitable for a wide variety of navigation tasks. For example, NuScenes [2] sequences are short (20 s) and lack loops, making them unsuitable for SLAM or odometry evaluation, and KITTI [3] sequences only contain a single stereo camera, so cannot be used to evaluate multi-view camera approaches.

Lastly, datasets that contain paired or unpaired simulated data (such as Virtual KITTI) are useful for domain transfer and image translation tasks, but the code used to produce the images is rarely provided. This means that the simulated data cannot be easily extended or adapted to cater to new research trends.

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

In this work, we introduce a new dataset and related tools to address these drawbacks and provide data that is immediately useful for a wide variety of navigation tasks. The dataset comprises 16 sequences in a car park and 6 around a university campus. Each sequence contains a variety of general driving and parking manoeuvres.

We use an NVIDIA Drive PX2 with 6 cameras mounted around the ego vehicle, capturing RGB images at 28fps. LiDAR packets, point clouds and GPS measurements are included. Ground truth trajectories have been reconstructed and are also provided. Long trajectories with loops make the data ideal for the evaluation of single and multicamera SLAM and odometry systems - Fig. 1 shows an example of a LiDAR point cloud constructed from just one of the 22 sequences.

To address the challenge of producing BEV ground truth maps, we present a purpose-built simulation environment to match the real-world layout of the car park. From this environment, we can render a digital double — paired simulated and real data — and also detailed top-down semantic maps for BEV prediction. We make the simulator publicly available alongside the dataset and provide tools to render the simulated counterparts of real sequences. Currently, it is able to render RGB images, depth maps, optical flow and BEV maps. Researchers can use these tools to adapt and extend the provided data to meet their own needs.

The dataset is immediately useful for SLAM, odometry, BEV estimation and domain transfer, and the provided tools make it notably more flexible and future-proof than existing offerings.

The main contributions of this work are:

1) The use of multi-view cameras capturing long trajectories with many loops, making the dataset ideal for the evaluation of multi-camera SLAM and odometry systems, or other approaches requiring revisitation.
2) The inclusion of semantic top-down Bird's Eye View maps, rendered using a novel simulation-based approach for accurate and crisp ground truth.
3) The addition of simulation tools and a digital twin, so researchers can add to the dataset and adapt it to their own specific needs.
4) A benchmark evaluation of techniques for SLAM and BEV estimation to support research in autonomous driving and parking.

II. Related Work

There are multiple popular automotive datasets used by the community to develop and evaluate autonomous driving technologies. Early autonomous driving datasets, such as Cityscapes [4], included collections of thousands of images to improve semantic understanding. However, more recent large-scale datasets, such as RobotCar [5] and NuScenes [2] include driving sequences to enable the evaluation of a greater range of systems.

A. SLAM and Visual Odometry

KITTI [3] and Nuscenes [2] are among the most widely used, but have limitations. KITTI was primarily intended for SLAM and odometry development, and therefore includes long trajectories with loops for evaluating such systems. However, it lacks images from multiple cameras, making it unsuitable for the evaluation of modern multi-view SLAM and odometry systems. KITTI's reliance on GPS and IMU for ground truth trajectory creation also makes it susceptible to GPS outages.

Conversely, Nuscenes, which is intended for object detection and tracking, includes multiple cameras and resolves GPS issues by incorporating LiDAR, but contains short 20 s sequences with no loops. Similarly, Lyft Level 5 [6], a smaller autonomous driving dataset aimed at motion prediction, consists of short sequences (25 s). None of these offerings are therefore suited to supporting modern autonomous driving research, where long sequences with multiple loops are essential.

Some datasets do attempt to meet this need. Oxford's Robotcar [5] is one example that breaks the short-sequence trend, but follows the same trajectory repeatedly. This is useful for some tasks, such as generalising over seasonal changes and weather conditions, but not for evaluating visual odometry or SLAM systems, since the trajectory is always identical.

Yang et al. [7] introduce a dataset with a focus on multi-camera SLAM with loops that could also fill this gap, but the focus is on distance driven rather than sequence duration, and the data is not yet available. Lastly, the EU Long-term dataset [8] includes long (approx. 16 min) sequences, but focuses on the inclusion of a variety of sensor types rather than large quantities of data for training learning-based approaches. To our knowledge, there is no modern, large-scale dataset that includes long sequences, loops and multiple sensors and views for evaluation of modern learning-based SLAM and odometry systems.

All of these datasets also focus on general driving; recent commercial interest in autonomous parking has resulted in parking data being used to train localisation systems (such as in [9]), but no parking dataset is publicly available. The only public car park datasets include static overhead views [10], [11], making them unsuitable for SLAM.

B. BEV

The production of top-down Bird's Eye View semantic maps from monocular images has received significant interest in recent years, and geometry-based Inverse Perspective Mapping (IPM) approaches [16], [17] have been superseded by modern learning-based techniques [18]–[20]. Yuexin et al. [21] provide a comprehensive summary of recent advances in vision-based BEV prediction.

However, obtaining suitable ground truth for training these systems presents a significant logistical challenge, and, as a result, very few datasets are available. Typically, BEV systems are evaluated by augmenting existing datasets. For example, Rashed et al. [22] created a moving-object BEV dataset with 12.9k samples by projecting KITTI bounding box annotations

TABLE I
COMPARISON OF PUBLIC DATASETS: NUMBER OF SEQUENCES AND IMAGES, SEQUENCE LENGTHS, AND INCLUDED FEATURES.

	Feature								
Dataset	Seqs	RGB Imgs	Seq Len	LiDAR	Loops	Multi-View	Digi Double	BEV GT	Parking
Cityscapes [4]	N/A	25k	N/A	-	-	-	-	-	-
KITTI [3]	22	15k	2 min 30 s[c]	Y	Y	-	-	-	-
Virtual KITTI [12]	50[a]	21k[a]	2 min 30 s[c]	-	Y	-	Y	-	-
Argoverse[b] [13]	113	490k	15-30 s	Y	-	Y	-	-	-
Lyft Perception [6]	366	323k	60 min	Y	-	Y	-	-	-
Waymo Open [14]	1k	1.4M	20 s	Y	-	Y	-	-	-
nuScenes [2]	1k	1.4M	20 s	Y	-	Y	-	-	-
Ford Multi-AV [15]	18	?	?	Y	Y	Y	-	-	-
RobotCar [5]	100	3M	?	Y	Y	Y	-	-	-
EU Long-Term [8]	13	?	16 min[c]	Y	Y	Y	-	-	-
Campus Map (Ours)	22	2M	avg. 10 min	Y	Y	Y	Y	Y	Y

[a]35 sequences with 17k images at time of original paper publication. [b]3D tracking dataset. [c]Approximate. ? = unpublished.

from the top-down (although this is not currently publicly available), and KITTI now provides a BEV benchmark with 7481 training samples. Roddick [18] provides scripts to produce labels from NuScenes and Argoverse data.

However, this approach suffers from two significant issues. Firstly, it is reliant on the accuracy of bounding box annotation. One of the key motivators of BEV research is the ability to predict occluded or partially-occluded objects, which are very difficult to accurately annotate on image frames. Secondly, this approach limits the types of classes that can be rendered in the BEV semantic map. Vehicles and sidewalks, for example, are easy to annotate, but road markings are significantly more challenging, especially when occlusions are taken into account. As a result, the Caltech Lanes Dataset [23] only contains 1224 images, the Road Marking Dataset [24] has 1443 images, and neither has top-down annotation. BEV ground truth is therefore typically dominated by different vehicle classes, which is useful but does not capture the entire geometry of the scene.

Although perception of other features is clearly essential for future autonomous driving systems, to our knowledge, there is no large-scale BEV dataset available that attempts to tackle these issues. A selection of public datasets that attempt to meet similar needs to ours is presented in Table I, with a comparison of the available data types and quantity.

III. DATASET

The sequences were collected in daylight between August-November 2022, focusing on scenarios that are rare or unavailable in other autonomous driving datasets. Each sequence includes a ground truth trajectory for SLAM and odometry evaluation. A simulation environment was also developed to match the car park area. This simulator is capable of producing paired data, including top-down semantic BEV maps for real images. Approximately 40k BEV maps are provided across 3 different scales, and the simulator code is included to render more at arbitrary scales.

The overall dataset statistics are presented in Table II, demonstrating the scale of the dataset and the types of data available. The presented camera acquisition frequency reflects the hardware-defined rate. The remainder of this section describes the data collection process and the methodology that was used to produce ground truth trajectories, BEV maps and paired simulated data.

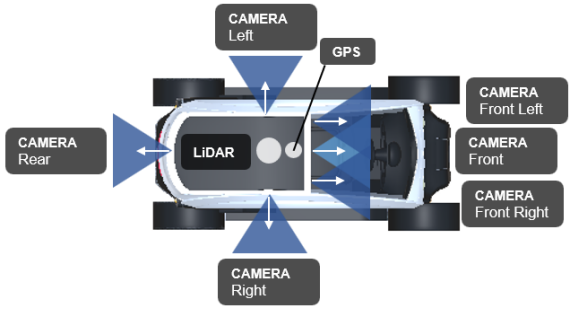

Fig. 2. Sensor configuration around real-world ego vehicle for data collection. An IMU was also present.

A. Vehicle Configuration

The real-world portion of the dataset includes images from 6 Sekonix AR0231 RGB cameras surrounding a Renault Twizy, scans from a 64-beam Ouster OS-1-64 LiDAR, and GPS data from a GlobalSat BU-353S4 5 Hz receiver. IMU data is also included.

Six Sekonix cameras each with 120 degree field-of-view provide 360 video around the vehicle. The front camera is mounted horizontally, whereas the side and back cameras are mounted with an inclination of 45 degrees downward. In addition, a front left and front right camera are configured as a stereo pair, and have a 9 degree inclination. The configuration of sensors around the ego vehicle is outlined schematically in Fig. 2. We use an NVIDIA PX2 Drive to capture and synchronise all data and include rosbags with ROS timestamps in the dataset. IMU data is included in the rosbags.

The simulator model provided with the dataset includes front, left, right and rear cameras with intrinsic and extrinsic calibration to match the real configuration. More cameras can be added to the simulator as required.

B. Camera Calibration

We provide intrinsic and extrinsic calibration matrices for all cameras. Intrinsic calibrations were obtained using a standard checkerboard pattern. Extrinsic calibrations were obtained using fiducial markers and a bundle-adjustment operation with

TABLE II
CAMPUS MAP INITIAL RELEASE STATISTICS.

Data Type	Occurrences	Acquisition Frequency
RGB Images	2 115 741	28.24 Hz
GPS Fixes	12 684	2 Hz - 5 Hz
LiDAR Packets	7 995 033	-
Point Clouds	124 941	-
BEV Maps	40 818	-

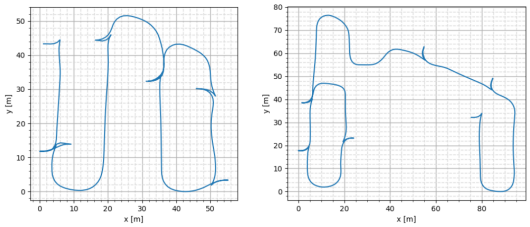

Fig. 3. Examples of two Campus Map car park ground truth trajectories (sequences T and U), showing loops and multiple parking manoeuvres.

COLMAP [25], [26], and verified by overlaying the LiDAR point cloud on the images. These qualitative overlays can be found alongside the dataset. The calibration parameters are provided as YAML files and can be accessed and used via the provided development kit.

C. Trajectory Planning

A key goal of the Campus Map dataset was to provide long sequences with more loops than are available in existing offerings. Trajectories are split into two groups: driving sequences around a car park with multiple parking manoeuvres, and long driving sequences around the university campus. 16 car parking sequences and 6 campus driving sequences are present in the initial release, for a total of 22 trajectories.

All trajectories include multiple loops. Fig. 1 shows a reconstructed LiDAR map of the campus from the sequences. Details of the collected sequences are presented in Table III. Two of the sequences are shown as examples in Fig. 3.

D. Ground Truth Trajectories

The ground truth trajectories are produced using a SLAM algorithm [27] employing LiDAR and GPS data. The optimisation was conducted offline to provide accurate trajectories. We publish the ground-truth trajectories for the LiDAR timestamps. Not that, given the trajectories were produced using LiDAR data, the dataset is not intended for LiDAR odometry evaluation. The ground-truth poses for the cameras can be produced with the development kit we provide.

E. Top-Down Semantic Ground Truth

Bird's Eye View ground truth is challenging to obtain using conventional methods. As previously discussed, most existing work relies on projecting manually-annotated 3D bounding boxes from the top-down. While this can be effective, it relies heavily on the accuracy of the manual annotations, and often means that occluded objects are not shown in the BEV map, forgoing one of the key advantages of BEV over

TABLE III
BREAKDOWN OF CAMPUS MAP SEQUENCES. SPLIT OF CAR PARK AND CAMPUS DATA, SCENARIO, TIME OF DAY AND DURATION.

Sequence	Scenario[a]	Time[b]	Duration
Campus			
A	D	A	9 m 31 s
B	D	A	5 m 35 s
C	D	A	6 m 52 s
D	D	A	10 m 40 s
E	D	A	6 m 44 s
Campus & Car Park			
F	D	M	13 m 5 s
Car Park			
G	D, P	A	49 m 22 s
H	D, P	A	38 m 41 s
I	P	M	6 m 25 s
J	P	M	2 m 30 s
K	P	M	4 m 32 s
L	P	M	5 m 42 s
M	P	M	2 m 2 s
N	P	M	2 m 13 s
O	P	M	5 m 25 s
P	P	M	3 m 12 s
Q	P	M	5 m 25 s
R	D, P	A	9 m 43 s
S	P	A	4 m 59 s
T	P	A	4 m 46 s
U	P	A	5 m 24 s
V	P	A	5 m 21 s
Total			3 h 28 m 8 s

[a]D = Driving, P = Parking. [b]M = Morning, A = Afternoon.

semantic segmentation in the image plane. To avoid these issues, we adopt a novel solution: we developed a virtual car park matching the dimensions of its real counterpart. High-fidelity ground truth at many scales can then be generated very quickly.

Parking spaces and road markings in the virtual car park are positioned according to satellite imagery and considered static. To produce BEV ground truth for real sequences, we must populate the car park according to the real-world arrangement of vehicles and reconstruct the real trajectory in the simulator. It is useful to have automated means to achieve both tasks, as this process must be repeated for each data capture session.

To achieve this, we begin by using hdl-slam [27] to generate a large point cloud for the real-world car park and the trajectory followed by the ego vehicle. Sequence G was chosen for this, as the whole car park is fully covered. We define X_l as poses in the resulting coordinate system such that $X_l \in SE(2)$ represents the trajectory of the vehicle through the car park.

Since the ego vehicle was fitted with a GPS receiver, we have known lat-lon coordinates in Cartesian space X_g. We can then determine the transformation

$$ {}^{g}T_l = \arg\min_{{}^{g}T_l} \sum |X_g - {}^{g}T_l X_l| \tag{1} $$

to map poses X_l to Cartesian GPS coordinates X_g. We then define $x_g, x_s \in \mathbb{R}^2$, a set of manual correspondences between X_g and points in the simulation environment's global coordinate frame X_s, and compute the transformation

$$ {}^{s}T_g = \arg\min_{{}^{s}T_g} \sum |x_s - {}^{s}T_g x_g|. \tag{2} $$

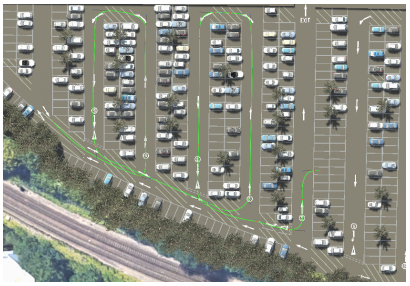

Fig. 4. Top-down view of car park simulator model populated with vehicles in bays matching their real-world counterparts. Real trajectory shown in green.

We can now define $P_l' \in \mathbb{R}^2$ as the points in the simulation coordinate frame such that

$$P_l' = {}^{s}T_g \, {}^{g}T_l \, P_l \tag{3}$$

where P_l are points along the trajectory interpolated to match mounted camera image timestamps. This allows us to recreate the original trajectory in simulation.

To produce paired data, the simulator must also be populated to match the real-world arrangement of parked vehicles, which requires automatically determining which spaces are occupied. To do this, we define a non-rigid transform ${}^{p}T_s$ between points in the simulation environment's global coordinate frame X_s and each parking bay's local coordinate system.

In practice, parking bays vary in length, and we accounted for this by scaling the transformation accordingly, but here we show the unscaled transformation without loss of generality. We can then test each point to determine if it is in a parking bay using

$${}^{p}T_s P_l' < \begin{pmatrix} W \\ D \end{pmatrix} \tag{4}$$

where the width W and depth D of the parking bay are known. A simple point threshold is used to determine which bays are occupied, and the simulation environment populated accordingly.

By moving the virtual ego vehicle to each point in the trajectory in turn, we can create a digital double of the real data. This includes mounted camera images and top-down BEV ground truth. The simulator is implemented in the Unity engine, and code is provided, including tools to populate the car park both automatically and manually as well as load trajectories or plot them by hand. Additional implementation details and suggestions for extensions and modifications are available in a readme file with the dataset. The advantage of this novel approach to BEV ground truth production is that maps are correct, temporally consistent and can include smaller classes that would be difficult to annotate (such as road markings).

BEV ground truth is provided at 3 different scales to test BEV prediction at different ranges. The provided scales are short (0.04 m/pixel), medium (0.08 m/pixel) and long (0.12 m/pixel). All maps are rendered at 200x200 resolution, so the perception distances are 8 m, 16 m and 24 m respectively. Approximately 40k of these maps are provided, split into a 36k training set and 4k validation. Five semantic classes are rendered by default (although this can be changed with the provided tools). A "null" class is also added to encapsulate any area that falls outside of the simulation environment; this is used when the vehicle is positioned close to the edge of the car park. It is suggested that this class should not be supervised when training BEV predictors.

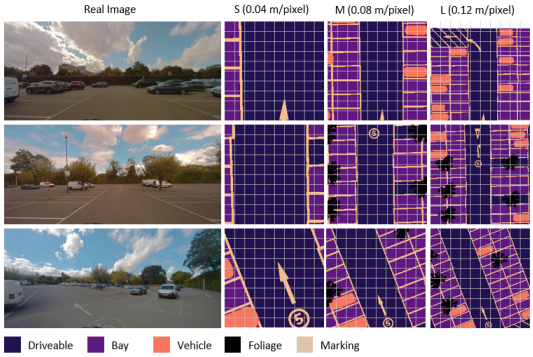

Fig. 5. Examples of BEV semantic maps rendered from simulation at multiple scales and paired with real-world car park images. Additional scales can be rendered using the provided simulation kit.

TABLE IV
COVERAGE OF LABELS INCLUDED IN BEV GROUND TRUTH (PERCENTAGE OF SEMANTIC MAPS COVERED FOR EACH PROVIDED SCALE).

Class	Percentage across split (%)					
	Training			Validation		
	S	M	L	S	M	L
Driveable	71.4	39.3	30.8	76.7	42.4	33.2
Bay	14.9	32.1	35.3	12.1	32.6	36.0
Vehicle	3.2	7.7	8.4	1.4	5.4	6.0
Foliage	0.2	6.2	7.7	0.1	3.8	4.1
Marking	7.7	9.4	9.4	6.4	8.8	9.1
Null	2.6	5.3	8.4	3.3	7.0	11.6

A top-down view of the simulation environment is shown in Fig. 4. See Fig. 5 for example semantic maps. Table IV shows the percentage of the dataset occupied by each semantic label across the different splits. Notice that the percentage of driveable area decreases with larger scales; this is because the area immediately around the vehicle is typically driveable. The validation statistics also approximately match the training statistics, justifying the choice of split.

F. Development Kit

We provide a Python Development Kit for easy downloading, pre-processing and manipulation of the dataset, accessing calibration data and processing the raw sensor data. Details are included in the dataset's readme file.

G. Release

The dataset and Python Development Kit are available at cvssp.org/data/twizy_data. Rosbags, split into sequences, are available for those interested in using, for example, raw LiDAR packets and IMU data. For ease of use, RGB images have been extracted, undistorted and stored in databases, and Python tools are provided to extract them. Ground truth trajectories are also provided.

TABLE V
SLAM AND VISUAL ODOMETRY BASELINES FOR CAMPUS MAP FOR SELECTED SEQUENCES.

	Seq A		Seq B		Seq C		Seq D		Seq E	
	RPE	ATE	RPE	ATE	RPE	ATE	RPE	ATE	RPE	ATE
ORB-SLAM 2 [28]	4.199	142.979	0.818	1.632	1.253	154.813	0.742	78.638	0.790	31.305
ORB-SLAM 3 [29]	0.718	64.567	1.387	92.127	1.232	98.876	1.095	82.355	0.547	186.601
COLMAP [25], [26]	1.620	148.98	2.834	51.780	2.095	95.886	1.354	78.637	1.393	70.356
PySLAM [30]	0.381	18.214	1.706	75.070	0.958	8.076	1.282	33.603	1.264	95.549
Single-Cam BEV-SLAM [31]	0.826	28.041	1.800	76.329	1.477	15.946	1.128	38.994	1.451	70.220
Multi-Cam BEV-SLAM [31]	0.403	21.366	1.657	57.630	1.012	8.065	0.889	33.211	0.947	56.931

TABLE VI
BASELINE PER-CLASS IOU(%) ON BEV VALIDATION SPLIT AT MULTIPLE SCALES. S = 8 M, M = 16 M, L = 24 M PERCEPTION DISTANCE.

	Driveable			Bay			Vehicle			Foliage			Marking			Mean		
	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
PON [18]	56.7	58.9	60.2	23.3	32.4	24.2	12.4	18.3	16.9	6.8	7.1	6.9	18.4	26.3	24.1	23.5	28.6	26.5
Saha [1]	69.5	69.9	70.1	27.1	37.4	33.6	8.6	23.7	18.7	6.9	7.6	7.0	19.9	30.6	23.0	26.4	33.8	30.5

The simulation environment is available separately at gitlab.surrey.ac.uk/cogvis/automotive-sim, and can be used to produce additional semantic segmentation, depth maps etc., and to render entirely new sequences.

IV. BASELINES

To improve the usefulness of the dataset, we present baselines for SLAM and BEV estimation tasks.

A. SLAM & Visual Odometry

We use ORB-SLAM [28], [29] to produce SLAM and visual odometry baselines for Campus Map. In addition, we produce trajectories using COLMAP [25], [26] and PySLAM [30]. Front-camera images are used in the experiments. BEV-SLAM [31] is also evaluated to demonstrate the utility of including BEV maps alongside trajectories. The medium (16 m) BEV scale was used for this evaluation. We use the *evo* evaluation toolkit [32] and calculate the RMSE of the Relative Pose Error (RPE) and Absolute Trajectory Error (ATE) to evaluate the performance of the baseline methods.

B. BEV

For the Bird's Eye View prediction tasks, we train two learning-based BEV predictors: a Pyramid Occupancy Network [18] and a spatial transformer network [1]. We train for 120 epochs and present results on the BEV validation split for the three scales provided in the initial release of the dataset.

V. RESULTS

The results of the baseline methods are shown in Table V. ORB-SLAM performs well for sections of the sequences; however, it frequently loses track and creates a new map from scratch. In general, ORB-SLAM3 outperforms ORB-SLAM2, which can be expected. PySLAM and COLMAP offer additional baselines with comparable performance.

A. BEV

We evaluate the two baselines across the 3 BEV validation splits provided in the dataset: S (0.04 m/pixel), M (0.08 m/pixel) and L (0.12 m/pixel). The per-class IoUs are presented in Table VI. We observe that the approaches perform comparably to results on the NuScenes validation split, with some improvement; we suggest that this is due to the accuracy of the BEV maps produced by our novel simulation method, rather than the use of bounding boxes projected from the top-down. We also observe that the networks perform best at the medium scale. We suggest this is due to lack of context at the small scale, and a greater number of occlusions at the large scale, making the BEV prediction task challenging.

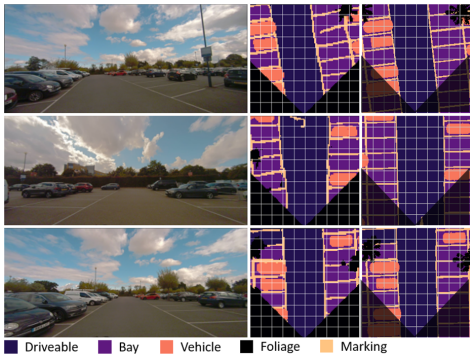

Fig. 6. Qualitative BEV validation results on the Campus Map dataset. Input images, prediction (logits > 0.5) and ground truth.

Lastly, we present a selection of qualitative examples of BEV prediction on the Campus Map dataset in Fig. 6. Notice the benefit to overall scene understanding and qualitative map quality obtained by predicting road markings in addition to vehicle placements.

VI. CONCLUSION

We have introduced Campus Map, a new dataset with a focus on SLAM, Visual Odometry and BEV prediction. Its long trajectories with multiple loops and parking sequences are unique, and we have adopted a novel approach to produce BEV ground truth to match real-world images, resulting in more accurate and higher-fidelity BEV maps than are available in other datasets.

We hope that the introduction of this dataset and related tools will encourage research in learning-based navigation systems for autonomous driving, and we look forward to seeing the dataset used by the community.

REFERENCES

[1] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 9200–9206, 2022.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, pp. 11618–11628, 2020.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, pp. 3354–3361, 2012.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, pp. 3213–3223, 2016.

[5] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, 11 2016.

[6] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Level 5 perception dataset 2020." https://level-5.global/level5/data/, 2019. Accessed: 2023-01-07.

[7] A. J. Yang, C. Cui, I. A. Bârsan, R. Urtasun, and S. Wang, "Asynchronous multi-view slam," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 5669–5676, 2021.

[8] Z. Yan, L. Sun, T. Krajník, and Y. Ruichek, "Eu long-term dataset with multiple sensors for autonomous driving," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 10697–10704, 2020.

[9] O. Mendez, M. Vowels, and R. Bowden, "Improving robot localisation by ignoring visual distraction," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 3549–3554, 2021.

[10] M. Marek, "Image-based parking space occupancy classification: Dataset and baseline," 07 2021.

[11] M.-R. Hsieh, Y.-L. Lin, and W. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *ICCV*, pp. 4165–4173, 10 2017.

[12] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, pp. 4340–4349, 2016.

[13] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, pp. 8740–8749, 2019.

[14] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, pp. 2443–2451, 2020.

[15] S. Agarwal, A. Vora, G. Pandey, W. Williams, H. Kourous, and J. McBride, "Ford multi-av seasonal dataset," *The International Journal of Robotics Research*, vol. 39, pp. 1367–1376, 09 2020.

[16] S. Sengupta, P. Sturgess, L. Ladický, and P. H. S. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 857–862, 2012.

[17] S. Abbas and A. Zisserman, "A geometric approach to obtain a bird's eye view from an image," in *IEEE/CVF Int. Conf. Computer Vision Workshop (ICCVW)*, pp. 4095–4104, 10 2019.

[18] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *CVPR*, pp. 11135–11144, 06 2020.

[19] C. Lu, G. Dubbelman, and M. Molengraft, "Monocular semantic occupancy grid mapping with convolutional variational auto-encoders," *IEEE Robotics and Automation Letters*, vol. PP, 04 2018.

[20] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, pp. 4867–4873, 07 2020.

[21] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, and X. Zhu, "Vision-centric bev perception: A survey," 2022.

[22] H. Rashed, M. Essam, M. Mohamed, A. Ei Sallab, and S. Yogamani, "Bev-modnet: Monocular camera based bird's eye view moving object detection for autonomous driving," in *IEEE Int. Intelligent Transportation Systems Conference (ITSC)*, pp. 1503–1508, 2021.

[23] M. Aly, "Real time detection of lane markers in urban streets," in *2008 IEEE Intelligent Vehicles Symposium*, pp. 7–12, 2008.

[24] T. Wu and A. Ranganathan, "A practical system for road marking detection and recognition," in *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 25–30, 06 2012.

[25] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, pp. 4104–4113, 2016.

[26] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *ECCV*, pp. 501–518, Springer, 2016.

[27] K. Koide, J. Miura, and E. Menegatti, "A portable three-dimensional lidar-based system for long-term and wide-area people behavior measurement," *International Journal of Advanced Robotic Systems*, vol. 16, 02 2019.

[28] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[29] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[30] L. Freda, "https://github.com/luigifreda/pyslam," 2020.

[31] J. Ross, O. Mendez, A. Saha, M. Johnson, and R. Bowden, "Bev-slam: Building a globally-consistent world map using monocular vision," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3830–3836, 2022.

[32] M. Grupp, "evo: Python package for the evaluation of odometry and slam.." https://github.com/MichaelGrupp/evo, 2017.