MCD: Diverse Large-Scale Multi-Campus Dataset for Robot Perception

Thien-Minh Nguyen[1]✉, Shenghai Yuan[1]✉, Thien Hoang Nguyen[1], Pengyu Yin[1], Haozhi Cao[1], Lihua Xie[1], Maciej Wozniak[2], Patric Jensfelt[2], Marko Thiel[3], Justin Ziegenbein[3], Noel Blunder[3]

[1] School of EEE, NTU, Singapore, [2] Division of RPL, KTH, Sweden, [3] ITL, TUHH, Germany

mcdviral.github.io

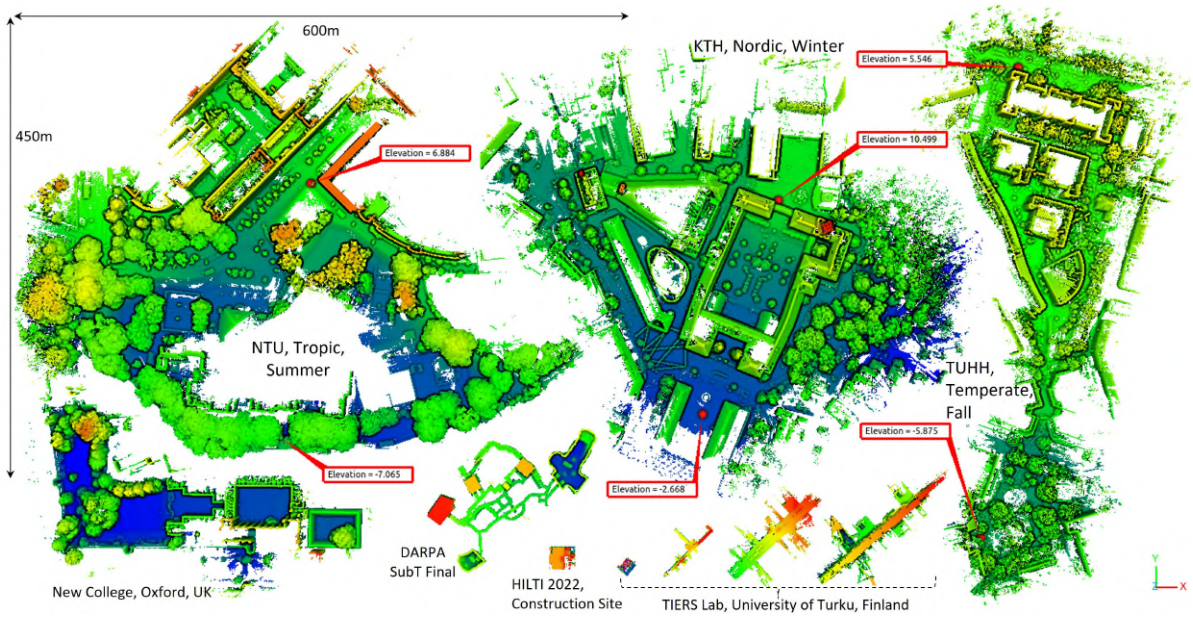

Figure 1. Up-to-scale comparison of survey-grade prior maps in MCD with other works.

Abstract

Perception plays a crucial role in various robot applications. However, existing well-annotated datasets are biased towards autonomous driving scenarios, while unlabelled SLAM datasets are quickly over-fitted, and often lack environment and domain variations. To expand the frontier of these fields, we introduce a comprehensive dataset named MCD (Multi-Campus Dataset), featuring a wide range of sensing modalities, high-accuracy ground truth, and diverse challenging environments across three Eurasian university campuses. MCD comprises both CCS (Classical Cylindrical Spinning) and NRE (Non-Repetitive Epicyclic) lidars, high-quality IMUs (Inertial Measurement Units), cameras, and UWB (Ultra-WideBand) sensors. Furthermore, in a pioneering effort, we introduce semantic annotations of 29 classes over 59k sparse NRE lidar scans across three domains, thus providing a novel challenge to existing semantic segmentation research upon this largely unexplored modality. Finally, we propose, for the first time to the best of our knowledge, continuous-time ground truth based on optimization-based registration of lidar-inertial data on three survey-grade prior maps, each several times larger than the next largest publicly available ones. We conduct a rigorous evaluation of numerous state-of-the-art algorithms on MCD, report their performance, and highlight the challenges awaiting solutions from the research community.

1. Introduction

Understanding the world and estimating egomotion are fundamental in computer vision and robotics, enhancing applications such as autonomous driving, logistics, delivery, and AR/VR (Augmented/Virtual Reality). Over the years, public datasets have played a crucial role in promoting broader

participation in the research of environment perception and egomotion estimation, especially for those with limited resources. However, we find that existing annotated multi-modality large-scale datasets are biased toward autonomous driving scenarios, and rely heavily on costly classical lidars and cameras. These datasets are also carefully curated to isolate users from some real-world issues. For instance, cameras are often collected in conducive lighting conditions and with little background noises, while lidar point clouds are often pre-processed to remove distortion. These ideal benchmarks are one of the main reasons causing the performance gap between the lab and real-world scenarios.

Recently, newer modalities for egomotion estimation and environment perception have surfaced, featuring low-cost NRE (Non-Repetitive Epicyclic) lidar and UWB (Ultra-WideBand) technology. These innovations address issues like cost, illumination, privacy, and robustness, but they also present new challenges, such as scan sparsity and limited field of view of NRE lidar, and NLOS (Non-Line-Of-Sight) observations by UWB. These factors complicate environment perception tasks, but also open up new research opportunities. To facilitate this research, we introduce MCD, a multi-campus dataset designed to address various robotics perception challenges. Our primary goal is to foster the development of next-generation, cost-effective robots, AI (Artificial Intelligence), and AR/VR systems capable of working in diverse environments, not limited to public roads. *The MCD boasts an extensive collection of 18 sequences, over 200k lidar scans, 1500k camera frames, and high-frequency IMU (Inertial Measurement Unit) and UWB data*. Notably, *the dataset includes point-wise annotations for typical outdoor and indoor objects*, making it one of the pioneering works to provide such annotations for affordable NRE lidar systems. Moreover, we also propose a novel *continuous-time ground truth data*, which demonstrates superior accuracy when compared to other datasets of similar scales. In summary, the contributions and features of MCD are listed as follows:

Multiple Sensing Modalities. MCD includes both classical rotating lidar and newer MEMS NRE lidar, as well as traditional cameras with different baselines. Additionally, it incorporates useful modalities like IMUs and UWBs, which were not present in widely used benchmarks like KITTI [14] [2], EuRoC [3], Newer College [34]. The manual for using these modalities can be found at `mcdviral.github.io/UserManual.html`.

Pointwise Annotations for NRE Points Clouds. Low-cost, accurate NRE lidars have gained attention but are underexplored in semantic segmentation. Their sparsity and irregular scanning pose challenges for both learning methods and human annotators, as depicted in Fig. 2. Despite this, we have carefully annotated 59k lidar scans, surpassing SemanticKITTI [2] and Nuscenes [4]. MCD is the first extensive dataset with semantic annotations for NRE lidars. The NRE scans and preview of annotated sequences can be found at `mcdviral.github.io/AnnotatedLidar`.

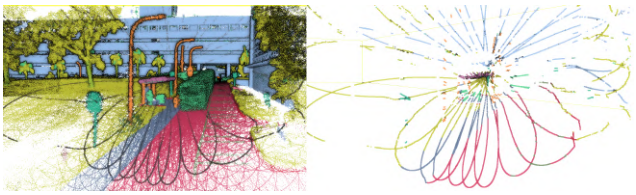

Figure 2. Concatenation of annotated NRE lidar scans in motion over a 20s period (left) and a single scan (right). This modality remains untested in semantic segmentation research to the best of our knowledge.

Wider Domain Coverage. We collect data from three university campuses across Eurasia, covering a broader latitude range than most autonomous driving datasets [14] [2] [6] [4] [39], as shown in Fig. 3. This diversity leads to notable variations in feature prior distribution, providing challenges for general robot learning and comprehension.

Continuous-time Ground Truth. Accurate ground truth is vital for SLAM and localization studies, especially in larger environments. Existing datasets offer only discrete-time ground truth poses or processed lidar scans to reduce motion distortion, limiting research possibilities. To address these issues, we propose continuous-time ground truth based on survey-grade prior map registration. This continuous-time ground truth allows arbitrary time and density sampling, which can be used to directly undistort the lidar point clouds. Detailed explanations and usage of ground truth can be found at `mcdviral.github.io/Groundtruth`.

Embrace the Challenges in Perception. We believe understanding motion distortion is critical in real-world end-to-end training scenarios and should not be obscured, as it has significant implications for research. Researchers can choose to retain or remove motion distortion from raw lidar point clouds, as shown in Fig. 4. MCD also tackles other challenges like extreme lighting, glass reflection, and solar interference, as illustrated in Fig. 5, which have led to real-world accidents due to developers overlooking them. However, these issues are often inadequately addressed in existing datasets. We incorporate those noise classes in annotations and images as part of our efforts to train robotics systems to handle such corner cases.

Extensive Benchmarks. To facilitate and advance future research endeavors, we provide the dataset, data-loading scripts, and benchmarking instructions for both semantic segmentation and SLAM studies, providing in-depth understanding and analysis of each method. These benchmarks highlight the considerable room for improvement and the opportunities for innovation in each field. The instructions can be found at the SLAM and Semantic Segmentation Tutorials at `mcdviral.github.io`.

Figure 3. Example camera views at different campuses in MCD, with various terrains, backgrounds and lighting conditions. All faces are anonymized in compliance with local regulations.

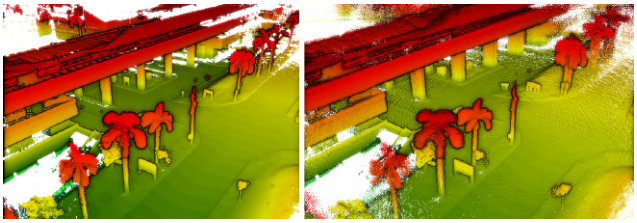

Figure 4. The lidar point clouds with (left) and without (right) motion undistortion.

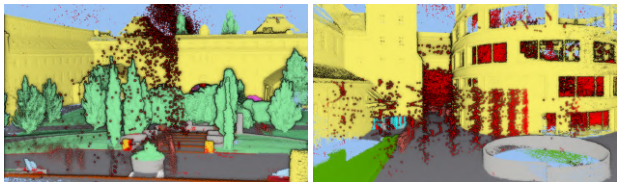

Figure 5. The sun glare and reflection outliers in the lidar scans are uncommon in lidar datasets. Significant effort were needed to segment out these outlier points.

2. Related Works

The MCD draws inspiration and insights from several preceding works, each representing a unique milestone in developing autonomous systems. A comprehensive table that compares existing multi-modal perception datasets with MCD can be found in the supplementary material. Below, we shall provide a short synopsis of the evolution of multi-modal datasets in the last decade.

Released more than a decade ago, the KITTI dataset [13, 14] stands out as a classic and is widely recognized in autonomous driving research. The KITTI dataset and its subsequent publications have enabled the evolution of numerous studies in areas such as stereo matching, optical flow [24], depth perception [41], object detection [12], tracking [22], and semantic segmentation [1]. Later, SemanticKITTI [2] was introduced to offer semantic annotations for all sequences within the KITTI Odometry Benchmark. However, the key limitation of the KITTI dataset lies in the absence of IMU data and low diversity in terms of environment and terrain, thereby impeding comprehensive research in domains like domain adaptation and inertial-based SLAM. Numerous autonomous driving datasets, such as Cityscapes [9], Urban Loco [45], NuScene [4], Waymo [39], ApolloScape [17], A*3D [31] and H3D [29], have emerged recently, each offering data across various locations. These datasets predominantly focus on road-centric driver views, thus lacking diversity in the environment. They often fail to deliver millimeter-level ground truth precision due to GPS/INS fusion limitations.

NCLT [6] aims to enhance diversity by providing a long-term, domain-shifted dataset that extends over a year. Nonetheless, this dataset still has challenges related to ground truth accuracy and is restricted to a single location. The challenge of geographical limitation is a common problem in dataset development. For example, KAIST Urban [18], TorontoCity [43], and Oxford RobotCar [23], also have similar limitations, operating exclusively within one city and lack millimeter-level ground truth.

To address perception and scene understanding in a general environment with higher ground truth accuracy, the Newer College Dataset [34] was introduced, utilizing a ground truth constructed from scanned survey maps for both indoor and outdoor scenes. However, the ground truth data contain noticeable errors due to failures in the ICP (Iterative Closest Point) matching of the lidar scan with the prior map. This issue primarily arose due to static maps and moving scans collected at different times of the year, resulting in noticeable differences in vegetation growth. Additionally, this dataset lacks ground truth in semantic labels and covers very small areas, significantly limiting its impact on AI and SLAM research. The Hilti SLAM Challenges [16, 47] also employ a comparable method for ground truth generation. However, their keyframe poses are prone to overfitting when using reinforcement learning-like hyperparameter tuning [21] for scoring. Additionally, these datasets do not include any annotations.

Most well-established datasets struggle with providing high-accuracy position ground truth, and only a few studies attempt to address these issues. The UAV flying dataset named NTU VIRAL [27] aims to overcome this accuracy limitation by incorporating two lidars, stereo vision, inertial sensors, and UWB on a single drone, supplemented with survey-grade tracking tools. This approach draws significant inspiration from the EUROC dataset [3]. However, these datasets still suffer from limited domain coverage and lack of annotations. Additionally, they require a line of sight

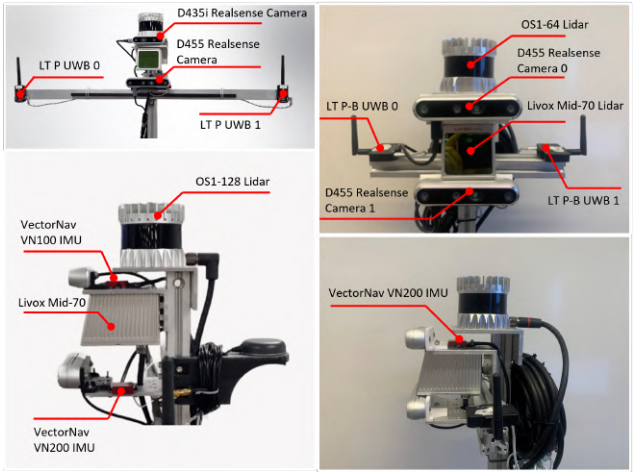

Figure 6. Sensing modalities on the sensor suites: ATV (left) and HHS (right).

to ground survey measuring stations, which makes them unsuitable for large-scale dataset applications.

Simulated datasets such as CARLA [11] and TartanAir [44] offer a wide range of environments along with absolute ground truth data. However, they exhibit a noticeable perceptual domain shift due to their simulator models and are limited by their environmental settings. Furthermore, their inability to completely replicate real-world physics and dynamics, such as sun glare and mirror reflections, often limits their relevance in practical real-world scenarios.

3. Dataset Statistical Analysis

3.1. Sequence Characteristics

The MCD dataset consists of 18 sequences, with 6 sequences from each campus. Within each campus, 3 sequences are captured during the day and 3 at night. Tab. 1 lists the names of these sequences and some of their characteristics. Visualization of the paths can be viewed at `mcdviral.github.io/Download`.

The route of each sequence is decided with deliberation. For each campus, we start by collecting one sequence on a long path covering most of the environment, so-called the main sequences. Then, we collect two shorter sequences that largely overlap with parts of the main sequence, though each still features some exclusive tracks. This arrangement facilitates researchers in analyzing the limitations of SLAM methods within the visual and geometrical contexts of the environment. Moreover, we collect another set of sequences in the nighttime, mainly following the same path as daytime sequences. Hence, a wide range of algorithms can be evaluated for their robustness in the context of visual methods under challenging lighting conditions.

Fig. 7 shows the velocity histograms of images captured under a camera topic across three main sequences.

Table 1. Sequence names, their duration (Dur.), traversed distance (Dist.), maximum and median velocity (max(V) and med(V)). The sequences in **bold** have their livox point clouds annotated.

Seq. Name	Dur. (mm:ss)	Dist. (m)	max(V) (km/h)	med(V) (km/h)
ntu_day_01	10:02	3198	36.70	18.69
ntu_day_02	03:49	642	21.02	11.37
ntu_day_10	05:25	1783	33.05	20.02
ntu_night_04	04:56	1459	27.18	18.27
ntu_night_08	07:47	2421	34.38	19.01
ntu_night_13	03:54	1231	29.40	20.27
kth_day_06	14:51	1403	7.54	5.82
kth_day_09	12:47	1076	7.82	5.25
kth_day_10	10:15	920	8.12	5.82
kth_night_01	16:09	1416	7.53	5.49
kth_night_04	12:26	1052	7.15	5.51
kth_night_05	11:05	919	7.37	5.52
tuhh_day_02	08:20	749	8.02	5.95
tuhh_day_03	13:59	1137	6.51	5.05
tuhh_day_04	03:08	297	7.77	6.32
tuhh_night_07	07:24	742	8.55	6.57
tuhh_night_08	11:49	1128	7.73	6.10
tuhh_night_09	03:05	290	7.66	6.21

The maximum velocity can approach 10m/s in the NTU sequences, which is almost five times that of the KTH and TUHH sequences. This difference is due to NTU sequences being collected from the ATV (All Terrian Vehicle) and the others from Handheld Setup (HHS), as shown in Fig. 6. These velocity profiles will allow users to test their algorithms against low or high-speed conditions. This issue is also discussed in Sec. 5.1

3.2. Semantic Annotations

A total of 29 semantic classes of 6 groups were applied on the Livox point clouds of 11 sequences. From Fig. 8, we can notice the domain shift of the three environments. For example, the number of points in poles, info-sign, vegetation, and buildings is roughly the same among the campuses. However, the data points for chairs and trash bins in KTH and TUHH are an order of magnitude larger than those in the NTU campus. This discrepancy is primarily attributed to the geographical and climate differences. NTU, being situated in a low-latitude region with a tropical climate, has fewer people hanging outdoors, in contrast to the other two campuses located in mid to high-latitude areas where the weather is more temperate. Additionally, NTU has more hydrants, likely influenced by the same climatic factors. The scarcity of bikes and riders in NTU is due

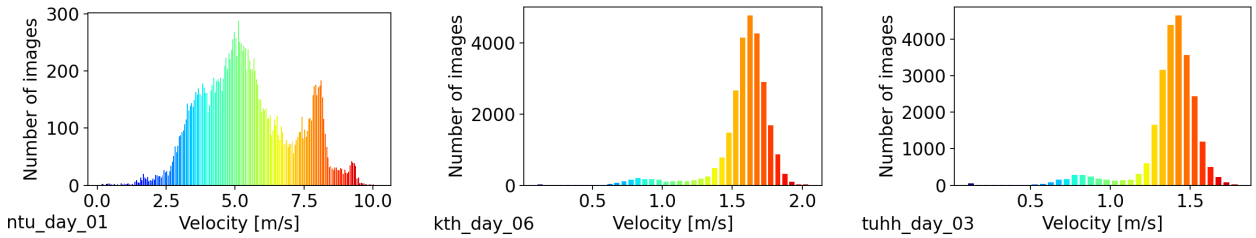

Figure 7. Velocity histograms of the images under topic /d455b/color/image_raw in the three main sequences.

to the combination of its harsh tropical weather and legal restrictions on e-scooters. Similar contrasts are observed in the structure group (stairs vs barriers) due to different styles of structures. The traffic cone reflects the different levels of construction, whereas the road/sidewalk ratio also reflects urban planning priorities. Differences in lane markings imply varying policies on pedestrian and vehicle regulations among the three campuses, with the Asia campus being more stringent and the Europe two campus being more lenient on traffic rules. While parking lot point distribution is similar, the Asia campus car park is more centralized, while the Europe campus car park is more distributed along the roadside. These domain shifts lead to unique lidar point cloud prior distributions across the campuses. Ignoring prior differences can impact prediction accuracy significantly, which is discussed in Sec. 5.4.

4. Survey Map Continuous-Time Registration

In previous works, such as the Newer College Dataset (NCD) [34], Multi-lidar Benchmark [37], and Hilti-SLAM-Challenge[47], simple ICP matching of SLAM-based undistorted point clouds with a prior map was used to provide discrete-time ground truth. However, we notice that these point clouds were generated through imperfect SLAM processes, introducing an inherent observational bias. This bias can then propagate down the line. In the case of NCD [34], we can identify instances of ground truth errors, primarily stemming from ICP matching failures in certain sequences. These failures arise due to the six-month time gap between the prior map data collection and the sequence data collection, resulting in observable discrepancies in the scanned data consistency, largely attributable to vegetation growth. In contrast, MCD data collection at each site spans less than one month, contributing to better ground truth consistency. Unlike the simple discrete-time ICP process, our method directly registers the raw point cloud with the prior map through a continuous-time optimization scheme that explicitly considers the deskew process during optimization. The cost function consists of pose priors, lidar-matching, and IMU factors as follows

$$f\left(\{\mathcal{Z}_{t_n}\}, \hat{\mathbf{T}}_t, \hat{\mathbf{b}}\right) = \sum f_{\text{pose}} + \sum f_{\text{lidar}} + \sum f_{\text{gyro}} + \sum f_{\text{acce}}, \quad (1)$$

where $\{\mathcal{Z}_{t_n}\}$ is the set of all lidar and inertial measurements, $\hat{\mathbf{T}}_t$ is the continuous time trajectory estimate based on B-spline formulation [38], $\hat{\mathbf{b}}$ is the IMU bias estimate. The individual factors in (1) are defined as

$$\begin{aligned}
f_{\text{pose}} &= \left\| \text{Log}(\bar{\mathbf{R}}_{t_k}^{-1} \hat{\mathbf{R}}(t_k)) \right\|_{\text{G}_\mathbf{R}}^2 + \left\| \hat{\mathbf{p}}(t_k) - \bar{\mathbf{p}}_{t_k} \right\|_{\text{G}_\mathbf{p}}^2, \\
f_{\text{lidar}} &= \left\| \mathbf{n}_{\text{pm}}^\top [\hat{\mathbf{R}}(t_j)\, {}^{t_i}\mathbf{f} + \hat{\mathbf{p}}(t_i)] - \mu_{\text{pm}} \right\|_{\text{G}_{\text{lidar}}}^2, \\
f_{\text{gyro}} &= \left\| \hat{\mathbf{R}}^{-1}(t_j)\, {}^{\text{W}}\hat{\omega}(t_j) + \hat{\mathbf{b}}_{\text{gyro}} - {}^{\text{B}_{t_j}}\breve{\omega} \right\|_{\text{G}_{\text{gyro}}}^2, \\
f_{\text{acce}} &= \left\| \hat{\mathbf{R}}^{-1}(t_j) \left[{}^{\text{W}}\hat{a}(t_j) + \mathbf{g} \right] + \hat{\mathbf{b}}_{\text{acce}} - {}^{\text{B}_{t_j}}\breve{a} \right\|_{\text{G}_{\text{acce}}}^2,
\end{aligned}$$

where $\text{G}_\mathbf{R}$, $\text{G}_\mathbf{p}$, G_{lidar}, G_{gyro}, G_{acce} are the covariance of the measurement error, $\mathbf{n}_{\text{pm}}$ and μ_{pm} are the coefficients of the plane contained within the voxel associated with the lidar point ${}^{t_i}\mathbf{f}$, $\mathbf{g}$ is the gravity constant vector, and ${}^{\text{W}}\hat{\omega}(t_j)$, ${}^{\text{W}}\hat{a}(t_j)$ are the angular velocity and acceleration estimates, whereas ${}^{\text{B}_{t_j}}\hat{\omega}$, ${}^{\text{B}_{t_j}}\hat{a}$ are the angular velocity and acceleration measurements themselves.

5. Benchmarking on SOTA algorithms

In this section, we conduct experiments on SOTA (state-of-the-art) SLAM methods with MCD to demonstrate its complexity as well as find out the capability of these methods.

5.1. Lidar-Inertial SLAM

Four SOTA open source lidar-inertial odometry methods of different paradigms are tested against MCD: LIO-SAM [36], Fast-LIO [46], DLIO [7], and SLICT [28]. LIO-SAM is a loosely coupled method that fuses lidar odometry and IMU data over GTSAM factor graph, with traditional plane-edge geometric features and ICP-based loop closure. Fast-LIO is a filter-based method that uses direct feature association, and incremental global map based on the i-kd tree.

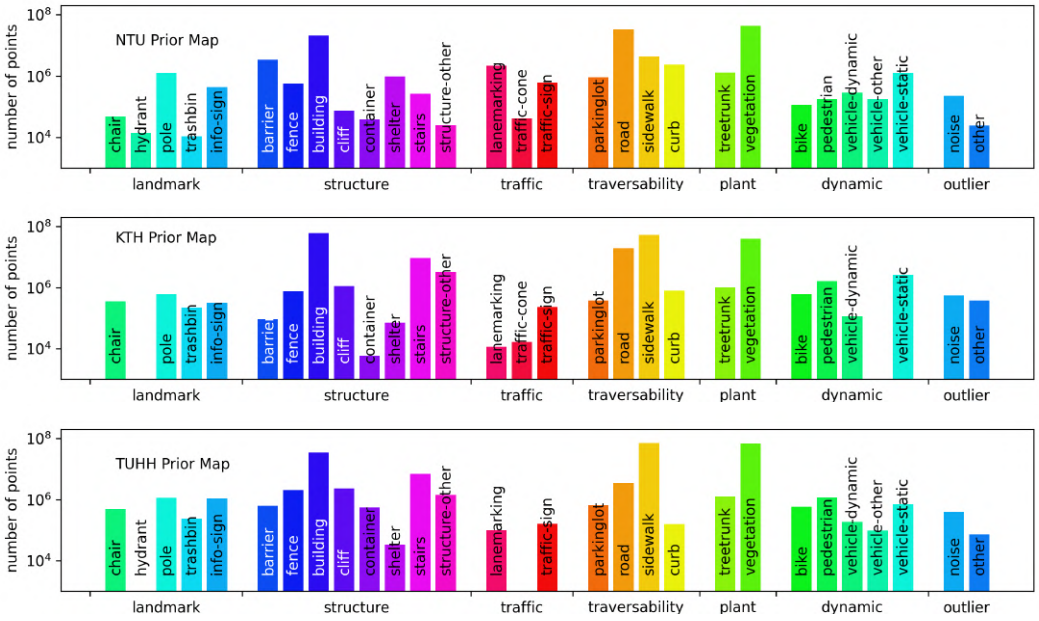

Figure 8. Distribution of the semantic classes in the three campuses.

Table 2. ATE* of SOTA SLAM methods.

sequence	L.SAM	F.LIO	DLIO	SLICT	VINS	KMR	OVINS	RO	VRO
ntu_day_01	6.658	**1.510**	1.925	1.890	15.510	x	**10.605**	14.965	**9.430**
ntu_day_02	0.227	0.272	0.636	**0.168**	4.373	x	**1.683**	9.175	**3.408**
ntu_day_10	2.845	2.084	3.052	**1.429**	**16.021**	x	17.604	14.156	**7.360**
ntu_night_04	1.134	1.599	2.373	**1.002**	**5.262**	x	5.890	25.584	**13.628**
ntu_night_08	3.672	1.425	2.056	**0.822**	x	x	**13.662**	**12.993**	x
ntu_night_13	0.923	0.903	1.928	**0.574**	x	x	**21.794**	**10.777**	49.067
kth_day_06	**0.524**	1.005	0.562	0.633	**20.045**	83.393	72.445	30.771	**15.490**
kth_day_09	**0.235**	0.733	0.326	0.262	x	x	**22.494**	26.740	**17.366**
kth_day_10	**0.255**	2.176	0.665	0.737	21.022	17.092	**10.510**	26.691	**11.550**
kth_night_01	12.240	1.040	**0.414**	0.540	x	x	x	**27.284**	42.581
kth_night_04	**0.234**	0.567	0.376	0.441	66.256	**14.947**	x	27.678	**24.719**
kth_night_05	**0.298**	2.158	0.903	0.855	x	x	x	**30.421**	50.539
tuhh_day_02	0.256	0.273	0.283	**0.236**	16.349	**3.735**	6.621	30.846	**9.268**
tuhh_day_03	1.134	0.970	**0.731**	0.743	28.103	**2.180**	8.681	45.054	**20.430**
tuhh_day_04	0.142	**0.077**	0.232	0.084	6.505	**2.583**	2.698	6.599	**0.903**
tuhh_night_07	40.566	0.279	0.436	**0.227**	x	x	x	46.180	**4.085**
tuhh_night_08	x	0.749	**0.685**	0.740	x	x	x	54.389	**15.067**
tuhh_night_09	0.103	**0.057**	0.375	0.094	x	**4.848**	x	**5.028**	8.429

*All values are in meters. The best results are in **bold**, and the second best are underlined. 'x' denotes a divergent result. L.SAM, F.LIO, KMR and OVINS are shorthands for LIO-SAM, Fast-LIO, KIMERA and OPENVINS.

DLIO is also a direct method with a key-frame-based mapping scheme. Finally, SLICT is a piecewise continuous-time optimization-based method, which uses multi-scale association and incremental global surfel map. Examples and code samples of these experiments can be found on the dataset's website.

Table 2 reports the Absolute Trajectory Error (ATE) of each method. We first notice that at least half of the sequences have an ATE above 0.5m for any method, and the ATE can vary largely even for the sequences of the same campus. Second, with LIO-SAM, we find it performs quite well for the HSS data (the KTH and TUHH sequences). This can be attributed to the use of an external orientation estimate from the expensive VectorNav IMU, which can be quite accurate at walking speed. However, this advantage is lost in the high-speed sequences (ATV). Compared to LIO-SAM, Fast-LIO gives a more consistent accuracy, thanks to the robust IMU propagation and global map association, at the cost of lower accuracy due to the association strategy not leveraging edge features. DLIO has similar accuracy as Fast-LIO, with a slightly larger computational footprint. Finally, SLICT appears to perform well and consistently in both HHS and ATV bags. This can be attributed to the continuous-time formulation, which can take care of higher speed, and the point-to-surfel association strategy on a multi-scale global map. However, SLICT can be quite computationally expensive and cannot run at high-speed lidar-rate like LIO-SAM and Fast-LIO.

5.2. Visual-Inertial SLAM

Three SOTA visual-inertial SLAM methods are tested against MCD: VINS [33], KIMERA [35], and OPENVINS [15]. VINS is a full-fledged visual SLAM method with sliding window optimization and loop-closure + bundle adjustment. KIMERA has a VIO backbone with extra modules such as incremental mesh generation and semantic representation. OPENVINS is a filter-based method with a KLT-based feature detector and tracker, online calibration, and multi-camera fusion. All three methods are configured to run the same stereo-camera-IMU suite.

The ATE of VI SLAM methods are reported in Tab. 2. At first glance, it can be seen that all methods struggle with the nighttime sequences. For VINS, all non-divergent experiments have an ATE beyond 4m, whereas KIMERA achieves better accuracy in several sequences. However, due to its additional geometric verification steps, KIMERA is more prone to losing track than VINS. OpenVINS seems to strike a better balance between robustness and accuracy, as it can complete all daytime sequences and is only compromised in nighttime HHS sequences. This compromise is primarily due to the discrete-time histogram equalization step that causes the feature to lose association in extreme lighting conditions. This result emphasizes the domain shifts and challenges in night video data, highlighting the difficulties for algorithms to adapt across diverse locations.

5.3. Range-Aided Localization

Several SOTA range-aided localization methods with different sensor configurations are tested against MCD: RO (range-only odometry) [42], and VRO (visual-ranging odometry) [26]. Their results are shown in the last two columns in Tab. 2.

First, we notice that range-based methods are less likely to diverge than VIO methods because the range observations restrict the estimates to diverge too far from the anchors. However, the ATEs in many sequences are quite large compared to VIO, reflecting that existing range-based methods have only been tested in small environments with consistent LOS (line-of-sight) to all anchors. In MCD, UWB anchors are deployed sparsely over the kilometer-long routes. Thus, the robot can lose LOS to some anchors frequently. In contrast, thanks to the visual factors filling in when UWB LOS is lost, the ATE is much better with the VRO method. Nevertheless, the accuracy can still be improved with more robust algorithms.

5.4. Semantic Segmentation

3D semantic segmentation is a long-standing topic that has attracted increasing attention in recent years. While existing methods are mostly tested on point clouds captured by CCS lidar, we investigate whether their state-of-the-art performance can be maintained on point clouds from NRE lidar.

Baselines. Eight SOTA methods for 3D semantic segmentation are tested on MCD, including SalsaNext [10], MKNet (MinkowskiNet) [8], SPVCNN [40], SDSeg3D [20], Cylinder3D [48], WaffleIron [32], and S-Former (SphereFormer) [19]. Specifically, SalsaNext is a projection-based method utilizing a CNN-based backbone with faster inference speed, which introduces an additional residual dilated convolution module compared to the previous range-image-based method [25]. On the other hand, MKNet, SPVCNN, and SDSeg3D are regarded as voxel-based methods that leverage sparse convolution [8] as their main feature extractor given 3D voxels as input. SPVCNN [40] further presents an efficient point-voxel feature fusion to embed dense point-wise features while SDSeg3D introduces self-distillation and test-time augmentation for performance boost. The rest of the baseline methods consider various 3D representations to cope with the 3D pattern of the raw input, such as cylindrical partitions (Cylinder3D [48]), spherical partitions (SphereFormer [19]), and high-level point-wise representations (WaffleIron [32]).

Results and Analysis. All baseline methods are trained on the NTU subset of MCD with fully supervised learning. As presented in the upper table of Tab. 3, while training and testing within the same campus, existing 3D segmentation methods can not adapt well to the different point cloud patterns of NRE lidar, where all baseline methods suffer from serious mIoU drops of more than 25% relatively compared to training on SemanticKITTI. For methods based on cylindrical-specific 3D representation, such as

Table 3. Segmentation performance of existing state-of-the-art methods on the MCD NTU (upper) and MCD TUHH (bottom) sequences. The intersection over union (IoU) is utilized as our evaluation matrix. The figure in round brackets is the relative mIoU gap on MCD NTU compared to SemanticKITTI [2] (upper) and the one on MCD TUHH compared to MCD NTU (bottom), respectively. Note that "Traffic-C" and "Structure-O" are short for "Traffic Cone" and "Structure-Other" classes, respectively.

Methods	*Barrier*	*Bike*	*Building*	*Chair*	*Cliff*	*Container*	*Curb*	*Fence*	*Hydrant*	*Sign*	*Lanemark*	*Other*	*Parkinglot*	*Pedestrian*	*Pole*	*Road*	*Shelter*	*Sidewalk*	*Stairs*	*Structure-O*	*Traffic-C*	*Trunk*	*Vegetation*	*Vehicle*	mIoU
MKNet [8]	57.5	59.7	79.9	34.5	1.8	11.5	45.7	31.6	17.3	62.9	46.9	1.6	9.2	48.9	53.3	89.7	52.4	56.6	42.2	6.7	8.6	45.1	87.7	67.5	46.2 (-26.9%)
SalsaNext [10]	49.6	0.2	81.3	37.2	6.4	0.0	48.7	36.7	9.3	35.5	45.7	30.2	12.2	11.8	27.5	91.4	43.7	66.3	0.0	0.0	8.1	31.1	84.0	48.2	33.5 (-43.6%)
SPVCNN [40]	65.5	56.1	81.4	44.9	8.6	16.6	50.6	37.0	22.0	65.8	49.7	1.3	28.2	46.3	58.1	91.7	62.9	93.8	49.1	6.1	8.9	51.3	89.5	71.4	48.2 (-27.5%)
Cylinder3D [48]	70.5	1.4	90.1	42.3	9.2	19.5	43.4	31.6	16.2	65.4	48.5	1.4	13.6	41.1	65.1	90.5	60.7	58.7	41.3	0.6	10.3	57.2	91.6	72.1	43.5 (-36.9%)
SDSeg3D [20]	72.2	**57.9**	82.2	**57.3**	0.8	0.0	47.6	31.4	24.4	67.1	49.1	0.0	12.5	45.3	61.0	91.6	69.0	64.8	49.0	0.0	11.1	56.0	89.4	80.0	46.7 (-31.4%)
WaffleIron [32]	76.7	28.3	81.6	43.9	6.9	6.9	57.2	37.7	14.7	57.9	53.5	**11.7**	39.7	52.5	56.7	91.5	67.3	77.0	65.8	**42.2**	12.4	59.0	88.2	75.9	50.2 (-29.1%)
S-Former [19]	**79.1**	46.1	**88.7**	50.4	**16.6**	**23.1**	**61.2**	**46.7**	**51.2**	**75.4**	**55.1**	10.4	**45.1**	**66.6**	**73.9**	**93.6**	**78.1**	**77.8**	**72.6**	37.6	**26.4**	**63.3**	**92.3**	**83.7**	**59.0** (-21.2%)

Methods	*Barrier*	*Bike*	*Building*	*Chair*	*Cliff*	*Container*	*Curb*	*Fence*	*Hydrant*	*Sign*	*Lanemark*	*Other*	*Parkinglot*	*Pedestrian*	*Pole*	*Road*	*Shelter*	*Sidewalk*	*Stairs*	*Structure-O*	*Traffic-C*	*Trunk*	*Vegetation*	*Vehicle*	mIoU
MKNet [8]	0.1	1.6	69.1	0.0	0.0	**0.1**	1.3	4.6	0.0	5.1	2.0	6.5	4.0	36.7	31.5	3.3	0.0	15.4	0.3	0.7	0.0	29.6	61.9	37.8	13.0 (-71.9%)
SalsaNext [10]	0.0	0.0	35.9	0.1	0.0	0.0	0.6	6.2	0.0	3.2	3.1	0.2	0.4	0.8	9.8	4.1	0.0	8.4	0.0	0.0	0.0	9.2	39.9	18.1	5.8 (-82.1%)
SPVCNN [40]	0.1	0.4	68.7	0.1	0.1	0.0	1.7	5.6	0.0	**6.4**	2.4	6.8	5.3	34.0	27.2	3.9	0.0	17.1	0.5	0.9	0.0	28.9	**64.0**	35.1	12.9 (-73.3%)
Cylinder3D [48]	0.1	0.6	66.8	0.1	0.0	0.0	0.4	**8.9**	0.0	5.3	**6.3**	3.7	2.1	14.1	12.0	**6.8**	0.0	27.1	0.4	0.0	0.0	9.2	50.5	14.6	9.6 (-76.9%)
SDSeg3D [20]	0.1	3.3	**73.2**	0.1	0.0	0.0	1.1	8.1	0.0	5.3	2.6	0.1	3.7	**45.9**	**38.0**	4.2	**0.3**	**19.9**	1.1	0.0	0.0	**43.4**	61.2	**48.3**	**15.0** (-67.9%)
WaffleIron [32]	0.1	5.5	65.0	0.2	0.2	0.0	1.3	7.9	0.0	6.0	2.5	0.0	**5.5**	35.2	18.6	3.6	0.0	8.3	0.1	1.0	0.0	27.6	52.8	32.7	11.4 (-77.3%)
S-Former [19]	**0.2**	1.4	72.8	**0.5**	**0.1**	0.0	**1.8**	6.3	0.0	5.4	3.7	2.8	3.3	40.1	33.9	4.5	0.0	17.8	**5.0**	**0.9**	0.0	35.8	54.8	40.4	13.8 (-76.6%)

Cylinder3D with cylindrical partitions and projection-based SalsaNext, the gaps are more severe since their designed representations are ineffective due to the different point cloud patterns. Voxel-based methods, including MKNet, SPVCNN, and SDSeg3D, also suffer from a significant performance drop mainly caused by point density discrepancy. Among all baseline methods, WaffleIron and S-Former are the only two approaches that achieve more than 50% on mIoU, mainly because their pre-defined representation can be propagated to point clouds captured by NRE lidar. Nevertheless, they similarly suffer from poor segmentation performance on some rare but important objects, such as traffic cones ("Traffic-C") and hydrants.

When testing on a different campus MCD TUHH as in the bottom table of Tab. 3, the performance of all methods significantly decreases by relatively more than 70% compared to the scores they achieve on MCD NTU, which indicates that the existing methods seriously suffer from the cross-campus discrepancy. In fact, this cross-campus inference task can be viewed as a *domain adaptation* problem [5, 30], where the domain gap is mainly attributed to the infrastructure difference between different campuses and different weather conditions. It is worth further investigation to improve the generalizability of existing segmentation networks. Detailed experimental settings and further analysis are provided in our supplementary materials.

5.5. Discussion and Insights

Based on our experiment benchmarks, no single SOTA SLAM method consistently excels across all MCD sequences. Additionally, there's a noticeable gap in SOTA semantic segmentation research for the NRE lidar modality. This shows the complexity of MCD through several diverse scenarios and conditions. It is our hope that the challenges in MCD can inspire new accurate and robust methods covering a variety of environments. Our configurations for these experiments will be shared on the MCD website.

6. Conclusion

This work introduces MCD data suite, featuring a wide range of the latest sensors and extensive coverage of diverse campuses across the Eurasian continent, along with semantic annotation on NRE lidar scans, and novel high-accuracy continuous-time ground truth. Tested with various state-of-the-art SLAM and perception methods, the dataset uncovers numerous challenges, calling for robust and precise solutions from the research community.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Medium Sized *Center for Advanced Robotics Technology Innovation (CARTIN)*, and the *Wallenberg AI, Autonomous Systems and Software Program* via the *2020 Wallenberg-NTU Presidential Postdoctoral Fellowship*.

References

[1] Hassan Alhaija, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving

scenes. *International Journal of Computer Vision (IJCV)*, 2018. 3

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 2, 3, 8

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35 (10):1157–1163, 2016. 2, 3

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3

[5] Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, and Lihua Xie. Multi-modal continual test-time adaptation for 3d semantic segmentation. *arXiv preprint arXiv:2303.10457*, 2023. 8

[6] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, 2016. 2, 3

[7] Kenny Chen, Ryan Nemiroff, and Brett T Lopez. Direct lidar-inertial odometry: Lightweight lio with continuous-time motion correction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3983–3989. IEEE, 2023. 5

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 7, 8

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[10] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15*, pages 207–222. Springer, 2020. 7, 8

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 4

[12] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. 3

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2, 3

[15] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. Openvins: A research platform for visual-inertial estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4666–4672. IEEE, 2020. 7

[16] Michael Helmberger, Kristian Morin, Nitish Kumar, Danwei Wang, Yufeng Yue, Giovanni Cioffi, and Davide Scaramuzza. The hilti slam challenge dataset, 2021. 3

[17] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 3

[18] Jinyong Jeong, Younggun Cho, Young-Sik Shin, Hyunchul Roh, and Ayoung Kim. Complex urban dataset with multi-level sensors from highly diverse urban environments. *The International Journal of Robotics Research*, 38(6):642–657, 2019. 3

[19] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17545–17555, 2023. 7, 8

[20] Jiale Li, Hang Dai, and Yong Ding. Self-distillation for robust lidar semantic segmentation in autonomous driving. In *European Conference on Computer Vision*, pages 659–676. Springer, 2022. 7, 8

[21] Hyungtae Lim, Daebeom Kim, Beomsoo Kim, and Hyun Myung. Adalio: Robust adaptive lidar-inertial odometry in degenerate indoor environments. *arXiv preprint arXiv:2304.12577*, 2023. 3

[22] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision (IJCV)*, 2020. 3

[23] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 3

[24] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[25] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220. IEEE, 2019. 7

[26] Thien Hoang Nguyen, Thien-Minh Nguyen, and Lihua Xie. Tightly-coupled ultra-wideband-aided monocular visual slam with degenerate anchor configurations. *Autonomous Robots*, 44(8):1519–1534, 2020. 7

[27] Thien-Minh Nguyen, Shenghai Yuan, Muqing Cao, Yang Lyu, Thien H Nguyen, and Lihua Xie. Ntu viral: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint. *The International Journal of Robotics Research*, 41 (3):270–280, 2022. 3

[28] Thien-Minh Nguyen, Daniel Duberg, Patric Jensfelt, Shenghai Yuan, and Lihua Xie. Slict: Multi-input multi-scale surfel-based lidar-inertial continuous-time odometry and mapping. *IEEE Robotics and Automation Letters*, 8(4): 2102–2109, 2023. 5

[29] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557. IEEE, 2019. 3

[30] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021. 8

[31] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE, 2020. 3

[32] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Using a waffle iron for automotive point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3379–3389, 2023. 7, 8

[33] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 7

[34] Milad Ramezani, Yiduo Wang, Marco Camurri, David Wisth, Matias Mattamala, and Maurice Fallon. The newer college dataset: Handheld lidar, inertial and vision with ground truth. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4353–4360, 2020. 2, 3, 5

[35] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696. IEEE, 2020. 7

[36] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Rus Daniela. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5135–5142. IEEE, 2020. 5

[37] Ha Sier, Qingqing Li, Xianjia Yu, Jorge Peña Queralta, Zhuo Zou, and Tomi Westerlund. A benchmark for multi-modal lidar slam with ground truth in gnss-denied environments. *Remote Sensing*, 15(13):3314, 2023. 5

[38] Christiane Sommer, Vladyslav Usenko, David Schubert, Nikolaus Demmel, and Daniel Cremers. Efficient derivative computation for cumulative b-splines on lie groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11148–11156, 2020. 5

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 2, 3

[40] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702. Springer, 2020. 7, 8

[41] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 3

[42] Chen Wang, Handuo Zhang, Thien-Minh Nguyen, and Lihua Xie. Ultra-wideband aided fast localization and mapping system. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1602–1609. IEEE, 2017. 7

[43] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016. 3

[44] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 4

[45] Weisong Wen, Yiyang Zhou, Guohao Zhang, Saman Fahandezh-Saadi, Xiwei Bai, Wei Zhan, Masayoshi Tomizuka, and Li-Ta Hsu. Urbanloco: A full sensor suite dataset for mapping and localization in urban scenes. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2310–2316. IEEE, 2020. 3

[46] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053 – 2073, 2022. 5

[47] Lintong Zhang, Michael Helmberger, Lanke Frank Tarimo Fu, David Wisth, Marco Camurri, Davide Scaramuzza, and Maurice Fallon. Hilti-oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping. *IEEE Robotics and Automation Letters*, -1(-1):–1–1, 2022. 3, 5

[48] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021. 7, 8