2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

VBR: A Vision Benchmark in Rome

Leonardo Brizi* Emanuele Giacomini* Luca Di Giammarino Simone Ferrari

Omar Salem Lorenzo De Rebotti Giorgio Grisetti

***Abstract*— This paper presents a vision and perception research dataset collected in Rome, featuring RGB data, 3D point clouds, IMU, and GPS data. We introduce a new benchmark targeting visual odometry and SLAM, to advance the research in autonomous robotics and computer vision. This work complements existing datasets by simultaneously addressing several issues, such as environment diversity, motion patterns, and sensor frequency. It uses up-to-date devices and presents effective procedures to accurately calibrate the intrinsic and extrinsic of the sensors while addressing temporal synchronization. During recording, we cover multi-floor buildings, gardens, urban and highway scenarios. Combining handheld and car-based data collections, our setup can simulate any robot (quadrupeds, quadrotors, autonomous vehicles). The dataset includes an accurate 6-dof ground truth based on a novel methodology that refines the RTK-GPS estimate with LiDAR point clouds through Bundle Adjustment (BA). All sequences divided in training and testing are accessible at www.rvp-group.net/datasets/slam.**

I. Introduction

Computer vision communities have relied on standard datasets to enhance their techniques since the early days. When ground truth data was accessible for a specific task, these communities devised appropriate metrics to evaluate the accuracy of their algorithm's results. With the rapid advancement of machine learning, datasets equipped with ground truth have become essential inputs for algorithms designed to learn intricate, non-parametric models. After KITTI, several other multi-sensor datasets [13], [15], [11], [19] have been presented, but no one seemed to have the same impact on the robotics and computer vision community as the original work [7].

Whereas the merits of KITTI are undisputed, and the core ideas are still valid, the dataset shows its years. The available sensors in the last decade improved significantly, and the same holds for computing devices and ground truth systems. Perhaps the main shortcoming of many datasets [7], [11], [18], [2] is the limited positional ground truth that is purely based on RTK-GPS and IMU and suffers from synchronization issues. In addition to that, the work is targeted at autonomous driving, hence, the data cover only road-like scenarios.

Other works aimed at addressing other aspects, such as seasonal changes [13], offering hand-held motion with a more accurate ground truth [14]. Still, to our knowledge, none of the recent datasets seem to address multiple issues. In this work, we present a contribution that aims to approach all these aspects simultaneously. At the moment of writing, we propose 6 datasets acquired with a hardware-synchronized sensor setting consisting of a 3D LiDAR, a stereo camera with a large baseline, an RTK-GPS, and an inertial sensor. Our data covers some of the most characteristic areas of Rome, spanning over 40 km of trajectory in almost 4 hours of recording. The raw data have a footprint of about 2TB. The sequences have been recorded in different environments, covering urban, forest, and indoor scenarios, using the same kind of sensors but at different frequencies and modalities. Heterogeneous sequences have been intentionally recorded to create a more challenging dataset, preventing domain overfitting. Moreover, we illustrate a procedure for obtaining highly accurate ground truth in large environments combining an RTK-GPS with a Bundle Adjustment schema on the LiDAR data to obtain precise trajectories. The accuracy of our ground truth, validated with a Total Station is ± 3 cm along an indoor/outdoor trajectory of 1.5 km. For each dataset we provide two flavors, similar to KITTI: a training version with ground truth available and a benchmarking version where the ground truth is not publicly provided. The

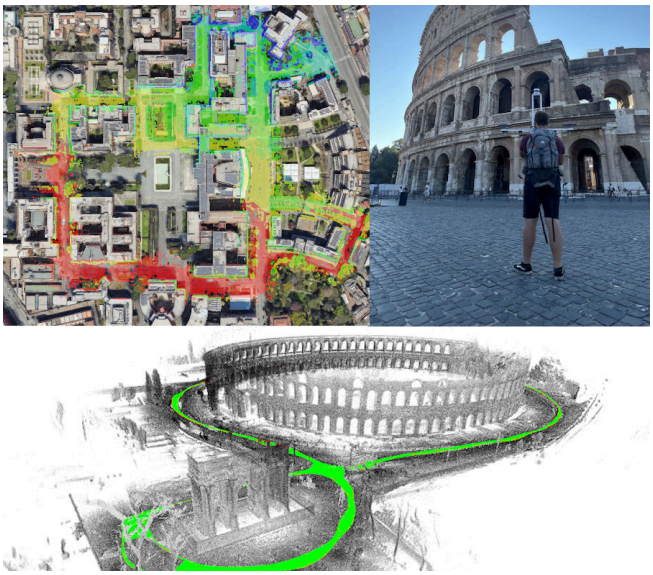


Fig. 1: A summary of our dataset. Data illustrating some of the sequences recorded (top). 3D mapping done with of our ground truth (bottom).

All authors are with the Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Italy, Email: {brizi, giacomini, digiammarino, s.ferrari, salem, derebotti, grisetti}@diag.uniroma1.it.

This work has been supported by PNRR MUR project PE0000013-FAIR.

*The authors contributed equally.

results of the community on the training datasets will be evaluated off-core. The public benchmark with the leading table will be available at our website.

II. Related Work

Within the domain of SLAM and 3D reconstruction, several datasets have played pivotal roles in benchmarking and advancing autonomous robotics systems and algorithms. These datasets vary in terms of their operational contexts, sensory configurations, and the accuracy of their associated ground truth data. One of the seminal car datasets in this field is KITTI. Its strength is providing different benchmarks (i.e. visual odometry, optical flow, stereo matching, and object detection). The KITTI datasets are publicly available and divided into training and evaluation sequences, fostering fair comparisons among the approaches. Despite being made available for more than a decade nowadays, KITTI is the most used benchmark in robotics perception and computer vision. However, KITTI presents synchronization issues between IMU readings and images, the ground truth data for visual odometry is produced only by fusing RTK GPS receiver and IMU, and the hardware used for recording data is nowadays outdated (Fig. 2, Fig. 3). Still, KITTI was pivotal in the development of many popular SLAM methods. Oxford RobotCar is another noteworthy car dataset [13]. In contrast to KITTI, it distinguishes itself by featuring the longest sequences among the datasets. Yet, the ground truth in the Oxford RobotCar dataset relies only on partial GPS and INS data, which makes the baselines unreliable for benchmarking the accuracy of SLAM and localization methods. Furthermore, approaches like Mulran use the same ground truth generation process, leading to the same issue. However, this dataset is renowned for embracing multimodal sensor data, including LiDAR and radar. While this adds diversity to the sequences, only the front half of the LiDAR field-of-view is included in the data collection [11].

In contrast to datasets collected from ground-based vehicles, certain research efforts have focused on data acquisition from micro aerial vehicles. For instance, the EuRoC dataset [3] stands out for its use of synchronized hardware and a laser tracking system to attain accurate ground truth data. However, the dataset does not provide LiDAR aquisition but only a stereo-camera and an IMU. In addition, data is recorded only in industrial environments.

Recent advancements in handheld datasets, exemplified by Newer College [14] and Hilti [19], have achieved exceptional levels of accuracy in ground truth generation through the use of 3D imaging laser scanners. This innovative technique involves acquiring a prior map and registering LiDAR point clouds using a localization approach. Nevertheless, a notable limitation in this case is the impracticality of applying this method to large-scale scenarios, which constrains its broader utility.

This paper introduces a diverse and heterogeneous dataset encompassing a wide range of environments. Our design accommodates various robotic platforms, including quadrupeds, quadrotors, and autonomous vehicles, making it a versatile resource for the robotics community.

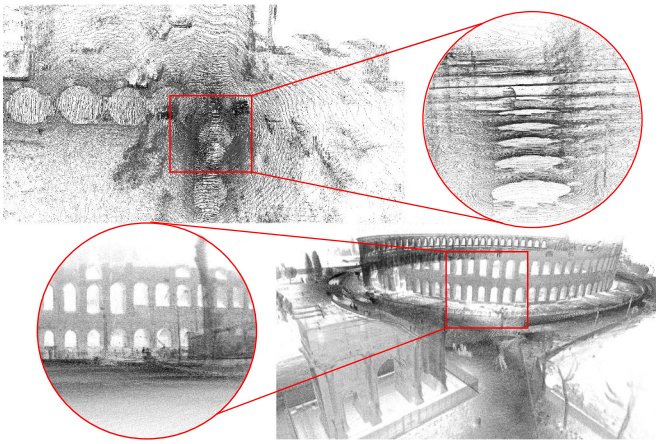


Fig. 2: Comparison between LiDAR clouds attached to ground truth trajectories of KITTI (up) and ours (down). The zoom shows the elevation view.

Fig. 3: Projection of the KITTI LiDAR point cloud into an image plane (up), projection of our LiDAR into an image plane (down). The many holes of the up image due to uneven distribution of the LiDAR beams and calibration issues make the KITTI LiDAR image unusable for computer vision tasks.

We maintain hardware synchronization to ensure data accuracy and reliability and employ stereo cameras with a wide baseline to capture robust visual information. Furthermore, we provide an accurate 6-dof ground truth even in large scale scenarios.

III. The Datasets

Creating comprehensive and authentic benchmarks for the tasks mentioned earlier is challenging. These challenges encompass collecting vast data in real-time, calibrating differ-

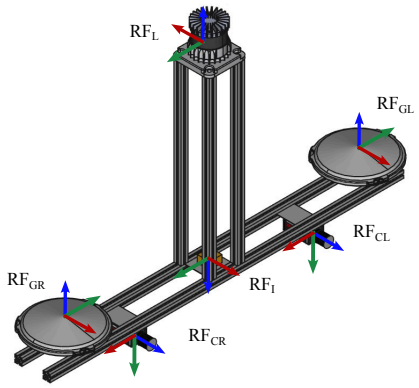


Fig. 4: Sensor setup and reference frames. Our ground truth is expressed in the LiDAR reference frame $\mathrm{RF_L}$. More details can be found in our website and supplementary materials.

	Accurate GT	Indoor	Outdoor	Various Motion	Large Scale	Benchmark
KITTI [7]			X		X	X
EuRoC [3]	X	X				
Oxford RobotCar [13]			X			
ETH3D [15]	X	X				X
MulRan [11]			X		X	
Newer College [14]	X		X			
Hilti [19]	X	X	X	X		X
Kitti-360 [12]	X		X		X	X
Ours	X	X	X	X	X	X

TABLE I: The table summarizes the most important datasets in robotics perception and computer vision related to odometry estimation and SLAM. For "Accurate GT" we mean any ground truth recorded with motion capture, Laser Total Station, or globally refined.

ent sensors operating at various speeds, producing accurate ground truths with minimal oversight, and choosing the right sequences and frames for every benchmark. The following section delves into our approaches to address these issues.

A. Sensors setup

Our sensor system is illustrated in Fig. 4 and consists of two RGB cameras, a 3D LiDAR, an RTK-GPS, and an IMU. The cameras are two global shutter Manta G-145 capturing in RGB and arranged in a wide stereo fashion, with a baseline of approximately 50 cm. The cameras have a horizontal FoV of $45°$ and a vertical one of $40°$. During the acquisition, we enable auto white-balance and auto-exposure, while maintaining a fixed focus. The maximum exposure is fixed at 20 ms. Each image is 1388×700 pixels and stored in Bayer pattern to reduce the memory footprint without losing information.

Two LiDARs were employed, tailored to the specific motion characteristics of the captured sequences. For hand-held sequences, an Ouster OS0-128 was used. This sensor offers a maximum range of 55 meters and a vertical FoV of $90°$ spanned by 128 beams. For car sequences, we used an Ouster OS1-64. This sensor provides a longer maximum range of 120 meters, a narrower vertical FoV of $45°$ spanned by 64 lasers.

The Inertial Measurement Unit (IMU) is an SBG Ellipse-E IMU, with 0.05 ° of roll/pitch accuracy, whose firmware supports GNSS integration. The RTK-GPS antennas are two Septentrio PolaNt-x MF GNSS antennas mounted in differential configuration. The GPS receiver supports multi-frequency GPS, GLONASS, Galileo, BeiDou, QZSS, NavIC, Compass and L-band signal reception with an accuracy open-sky condition of 0.6 cm horizontally and 1 cm vertically. All sensors are rigidly attached to an aluminum frame. The relative position of the sensor is the same for both the hand-held datasets and for the driving datasets. Fig. 1 shows the sensor placement on the car. Tab. II summarizes the devices used in our system.

B. Calibration

The accuracy of intrinsic and extrinsic sensor calibration is fundamental in achieving dependable ground truth data. Our calibration process is outlined below.

Sensor	Type	Details	Rate
LiDARs	Ouster, OS0-128	Vertical FOV: $90°$ Horizontal Res: 2048	10 Hz
	Ouster, OS1-64	Vertical FOV: $45°$ Horizontal Res: 1024	20 Hz
Cameras	Manta G-125B/C	Global Shutter Stereo configuration Wide baseline	20 Hz 30 Hz
IMU	SBG Ellipse-E	GNSS synchronization $0.05°$ precision	100 Hz

TABLE II: Summary of the devices in our sensor setup, and the accuracy of the temporal synchronization.

Initially, we calibrate the stereo camera intrinsically and extrinsically. Subsequently, we determine the $\mathbb{SE}(3)$ parameters that connect the coordinate systems of the laser scanner within the right camera. Finally, we align the LiDAR /camera system with GPS/IMU reference frame. To calibrate the camera's intrinsic and extrinsic parameters, we use an A3 checkerboard. Keeping the camera steady, we move the checkerboard and detect its corners in the calibration images. Minimizing the average reprojection error allows us to find optimal parameters for our setup [1]. Using the same target, we estimate the rigid transformation between the right camera ($\mathrm{RF}_{\mathrm{CR}}$) and the LiDAR. We achieve accurate results by minimizing plane-to-plane error [8]. For each recorded sequence we acquire the calibration data which are public available.

Determining the relative pose between the GPS/IMU and the LiDAR relies on the sensor systems' motion, since the two devices cannot observe a common target. To this extent, we recover a trajectory from the LiDAR/camera system using a LiDAR odometry based on point-to-plane ICP. During the process, we ensure a wide range of orientations and translations essential for addressing the minimization issue. This technique is known as *hand-eye* calibration [6] and aims at computing the sensor offset that results in the maximum overlap between two LiDAR trajectories: the one computed by the odometry, and the one obtained by computing the LiDAR motion from the GPS measurements (after applying the estimated offset).

C. Synchronization

A key challenge during acquisition involves synchronizing sensors to establish a shared temporal reference. Within our

platform, two separate subsystems are at play: one comprises the LiDAR and RGB stereo pair, while the other includes the GNSS receiver and IMU. In the first system, the LiDAR takes on the role of the master, generating synchronization pulses during its acquisition phase based on angle data from its encoder. For hand-held use, the LiDAR records at 10 Hz, and the synchronization pulse activates every 120 degrees, resulting in a 30 Hz signal. Moreover, in the automotive setup, the LiDAR operates at 20 Hz, with the synchronization pulse set to trigger once per revolution. The signal triggers the frame acquisition for both cameras, leading to sub-millisecond synchronization between the frames. Once the frames are received, their timestamp is overwritten with one of the LiDAR data packets received when the encoder was at the trigger angle. This ensures an accurate hardware synchronization between the two sensors.

In contrast, the GNSS-IMU system relies on Pulse Per Second (PPS) protocol for synchronization, which is directly addressed by the IMU firmware. The streams from the two subsystems are merged together by performing an offline time synchronization to determine the difference between the internal clocks of GPS and LiDAR. We exploit the internal LiDARIMU measurements to compute the temporal shift respect to the IMU, using cross-correlation. This technique allows us to obtain a maximum of 5 ms error considering possible un-observable phase shift between the IMUs signals that come every 10 ms. A temporal drift also affects the internal clocks of the LiDAR and GPS. In the longest sequences, we observed a maximum shift between the two clocks of about 10 ms, which is neglectable compared to the scan/image frequency and the velocity of the sensor.

D. Ground truth generation

Generating accurate trajectories is the main objective of this work; hence, major attention was dedicated to this task. Some work produces very accurate ground truth using ICP localization within 3D Total Station reconstruction ([14], [19]). This is not always possible when moving in a very large environment. In such cases, a GNSS RTK system is usually used. These systems can reach an accuracy of a few centimeters, but the signal quality is not always optimal. In addition, these systems provide good global estimation, but poor locality.

In the remainder, we assume a 3D pose $\mathbf{X} \in \mathbb{SE}(3)$ be represented as a homogeneous 4×4 matrix where $\mathbf{R}$ is the rotation matrix, and $\mathbf{t}$ is the translation vector. With the operator $\log(\mathbf{X})$, we refer to the conversion of a transform in minimal form (e.g. translation vector and unit quaternion for the orientations).

We combine the GPS priors $\mathbf{Z}_t^{\mathrm{g}} \in \mathbb{SE}(3)$ with a variation of BA formulation proposed in [5]. This allows us to combine the global accuracy of the GPS with the local precision of registration approaches. Our procedure works first by computing a LiDAR odometry that expresses the relative transform $\mathbf{Z}_{t,t+1}^{\mathrm{sm}}$ between subsequent frames. To this extent, we use the same ICP algorithm used for temporal synchronization mentioned in Sec. III-C. This odometry is accurate, reliable in the short term, and can cope with GPS outages. Subsequently, we determine a global alignment of all the poses using the RTK-GPS readings and considering the incremental measurements of the LiDAR odometry. In short, we solve the following optimization problem:

$$\mathbf{X}_{1:T}^{*} = \underset{\mathbf{X}_{1:T}}{\operatorname{argmin}} \sum \|\log(\mathbf{Z}_{t,t+1}^{\mathrm{sm}-1}\mathbf{X}_t^{-1}\mathbf{X}_{t+1})\|^2_{\mathbf{\Omega}_{t,t+1}^{\mathrm{sm}}} + \sum \|\log(\mathbf{Z}_t^{\mathrm{g}-1}\mathbf{X}_t)\|^2_{\mathbf{\Omega}_t^{\mathrm{g}}} \quad (1)$$

Here $\|\mathbf{v}\|^2_{\mathbf{\Omega}} = \mathbf{v}^T\mathbf{\Omega}\mathbf{v}$ denotes the Omega L2 norm of a vector $\mathbf{v}$, with $\mathbf{\Omega}$ representing the information matrix encoding the accuracy of the measurement. Accordingly, in Eq. (1) $\mathbf{\Omega}_t^{\mathrm{sm}}$ is the information matrix resulting from scan matching, while $\mathbf{\Omega}_t^{\mathrm{g}}$ encodes the GPS accuracy.

Once this process is completed and we have a reasonable initial guess, we look for the set of poses that is maximally consistent with all LiDAR scans. We solve the following problem using geometric and photometric error terms. The geometric part based on point-to-plane results in a configuration of the rigid motion which is close to the optimum. The second photometric step, increases the accuracy by ensuring subpixel consistency. Let $\langle i, j \rangle$ be the set of poses, $\langle k, l \rangle$ geometric associations, and u the image pixel generated by spherical projecting the LiDAR point cloud into an image (as illustrated in [5]). The total residual can be expressed as follows:

$$E^{\mathrm{ba}} = \sum_{i,j,k,l} \rho_{\mathrm{geo}} \|e_{k,l}^{\mathrm{geo}}\|^2_{\mathbf{\Omega}_{\mathrm{geo}}} + \sum_{i,j,u} \rho_{\mathrm{photo}} \|e_u^{\mathrm{photo}}\|^2_{\mathbf{\Omega}_{\mathrm{photo}}}. \quad (2)$$

We employ a point-to-plane metric for the geometric error term, explicitly relying on efficient KD-tree data association based on PCA splitting criteria. The geometric term can be compactly written as:

$$e_{k,l}^{\mathrm{geo}}(\mathbf{X}_i, \mathbf{X}_j) = (\mathbf{X}_i\mathbf{p}_{i,j} - \mathbf{X}_j\mathbf{p}_{k,l}) \cdot (\mathbf{R}_i\mathbf{n}_{i,k}) \quad (3)$$

with $\mathbf{p}_{i,k}$ and $\mathbf{p}_{j,l}$ denoting corresponding points between the poses $\mathbf{X}_i$ and $\mathbf{X}_j$, and $\mathbf{n}_{i,k}$ be the corresponding normal. The photometric term, instead, follows the formulation illustrated in [5], but relies only on range and intensity images. The overall error function to minimize, taking into account GPS information, will be therefore

$$\mathbf{X}_{1:T}^{gt} = E^{\mathrm{ba}} + \sum \|\log(\mathbf{Z}_t^{\mathrm{g}-1}\mathbf{X}_t)\|^2_{\mathbf{\Omega}_t^{\mathrm{g}}} \quad (4)$$

We measured the accuracy of our ground truth using a Total Station and 6 highly reflective markers disposed as a hexahedron in the scene, to lock all redundantly all degrees of freedom. Since ranges are invariant of reference frame, we measure the differences between the distances measured from the points acquired from Total Station and the one detected in our estimated map. Our 6-dof ground truth results in ± 3 cm accuracy on a trajectory of length of approximately 1.5 Km (indoor/outdoor). Our ground truth generation process has been shown to scale well to large environments while relying only on the onboard sensors.

The final estimated global clouds are usually in the order of billions of points.

We release the ground truth for each training sequence, always expressed in the LiDAR reference frame $\mathrm{RF_L}$.

E. Data selection

In the context of this research study, the OS0-128 LiDAR system offers extensive capabilities for collecting spatial data. Specifically, it delivers precise range measurements across the entire horizontal plane, covering large distances. Furthermore, it employs an dense array of vertical beams distributed over a $90°$, enabling comprehensive scans of a spherical area surrounding the sensor.

Given the nature of the chosen environments and to ensure a rich and diverse dataset, we used various configurations provided by the LiDARs with varying resolution and frequency. We used the OS1-64 for car sequences at 20 Hz to maximize the observation in wide scenarios and to reduce the skewing effect at higher speeds. Moreover, we employed the OS0-128 for hand-held sequences at 10 Hz to maximize the observation in narrow scenarios.

We provide 6 datasets split into different sequences. Among the datasets, 4 were acquired by walking using the hand-held device, while the other 2 collected by car. Each sequence was collected in a different environment, with different challenging scenarios such as dynamics, traffic, long sequences, and wide areas (Fig. 5). Tab. III summarizes some parameters for the sequences and provides some illustration.

In the remainder, we shortly review each sequence, describing the scenario:

a) Spagna: this sequence has been acquired in *Piazza di Spagna* and in the nearby streets. It features several large loops going up and down the stairs hence, the trajectory is non-planar. The narrow streets limit the FoV of the LiDAR, but the building facades are a rich source of structure.

b) Colosseum: consists of two rounds around the *Colosseum* and the *Arco di Costantino*. The range of the LiDAR is not always sufficient to capture vertical structure. In some cases, the maximum range of the sensor is not sufficient to measure the entire surroundings, and the environment is repetitive, making some chunks difficult for state estimation.

c) Pincio: several loops were collected in *Villa Borghese*. This dataset is characterized by rich vegetation and a repetitive environment.

d) DIAG: this sequence is a mixed indoor/outdoor dataset. We traveled inside and outside our building, walking inside the corridors, through the courtyard, up to the stairs, and on the roof, which is the only part where the reception of RTK-GPS was available.

e) Campus: in contrast to all other sequences, this has been acquired at the main Campus of Sapienza University using the equipped car, and features several loops, spanning approximately all the streets that can be traversed by car. There are several narrow passage and some tunnels that pass under buildings. The dynamic is composed mainly of people walking composing a low percentage of the recorded data.

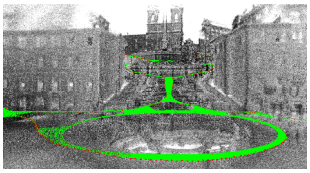
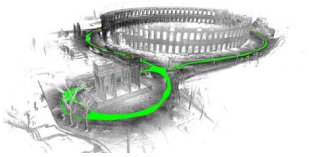
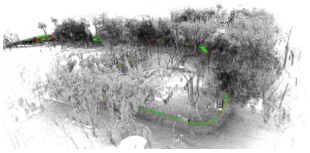
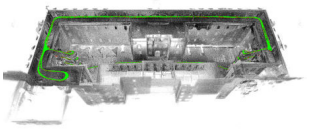
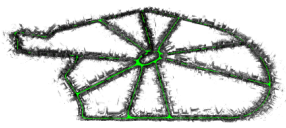
Dataset	Detail
Name: *Spagna* Motion: *hand-held* Type: *urban/vertical* Length: *2.045 km* Duration: *1827 s*	
Name: *Colosseum* Motion: *hand-held* Type: *urban/dynamics* Length: *2.159 km* Duration: *1383 s*	
Name: *Pincio* Motion: *hand-held* Type: *park/trees* Length: *2.541 km* Duration: *2064 s*	
Name: *DIAG* Motion: *hand-held* Type: *outdoor/indoor* Length: *1.480 km* Duration: *1458 s*	
Name: *Campus* Motion: *car* Type: *urban, underpasses* Length: *11.455 km* Duration: *2290 s*	
Name: *Ciampino* Motion: *car* Type: *urban/traffic* Length: *21.064 km* Duration: *3688 s*	

TABLE III: Summary of our sequences.

f) Ciampino: these sequences have been recorded in the city of Ciampino (Fig. 6). It is the longest sequence so far: the length of the total trajectory is about 21 km, while being subject to moderate dynamics.

IV. Benchmark

For a detailed assessment of SLAM and odometry estimation, we concentrate on the Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). As in [7], we assess rotation and translation errors independently rather than merging them into one metric.

ATE emphasizes SLAM performance over odometry. To compute its RMSE, we first align the estimated trajectory with the ground truth using a $\mathbb{SE}(3)$ transformation, matching poses with synchronized timestamps and employing the Horn method [9]. Subsequently, we calculate the RMSE of the ATE [m] among all the matched poses.

On the other hand, RPE emphasizes the odometry comparing local motion estimate chunks within the ground truth. It involves computing the RPE [%] (measured in percent-

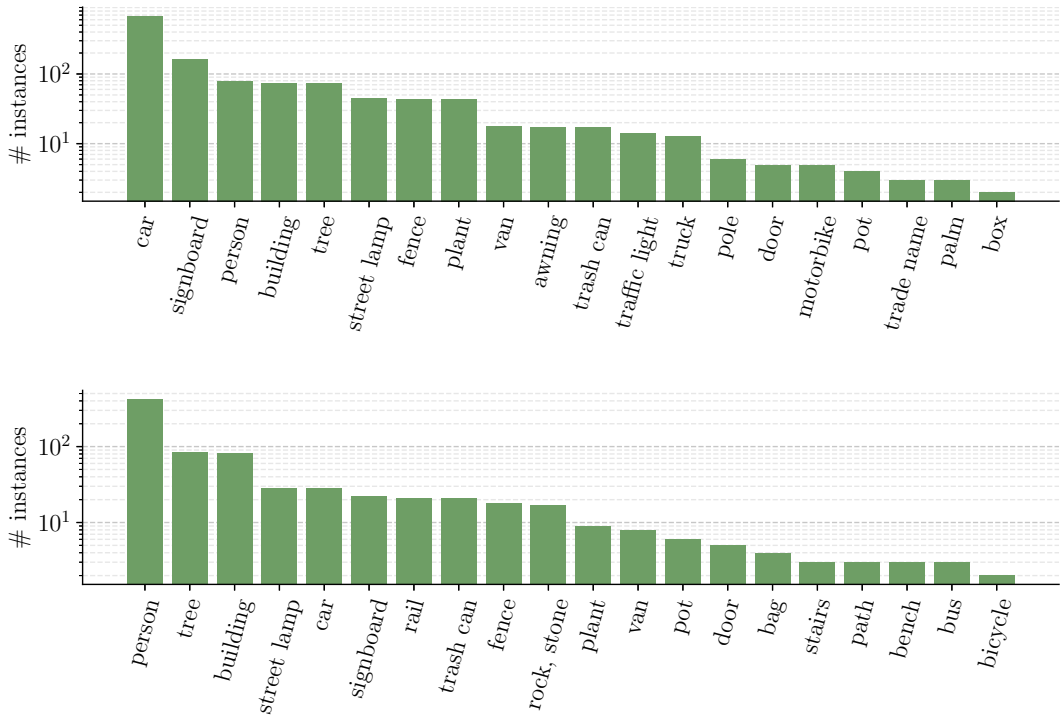


Fig. 5: Number of top 20 most frequent semantic instance for Ciampino (above) and Colosseum (below) sequences. The instances were counted using OneFormer [10] over a subset of images for each sequence and excluding the most predominant classes: sky, wall, road, grass, sidewalk, ground.

Fig. 6: The image shows the overlay of the 3D model obtained from our ground truth system and a view from Google Maps.

age), over a set of subsequences of different lengths, as proposed by [7]. Afterwards, the translational RPE [%] and the rotational RPE [deg/m] of a sequence are computed as the average of all chunks RPE. Differently than any other benchmark, we have chosen to make chunk lengths adaptive to the total sequence length. In fact, local and global accuracy of our ground truth trajectories allows us to choose subsequences of arbitrary lengths, without biasing evaluation results.

Given a trajectory estimate for each sequence, a cumulative error curve is computed, like those in Fig. 7. For a given error value on the x-axis, the y-axis shows in how many sequences a method achieves a lower error. Therefore, the method ranking is determined as the area under curve, up to the selected maximum error and, for this metric, the larger the better. This metric rewards the robustness of evaluated methods, since a successful result on a sequence usually adds much more area under the curve than slightly improving the accuracy on many sequences.

Additional information, supplementary materials, and the leading table can be accessed on our website.

A. Evaluation

To assess our recorded data's integrity, we evaluated different LiDAR odometry and visual SLAM systems on all our training sequences, specifically focusing on three notable solutions: KISS-ICP [16], F-LOAM [17] and ORB-SLAM3 [4]. The evaluation outcomes are reported in Fig. 7, showing cumulative RPE [%] and ATE RMSE [m], with threshold limits set to 10 % and 10 m respectively. As expected, LiDAR odometry methods are more robust and accurate than visual SLAM, in fact both KISS-ICP and F-LOAM perform successfully across all our 8 training sequences. The outcomes slightly change in the ATE RMSE [m] graph, where the area under the curve of every method is reduced, emphasizing the challenges in estimating the egomotion with global accuracy.

V. Conclusion

In this paper, we present a new vision and perception dataset, specifically targeted at SLAM and odometry estimation methods. Our sequences cover different environments

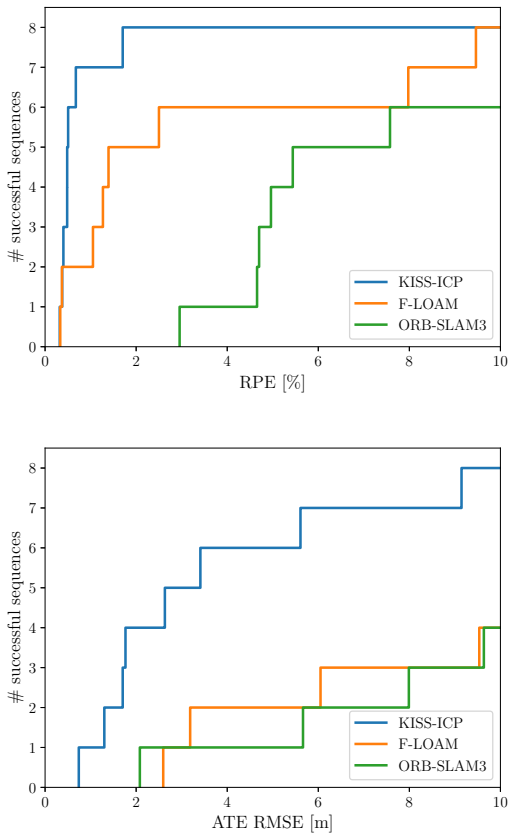


Fig. 7: Our benchmark. Cumulative RPE [%] and ATE RMSE [m] across all training sequences in our dataset for KISS-ICP, F-LOAM and ORB-SLAM3.

and are acquired in a hand-held fashion and by using a car. Our design is to accommodate various types of robotic platforms, including quadrupeds, quadrotors, and autonomous vehicles, making it a versatile resource for the robotics and vision community. Compared to existing datasets, we offer a variety of environments within our sequences. Moreover, this work presents a novel ground truth estimation, fusing an RTK-GPS with a LiDAR Bundle Adjustment schema. All the sequences are split into training and test sets. In addition we provide a public benchmark evaluation system, accessible from our website, that produces a leading table from the results submitted by the community.

As a further service to the community, we plan to extend our benchmark with other sequences, annotations and challenges in the area of computer vision and robotic perception (i.e. semantics, monocular and stereo dense depth estimation, object tracking, etc.).

Acknowledgement

We thank Juan D. Tardos for supporting us with ORB-SLAM. We are grateful to Roberto Mauroni and Francesco Spognardi for their invaluable support. Special thanks to Vincenzo Suriani for providing supplementary hardware. Last but not least, we appreciate Luca Callocchia's contribution in developing the website.

References

[1] G. Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[2] K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y. Leung, A. P. Schoellig, and T. D. Barfoot. Boreas: A multi-season autonomous driving dataset. *Intl. Journal of Robotics Research (IJRR)*, 42(1-2):33–42, 2023.

[3] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart. The euroc micro aerial vehicle datasets. *Intl. Journal of Robotics Research (IJRR)*, 35(10):1157–1163, 2016.

[4] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans. on Robotics (TRO)*, 37(6):1874–1890, 2021.

[5] L. Di Giammarino, E. Giacomini, L. Brizi, O. Salem, and G. Grisetti. Photometric lidar and rgb-d bundle adjustment. *IEEE Robotics and Automation Letters (RA-L)*, 2023.

[6] F. Dornaika and R. Horaud. Simultaneous robot-world and hand-eye calibration. *IEEE Trans. on Robotics and Automation*, 14(4):617–622, 1998.

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012.

[8] E. Giacomini, L. Brizi, L. Di Giammarino, O. Salem, P. Perugini, and G. Grisetti. Ca2lib: Simple and accurate lidar-rgb calibration using small common markers. *IEEE Sensors Journal*, 24(3), 2024.

[9] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *J. Opt. Soc. Am. A*, 5(7):1127–1135, Jul 1988.

[10] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim. Mulran: Multimodal range dataset for urban place recognition. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 6246–6253, 2020.

[12] Y. Liao, J. Xie, and A. Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[13] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The oxford robotcar dataset. *Intl. Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017.

[14] M. Ramezani, Y. Wang, M. Camurri, D. Wisth, M. Mattamala, and M. Fallon. The newer college dataset: Handheld lidar, inertial and vision with ground truth. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4353–4360, 2020.

[15] T. Schops, T. Sattler, and M. Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 134–144, 2019.

[16] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023.

[17] H. Wang, C. Wang, C.-L. Chen, and L. Xie. F-loam : Fast lidar odometry and mapping. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4390–4396, 2021.

[18] Z. Yan, L. Sun, T. Krajnik, and Y. Ruichek. EU long-term dataset with multiple sensors for autonomous driving. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[19] L. Zhang, M. Helmberger, L. F. T. Fu, D. Wisth, M. Camurri, D. Scaramuzza, and M. Fallon. Hilti-oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping. *IEEE Robotics and Automation Letters (RA-L)*, 8(1):408–415, 2022.