

# CroCoDL: Cross-device Collaborative Dataset for Localization

Hermann Blum<sup>1,3</sup>, Alessandro Mercurio<sup>1</sup>, Joshua O'Reilly<sup>1</sup>, Tim Engelbracht<sup>1</sup>,  
Mihai Dusmanu<sup>2</sup>, Marc Pollefeys<sup>1,2</sup>, Zuria Bauer<sup>1</sup>  
<sup>1</sup>ETH Zurich <sup>2</sup>Microsoft <sup>3</sup>Lamarr Institute / Uni Bonn

blumh@uni-bonn.de & mihaidusmanu@microsoft.com & {pomarc, zbauer}@ethz.ch

## Abstract

*Accurate localization plays a pivotal role in the autonomy of systems operating in unfamiliar environments, particularly when interaction with humans is expected. High-accuracy visual localization systems encompass various components, such as image retrievers, feature extractors, matchers, reconstruction and pose estimation methods. This complexity translates to the necessity of robust evaluation settings and pipelines. However, existing datasets and benchmarks primarily focus on single-agent scenarios, overlooking the critical issue of cross-device localization. Different agents with different sensors will show their own specific strengths and weaknesses, and the data they have available varies substantially. This work addresses this gap by enhancing an existing augmented reality visual localization benchmark with data from legged robots, and evaluating human-robot, cross-device mapping and localization. Our contributions extend beyond device diversity and include high environment variability, spanning ten distinct locations ranging from disaster sites to art exhibitions. Each scene in our dataset features recordings from robot agents, hand-held and head-mounted devices, and high-accuracy ground truth LiDAR scanners, resulting in a comprehensive multi-agent dataset and benchmark. This work represents a significant advancement in the field of visual localization benchmarking, with key insights into the performance of cross-device localization methods across diverse settings.*

## 1. Introduction

In recent years, a range of mixed-reality (MR) headsets appeared on the market, bringing them for the first time into private and public spaces with diverse applications, such as assisting in frontline activities, boosting the productivity of office workers, augmented / virtual tours, and games. Many of the foreseen use cases of these devices involve multiple users observing and interacting

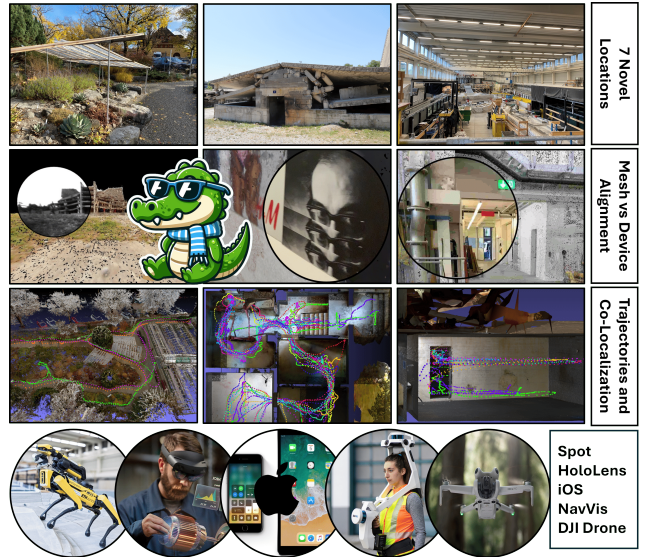
















































































Figure 1. **CroCoDL**: the first dataset to contain sensor recordings from real-world robots, phones, and MR headsets, covering a total of 10 challenging locations to benchmark cross-device and human-robot visual registration.

with the same virtual hologram through different MR interfaces. For the task of human-robot collaboration, MR unleashes the potential of visualizing the robot sensor readings, as well as planning its actions ‘at a glance’. The key technological challenge to interlace virtual or augmented reality experiences into shared digital spaces is to accurately localize all kinds of devices from different viewpoints with respect to each other. Localizing multiple devices with respect to a shared reference is a necessary condition for rendering shared content at the same location.

In this work, we investigate the problem of visual co-localization between pairs of devices. In particular, we are interested in device pairs that differ in their typical viewpoint, as well as their motion patterns and sensor configurations. Our main focus lies on mixed-reality headsets, hand-held smartphones, and legged robots. Our findings suggest that, while visual retrieval and registration have made considerable progress on

Table 1. **Commonly used datasets for visual localization and SLAM.** Legend:  inside,  outside environment;  Structural changes due to moving people,  long-term changes due to displaced furniture,  weather,  day-night,  construction work; Trajectory motion from sensors mounted on  ground vehicle,  legged robot,  drone,  car,  hand-held,  head-mounted, ‘syn.’ synthetic. (noted with \*: at most 2 devices are recorded in the same location; []: not aligned, due to safety / permission reasons - we could only capture drone footage in 8/10 locations)

Dataset	Motion	Env.	Locations	Changes	Sensors	GT pose accuracy	Seqs.
KITTI [7]			1		RGB, LiDAR, IMU	<10cm (RTKGPS)	22
TUM RGBD [17]			2		RGB-D, IMU	1mm (mocap)	80
EUROC [2]			2		RGB, IMU	1mm (mocap)	11
NCLT [3]		 	1	  	RGB, LiDAR, IMU, GNSS	<10cm (GPS + IMU + LiDAR)	27
UZH-FPV [5]		 	2	-	RGB, event camera, IMU	1cm (total station + VI-BA)	28
ETH3D SLAM [15]			1	-	RGB, depth, IMU	1mm (mocap)	96
OpenLoris-Scene [16]			5	 	RGB-D, IMU, wheel odom.	<10cm (2D LiDAR)	22
TartanAir [18]	syn.	 	30	-	RGB	perfect (synthetic)	30
UMA VI [21]	 	 	2	-	RGB, IMU	(visual tags)	32
Naver Labs [11]			5	 	RGB, LiDAR, IMU	<10cm (LiDAR SLAM and SfM)	10
HILTI SLAM [9]		 	8	-	RGB, LiDAR, IMU	<5mm (total station)	12
Graco [20]	 		1	 	RGB, LiDAR, GPS, IMU	~1cm (GNSS)	14
FusionPortable [10, 19]	2 ∈     *	 	9	-	RGB, event cameras, LiDAR, IMU, GPS	~1cm (GNSS RTK)	41
LaMAR [14]		 	3	   	RGB, LiDAR, depth, IMU, WiFi/BT	<10cm (LiDAR + PGO + PGO-BA)	500
<b>CroCoDL</b>	   [  ]	 	10	   	RGB, LiDAR, depth, IMU, WiFi/BT	~10cm (LiDAR + PGO + PGO-BA)	500 +800

datasets that are mostly recorded with a single type of device, visual localization of one against another can be very challenging even for state-of-the-art methods.

To investigate visual co-localization, we extend the landscape of visual localization data with a considerably larger and more diverse dataset and benchmark entitled “**CroCoDL**” illustrated in Figure 1. CroCoDL is the first dataset to contain sensor recordings from both robots and mixed-reality devices, and spans more real-world environments than any other existing cross-device visual localization dataset. In summary, our contributions are:

- The (to the best of our knowledge) largest real-world cross-device visual localization dataset, focusing on diverse capture setups and environments.
- A novel benchmark on cross-device visual registration that shows considerable limitations of current state-of-the-art methods.
- Integration of ROS-based robotic sensor streams into LaMAR’s pseudo-GT pipeline [14]. We will release the code for the data pre-processing and the required changes to the pipeline.

## 2. Related Work

We present an overview of the most relevant existing datasets used to evaluate visual localization and SLAM systems in Table 1. Most existing datasets focus on a single device type for data capture. The other datasets always combine at most two different devices per location: TUM RGBD [17] records handheld & ground robot, UMA VI [21] handheld and a few car

sequences, Graco [20] ground robot & drone. In Fusion-Portable [10, 19] 3 sequences overlap between handheld and a legged robot, one sequence legged robot with ground robot, and one sequence ground robot with car. Out of the four, LaMAR [14] is the only dataset that captures longterm, structural, day and night, and other changes. CroCoDL builds upon the efforts of LaMAR [14] by adding data recorded by a legged robot and a drone, and expanding the hand-held and head-mounted data from 3 to 10 locations. To the best of our knowledge, we introduce the largest cross-device real-world visual localization dataset, focusing on diverse devices and environments. Therefore, CroCoDL serves a different purpose than, *e.g.*, EUROC [2] or HILTI SLAM [9], where more accurate ground truth through motion-capture or line-of-sight tracking limits the variability of sensors, motion patterns, scale, and locations.

## 3. Dataset

The dataset consists of 10 distinct locations, featuring over **800** new sequences, totaling more than **100 hours** of original raw recording time with 5 different devices.

### 3.1. Locations

The first locations are recordings during an event for developing advanced robotics capabilities in hazardous environments (ARCHE [1]), held at a training village designed to safely simulate realistic disaster scenarios.

1. **ARCHE D2:** The intact basement of a semi-collapsed building.



Figure 2. **New locations of the CroCoDL dataset.** Each location has high-quality meshes, obtained from LiDAR, which are registered with numerous phone, AR headset, and robotic sequences. These locations were chosen to complement the existing locations in LaMAR in terms of diversity.

Table 2. **Sensor specifications.** The different platforms in CroCoDL with their recorded sensors. \*: GPS is only available outdoors, iOS only exposes anonymized BT GUIDs.

Platf.	Sensors		Cameras			Raw	Dev.	BT	WiFi	GPS
	#	FOV	RGB/GS	Depth	Freq.	IMU	Odom.			
Spot	4	103°	VGA	VGA	15Hz	✗	✓	✗	✗	✗
Azure Kinect	1	65°	FHD	VGA	15Hz	✓	✗	✗	✗	✗
ZED 2i	2	70°	QHD	HD	15Hz	✓	✗	✗	✗	✗
DJI Mini 4Pro	1	72°	4K	✗	30Hz	✗	✗	✗	✗	✓*
NavVis VLX	4	90°	FHD	(PC)	1-3m	✗	✓	✓	✓	✗
HoloLens2	4	83°	VGA	QVGA	30/5Hz	✓	✓	✓	✓	✗
iPad/iPhone	1	64°	FHD	FHD	10Hz	✓	✓	✓*	✗	✓*

2. **ARCHE B3:** The intact basement, bunker, and collapsed second story of a semi-collapsed building.
  3. **ARCHE B5:** Ruins exposed to the open sky.
  4. **ARCHE Grande Plaza:** An open plaza in front of a freight train, with an exhibition set up in tents.
- The next locations present visually unique challenges typically absent from SLAM datasets.
5. **Hydrology Lab:** An ETH experimentation hall and its underground facilities where tests on scaled models of water channels and dams are conducted.
  6. **Succulent Plant Collection:** A botanical museum in Zürich featuring connected greenhouses and a garden with a vast collection of succulents and cacti.
  7. **Design Museum Collection:** A poster storage room, basement, and staircases of an on-campus museum at Zürich University of the Arts.

Lastly, we enhance the existing LaMAR scenes with data collected using the Spot robot:

8. **HGE:** The main building of ETH Zürich.
9. **CAB:** The computer science building with offices and classrooms at ETH Zürich.
10. **LIN:** Part of the Zürich old town district.

### 3.2. Recorded sensors

The AR component of the dataset was recorded using a combination of iOS (iPad Pro and iPhone 13 mini) and HoloLens 2 devices; more details about their capturing application and sensors can be found in [14]. The robot system used for data capture is summarized in Table 2 and consists of a Boston Dynamics Spot quadruped with two additional front-facing camera systems, an Azure

Kinect Developer Kit and StereoLabs ZED2i. We also recorded with a DJI Mini 4 Pro drone which has an RGB camera with a gimbal for image stabilization purposes. However, the alignment of ground-truth pose data for the drone recordings is still under active development.

### 3.3. Ground-Truth Pipeline

While iOS and HoloLens 2 data pre-processing is implemented by LaMAR [14], Spot data pre-processing follows its own pipeline. First, rosbags are converted to the Capture format [12]. Extrinsic parameters are corrected using a calibration recording on-site, cameras which are recorded sideways or upside down are rotated upright to account for rotation-variant local features in the ground truth pipeline, sensor readings are sorted chronologically, and duplicates are removed. The ground truth pipeline requires chronologically synchronized sensor readings, which is achieved by creating virtual rigs. For a given set of temporally adjacent sensor readings, their timestamps are averaged, and the robot’s pose at this averaged timestamp is interpolated. For each sensor, the robot’s displacement between the averaged timestamp and the sensor reading’s actual timestamp is determined and incorporated into the transformation between that sensor and a selected rig base-frame. This process results in sensor measurements that share a common timestamp while adjusting their positions relative to the rig origin to account for their individual capture times. The precision of this is limited by the robot’s odometry rate of 50Hz.

Pre-processed data is fed into LaMAR [14] and follows their pipeline to generate accurate ground truth poses for all sessions. As shown in Figure 1, the footage from Spot is successfully aligned to the NavVis mesh, achieving comparable overall accuracy with respect to HoloLens and iOS devices results.

## 4. Evaluation

We focus on the evaluation of five of the new locations, which cover many of the challenging conditions present. ARCHE D2, Hydrology, and Design Museum are indoor locations with different features: The first is sparsely



query \ map												
	HL	iOS	Spot	NV	HL	iOS	Spot	NV	HL	iOS	Spot	NV
HL	0.88	0.69	0.48	0.75	0.90	0.75	0.54	0.86	0.96	0.86	0.73	0.97
iOS	0.64	0.80	0.31	0.68	0.67	0.82	0.36	0.78	0.81	0.91	0.56	0.90
Spot	0.65	0.60	0.96	0.60	0.70	0.63	0.97	0.68	0.85	0.83	0.98	0.90

(a) APGeM

(b) NetVLAD

(c) Overlap

Figure 3. **SuperPoint local features with LightGlue matcher combined with varying image retrieval methods.** Aggregated results of the five examined locations, normalized on the number of queries per device type. Percentage of correct pose estimation queries with rotation-translation thresholds of 5 degrees, 0.5 meters respectively. HL is HoloLens 2, NV is NavVis. Overlap uses the ground-truth to simulate a perfect retrieval (isolating feature matching).

illuminated, the last includes many repetitive hallways. The Succulent Plant Collection is both indoor and outdoor, with some occlusions from heavy foliage. ARCHE Grande Place features large structural changes as a series of large tents were set up and taken down.

**Methodology.** Per location and per device, a SfM model is created from each mapping set using [13]. Image retrieval followed by matching and pose estimation is then performed using each query set. In order to evaluate cross-device localization, the query set and mapping set belong to different devices, with all combinations evaluated. We start by retrieving top  $K$  mapping images for each query image using different retrieval methods. For devices with rigs (e.g., Spot, HoloLens), we retrieve  $K$  images for each camera in the rig. The query images are then matched to the retrieved ones to yield a set of tentative 2D-2D matches that are lifted to 2D-3D matches using the sparse SfM model. These matches are then provided to LO-RANSAC [4, 6] which robustly estimates the final pose using P3P [6] for single images or GP3P [8] for rigs. The rotation and translation recall correctness thresholds have been set to 5 degrees and 0.5 meters respectively due to limited ground truth accuracy of Spot poses in particular.

**Results with SOTA methods.** We report the aggregated results normalized by the number of queries per location in Figure 3.

**Is cross-device localization harder than same-device localization?** Across all retrieval methods, recall tends to be greatest when the same device is used for both mapping and querying, demonstrating the impact of heterogeneous sensor specifications such as field-of-view and viewpoint height on performance. We observe higher recall when the multi-camera systems, HoloLens 2 and Spot, are used as queries. Spot maps with iOS queries consistently demonstrate the lowest recall, demonstrating the limits of SOTA methods.

**Is retrieval or matching the bottleneck?** Figure 3 suggests that for many device pairs, retrieving images of a different device looking at the same location is something SOTA methods struggle with. If retrieval is

Table 3. **Comparison of SOTA methods for retrieval, extraction, and matching.** Recall of localization component combinations in the hardest tested scenario: iOS query images in a Spot-built map. Results from Hydrology Lab location. (noted with \*: unlike MAST3R, SuperPoint + LightGlue uses multi-view triangulation for the map)

Image Retrieval	Feature Extraction	Feature Matching	Recall (5 deg, 0.5 m)
APGeM			0.36
NetVLAD			0.42
Fusion	SuperPoint	LightGlue	0.43
OpenIBL			0.49
CosPlace			0.32
SALAD			<b>0.54</b>
NetVLAD	SIFT	LightGlue	0.25
NetVLAD	SuperPoint	LightGlue	0.42
NetVLAD	SuperPoint	SuperGlue	<b>0.43</b>
NetVLAD		LoFTR	0.40
Overlap (Top 10) (upper bound)	SuperPoint	LightGlue	0.67
Overlap (Top 1)	SuperPoint	LightGlue	0.57*
Overlap (Top 1)		MAST3R	0.34

skipped by using ground truth overlap, the aggregated recall improves for many device pairs, but localizing other devices in robot maps still has low recall. Table 3 investigates the pair of iOS queries in Spot maps further. We can conclude that cross-device image retrieval is overall more limiting right now, but even with perfect retrieval, there is a large potential for improvement of feature matching for relative pose estimation between robots and mixed-reality devices.

## 5. Conclusion & Outlook

We present a novel benchmark for cross-device visual registration that shows the considerable limitations of current state-of-the-art methods for both image retrieval and feature extraction and matching when using different devices for mapping and localization. CroCoDL presents a solid foundation on which to explore cross-device localization in more depth, and we intend to further extend both the dataset and benchmark.

**Acknowledgments** We thank Philipp Lindenberger and Petar Lukovic for their help, and ETH Foundation Project 2025-FS-352, SNSF Advanced Grant 216260, and Lamarr Institute for Machine Learning and AI for financial support.

## References

- [1] armasuisse. Advanced Robotic Capabilities for Hazardous Environments (ARCHE). <https://www.admin.ch/en/arc2023-en>, 2023. 2
- [2] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W. Achtelik, and Roland Siegwart. The EuRoC Micro Aerial Vehicle Datasets. *International Journal of Robotics Research*, 2016. 2
- [3] Nicholas Carlevaris-Bianco, Arash K. Ushani, and Ryan M. Eustice. The University of Michigan North Campus Long-Term Vision and LiDAR Dataset. *International Journal of Robotics Research*, 2015. 2
- [4] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally Optimized RANSAC. In *Joint Pattern Recognition Symposium*, 2003. 4
- [5] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are We Ready for Autonomous Drone Racing? The UZH-FPV Drone Racing Dataset. In *International Conference on Robotics and Automation*, 2019. 2
- [6] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 1981. 4
- [7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research*, 2013. 2
- [8] Gim Hee Lee, Bo Li, Marc Pollefeys, and Friedrich Fraundorfer. Minimal Solutions for Pose Estimation of a Multi-Camera System. *International Journal of Robotics Research*, 2016. 4
- [9] Michael Helmberger, Kristian Morin, Beda Berner, Nitish Kumar, Giovanni Cioffi, and Davide Scaramuzza. The Hilti SLAM Challenge Dataset. *IEEE Robotics and Automation Letters*, 2022. 2
- [10] Jianhao Jiao, Hexiang Wei, Tianshuai Hu, Xiangcheng Hu, Yilong Zhu, Zhijian He, Jin Wu, Jingwen Yu, Xupeng Xie, Huaiyang Huang, et al. FusionPortable: A Multi-Sensor Campus-Scene Dataset for Evaluation of Localization and Mapping Accuracy on Diverse Platforms. In *IEEE International Conference on Intelligent Robots and Systems*, 2022. 2
- [11] Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guérin, Gabriela Csurka, et al. Large-scale Localization Datasets in Crowded Indoor Spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2
- [12] Microsoft. LaMAR Benchmark - CAPTURE Specification. <https://github.com/microsoft/lamar-benchmark/blob/main/CAPTURE.md>, 2022. 3
- [13] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 4
- [14] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. LaMAR: Benchmarking Localization and Mapping for Augmented Reality. In *European Conference on Computer Vision*, 2022. 2, 3
- [15] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2
- [16] Xuesong Shi, Dongjiang Li, Pengpeng Zhao, Qinbin Tian, Yuxin Tian, Qiwei Long, Chunhao Zhu, Jingwei Song, Fei Qiao, Le Song, Yangquan Guo, Zhigang Wang, Yimin Zhang, Baoxing Qin, Wei Yang, Fangshi Wang, Rosa H. M. Chan, and Qi She. Are We Ready for Service Robots? The OpenLORIS-Scene Datasets for Lifelong SLAM. In *International Conference on Robotics and Automation*, 2020. 2
- [17] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *IEEE International Conference on Intelligent Robots and Systems*, 2012. 2
- [18] 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 *IEEE International Conference on Intelligent Robots and Systems*, 2020. 2
- [19] Hexiang Wei, Jianhao Jiao, Xiangcheng Hu, Jingwen Yu, Xupeng Xie, Jin Wu, Yilong Zhu, Yuxuan Liu, Lujia Wang, and Ming Liu. FusionPortableV2: A Unified Multi-Sensor Dataset for Generalized SLAM Across Diverse Platforms and Scalable Environments. *International Journal of Robotics Research*, 2024. 2
- [20] Yilin Zhu, Yang Kong, Yingrui Jie, Shiyu Xu, and Hui Cheng. GRACO: A Multimodal Dataset for Ground and Aerial Cooperative Localization and Mapping. *IEEE Robotics and Automation Letters*, 2023. 2
- [21] David Zuñiga-Noël, Alberto Jaenal, Ruben Gomez-Ojeda, and Javier Gonzalez-Jimenez. The UMA-VI Dataset: Visual-Inertial Odometry in Low-textured and Dynamic Illumination Environments. *International Journal of Robotics Research*, 2020. 2