CroCoDL: Cross-device Collaborative Dataset for Localization Supplementary Material

Hermann Blum[1,3], Alessandro Mercurio[1], Joshua O'Reilly[1], Tim Engelbracht[1], Mihai Dusmanu[2], Marc Pollefeys[1,2], Zuria Bauer[1]
[1]ETH Zurich [2]Microsoft [3]Lamarr Institute / Uni Bonn
blumh@uni-bonn.de & mihaidusmanu@microsoft.com & {pomarc, zbauer}@ethz.ch

Contents

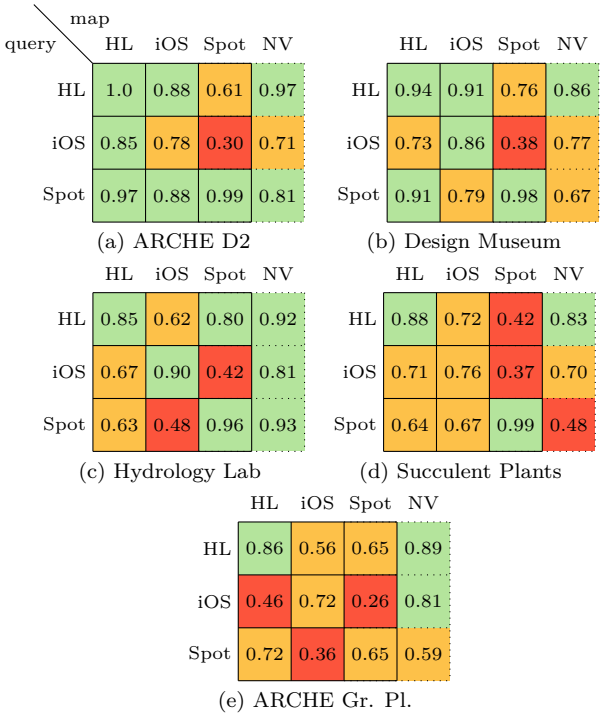
1. Additional quantitative and qualitative results

In this section we show additional quantitative device-pairs cross-localization results on each of the locations independently (Supplementary Section 1.1), as well as the best performing visual localization configuration showing the device queries plotted in the four available maps (Supplementary Section 1.2).

1.1. Confusion matrices

To benchmark visual localization on CroCoDL, we evaluated the configurations on each location independently, obtaining the following confusion matrices.

For all the matrices the percentages of correct pose estimation queries with rotation-translation thresholds is 5 degrees and 0.5 meters respectively. The axis of the matrices refer to HoloLens 2 as HL, and to the NavVis ground-truth scanner as NV.

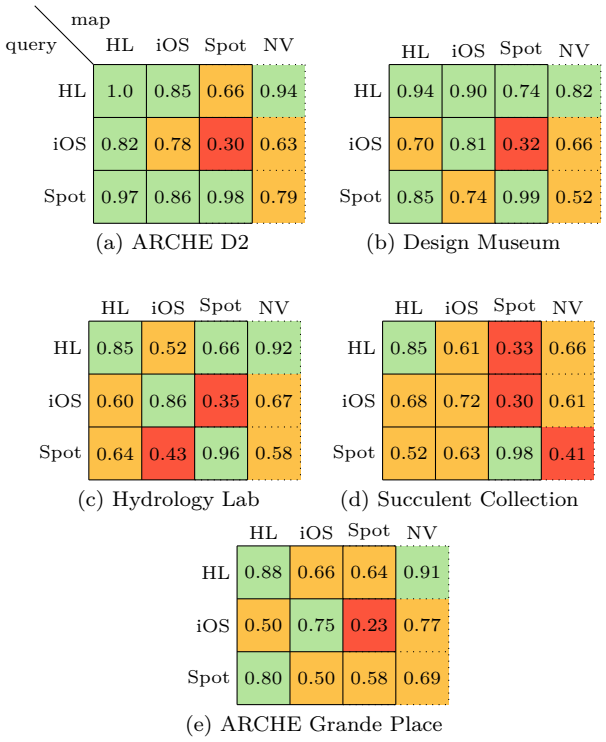


(a) ARCHE D2 (b) Design Museum

(c) Hydrology Lab (d) Succulent Plants

(e) ARCHE Gr. Pl.

Supp. Figure 1. **NetVLAD– SuperPoint– LightGlue.** Device-pairs cross-localization results in the five examined locations. Percentage of correct pose estimation queries with rotation-translation thresholds of 5 degrees, 0.5 meters respectively.

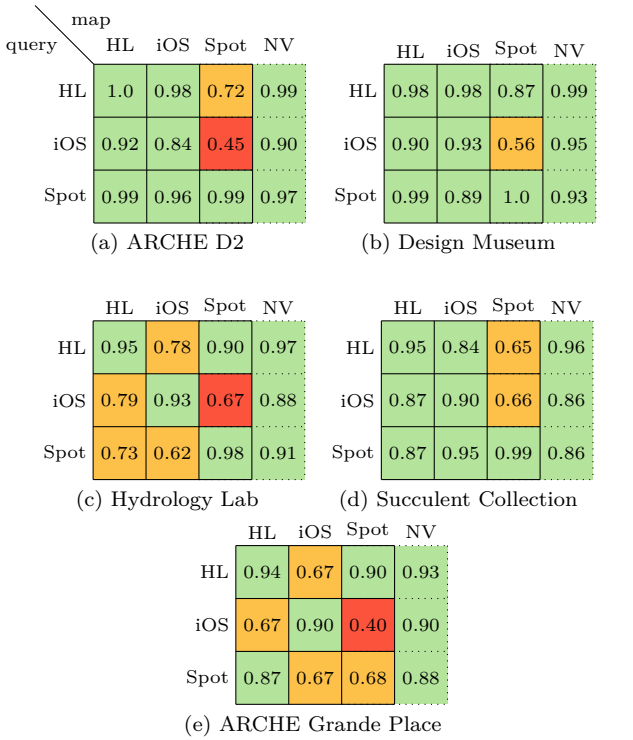


(a) ARCHE D2 (b) Design Museum

(c) Hydrology Lab (d) Succulent Collection

(e) ARCHE Grande Place

Supp. Figure 2. **APGeM–SuperPoint–LightGlue.** Device-pairs cross-localization results in the five examined locations.



(a) ARCHE D2 (b) Design Museum

(c) Hydrology Lab (d) Succulent Collection

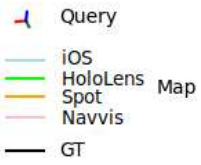
(e) ARCHE Grande Place

Supp. Figure 3. **Overlap–SuperPoint–LightGlue.** Device-pairs cross-localization results in the five examined locations. Since overlap-based image retrieval uses the ground-truth mesh, this focuses instead on feature extraction and matching.

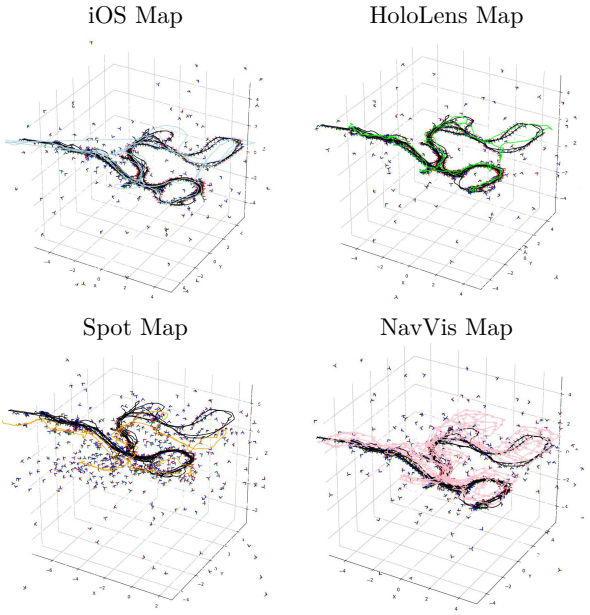
1.2. Pose estimation plots

This supplementary section contains 3D visualizations of mapping trajectories, ground truth (GT) query set poses, and query set estimated poses using the NetVLAD - SuperPoint - LightGlue visual localization configuration. We show results for two locations, the smaller ARCHE D2 and the larger, more challenging Succulent Plant Collection. For each location, we visualize the estimated poses of all three query devices (iOS, HoloLens, Spot) using the four available mapping devices (iOS, HoloLens, Spot, NavVis).

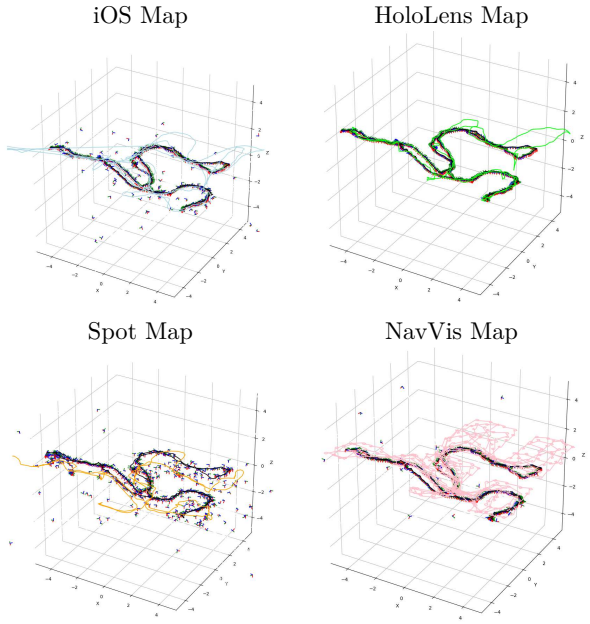
This is the common legend for all the plots in this section:



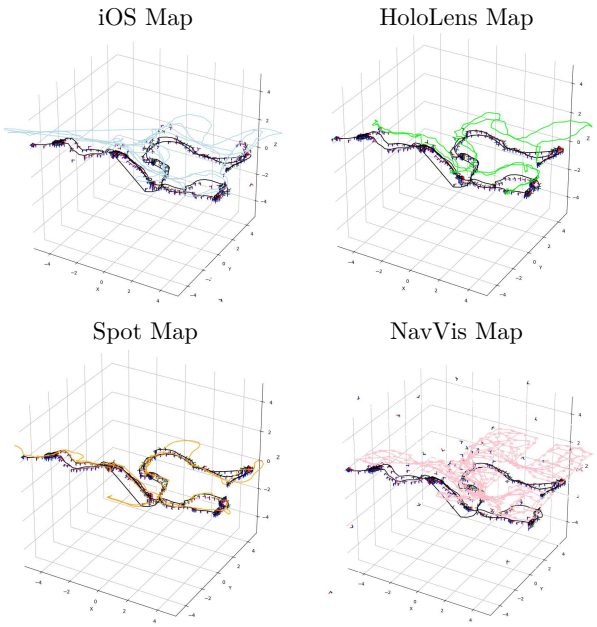
Supp. Figure 4. **Pose estimation legend.**



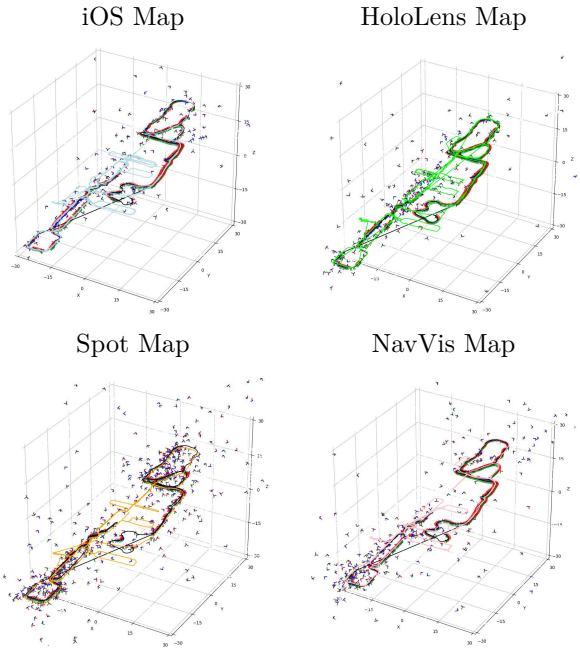
Supp. Figure 5. **Qualitative iOS pose estimation results in ARCHE D2.** Using the NetVLAD - SuperPoint - LightGlue visual localization configuration, we visualize the iOS queries in the four available maps.



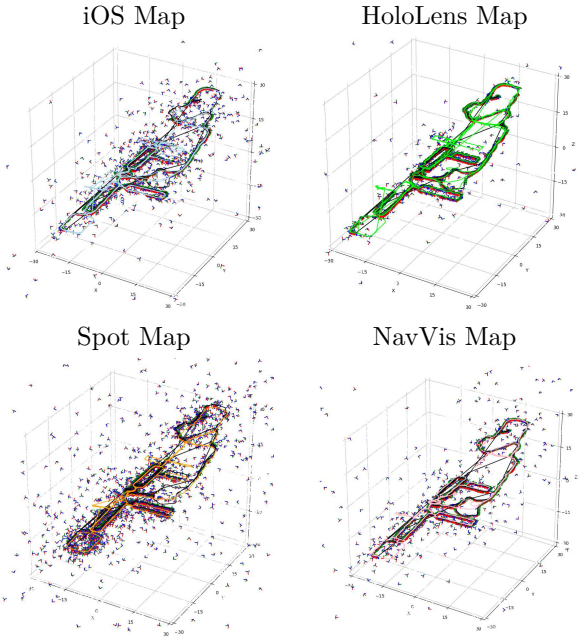
Supp. Figure 6. **Qualitative HoloLens pose estimation results in ARCHE D2.** Using the NetVLAD - SuperPoint - LightGlue visual localization configuration, we visualize the HoloLens queries in the four available maps.



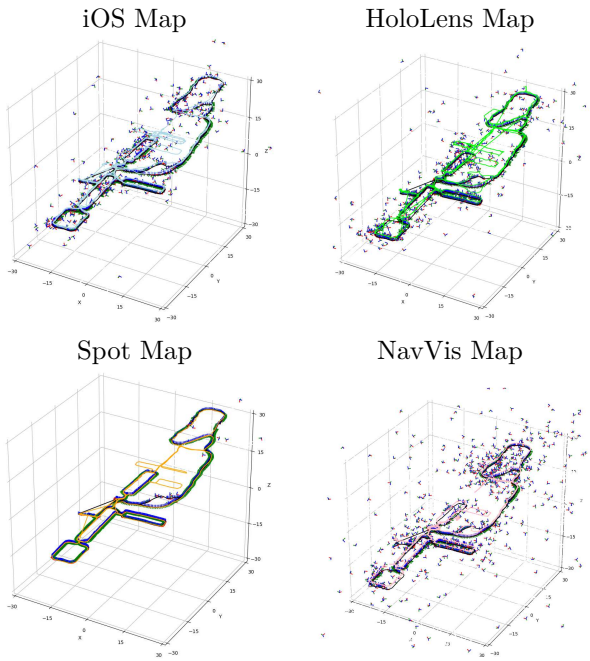
Supp. Figure 7. **Qualitative Spot pose estimation results in ARCHE D2.** Using the NetVLAD - SuperPoint - LightGlue visual localization configuration, we visualize the Spot queries in the four available maps.



Supp. Figure 8. **Qualitative iOS pose estimation results in the Succulent Plant Collection.** Using the NetVLAD - SuperPoint - LightGlue visual localization configuration, we visualize the iOS queries in the four available maps.



Supp. Figure 9. **Qualitative HoloLens pose estimation results in the Succulent Plant Collection.** Using the NetVLAD - SuperPoint - LightGlue visual localization configuration, we visualize the HoloLens queries in the four available maps.



Supp. Figure 10. **Qualitative Spot pose estimation results in the Succulent Plant Collection.** Using the NetVLAD - SuperPoint - LightGlue visual localization configuration, we visualize the Spot queries in the four available maps.

2. Additional visualizations

In this section, we group additional visualizations of the different locations (Supplementary Section 2.1), extra renderings from the collected dataset (Supplementary Section 2.2), and additional rendering alignments with the raw images from the various devices (Supplementary Section 2.3).

2.1. Images from the locations

Additional images of the different locations are shown in this section.

2.1.1. Hydrology Lab

Supp. Figure 11. **Additional images from the Hydrology Lab.** The left image showcases the aerial view of the Hydrology Lab, to portray the extensions of the recorded environment, while the right image shows the appearance of the basement below the lab.

2.1.2. Design Museum Collection

Supp. Figure 12. **Additional images from the Design Museum Collection.** The left image showcases the art covering the walls of the recorded Design Museum location. The right image is an archive room for posters. The Design Museum contains a mixture of wider spaces with art, and narrow archives with repetitive drawers on the walls.

2.1.3. Succulent Plant Collection

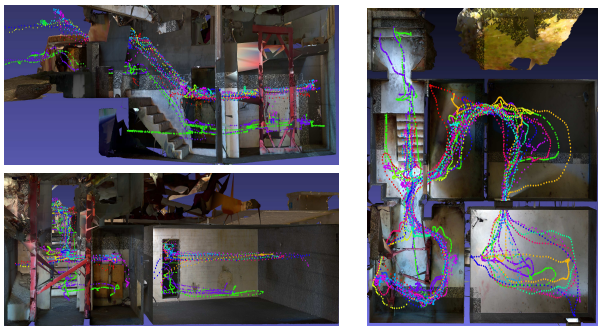
Supp. Figure 13. **Additional images from the Succulent Plant Collection.** Close-ups of various succulents, as well as one of the main hallways of the collection; this scene is composed of narrow hallways bordered by diverse plants indoors, and a plaza and garden outdoors.

2.2. Renderings and trajectories

This section contains point cloud renderings of the locations, as well as estimated ground truth trajectories inside them.

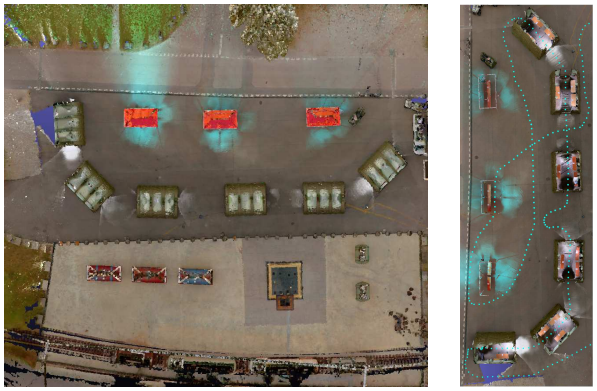
2.2.1. ARCHE D2

Supp. Figure 14. **Qualitative results of the point cloud renderings from the recordings in ARCHE D2.**



Supp. Figure 15. **Qualitative results showing recorded device trajectories in ARCHE D2.** Left: side-views of ARCHE D2 point cloud rendering. Right: top-down view of ARCHE D2 point cloud rendering. Device trajectories are shown in both; Spot trajectories are significantly closer to the ground than hand-held or head-mounted devices.

2.2.2. ARCHE Grande Place



Supp. Figure 16. **Qualitative results showing recorded device trajectories in ARCHE Grande Place.** Left: top-down view of entire location. Right: top-down section view of tent-covered segment of the scene with a Spot trajectory traversing it.

2.2.3. Hydrology Lab

Supp. Figure 17. **Qualitative results of the point cloud renderings from the recordings in the Hydrology Lab.** Left: main floor of the lab. Right: lab basement.

Supp. Figure 18. **Qualitative results showing recorded device trajectories in the Hydrology Lab.** Top-down view of the ground floor, with an iOS trajectory.

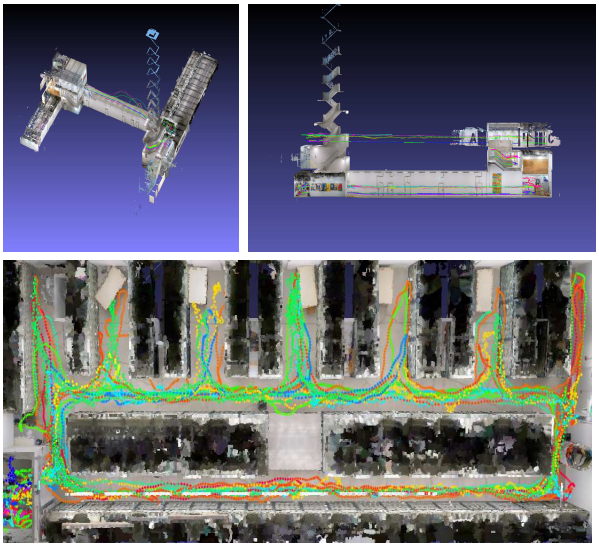
2.2.4. Succulent Plant Collection

Supp. Figure 19. **Qualitative results of the point cloud renderings from the recordings in the Succulent Plant Collection.** Left: entrance of collection. Right: aerial view of one of the rooms.

Supp. Figure 20. **Qualitative results showing recorded device trajectories in the Succulent Plant Collection.** Top-down view of plant collection with a device trajectory. Left: subset of scene interior. Right: subset of scene exterior.

2.2.5. Design Museum Collection

Supp. Figure 21. **Qualitative results of the point cloud renderings from the recordings in the Design Museum Collection.** Left: staircase with paintings and posters. Right: archive room.

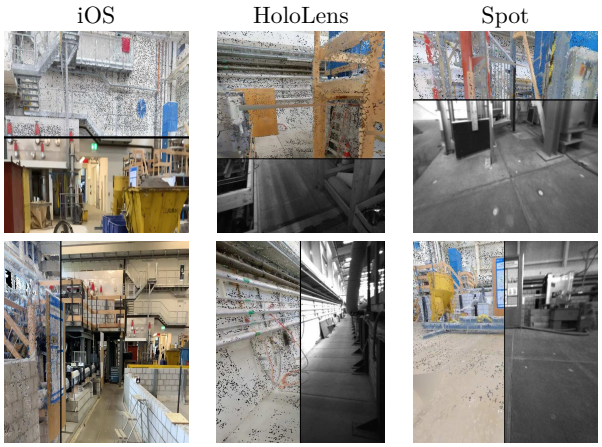


Supp. Figure 22. **Qualitative results showing recorded device trajectories in the Design Museum Collection.** All: point cloud renderings of location with multiple device trajectories. Top: aerial and side view of entire location. Bottom: top-down view of archive room.

2.3. Alignment with the renderings

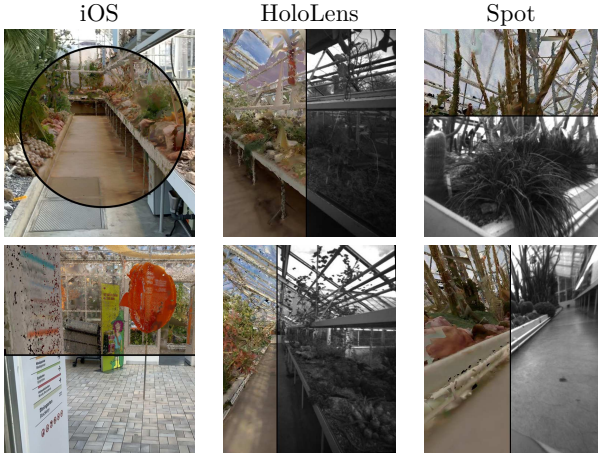
This section provides qualitative results of the GT alignment process, comparing images recorded by multiple devices (iOS, HoloLens, and Spot) to renderings from the NavVis mesh at the GT trajectory pose. This highlights the variety of the recorded dataset and the quality of the GT trajectory.

2.3.1. Hydrology Lab



Supp. Figure 23. **Qualitative results of the renderings vs the raw images recorded in the Hydrology Lab.** The images show the very good alignment between the rendered images and the recorded images by the different devices: iOS, HoloLens, and Spot. This is needed for accurate GT to evaluate against.

2.3.2. Succulent Plant Collection

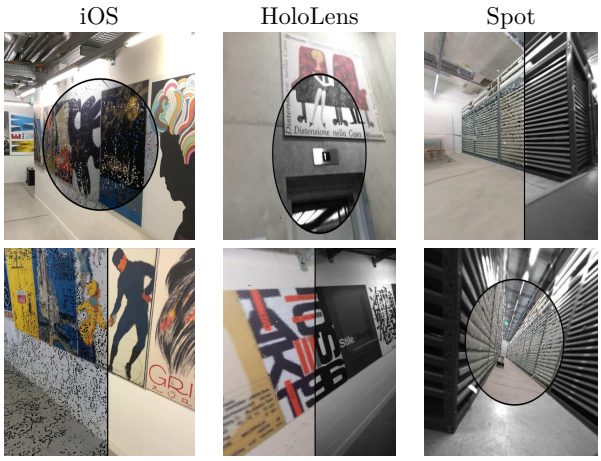


Supp. Figure 24. **Qualitative results of the renderings vs the raw images recorded in the Succulent Plant Collection.** The images show the very good alignment between the rendered images and the recorded images by the different devices: iOS, HoloLens, and Spot. This is needed for accurate GT to evaluate against.

```
CroCoDL
└─ SUCCULENT ..... Name of the location
   └─ sessions
      ├─ <timestamp1> ..... Session containing a NavVis scan
      ├─ <timestamp2>
      ├─ <timestamp3>
      ├─ <timestamp1>+<timestamp2>+<timestamp3> ..... Merged NavVis scans used for pseudo-GT
      ├─ ios_map ..... Session containing the iOS map set
      │  └─ trajectories.txt ..... GT aligned trajectory for the map
      ├─ ios_query ..... Session containing the iOS query set
      │  ├─ proc
      │  │  └─ aligned_trajectories.txt ..... GT aligned trajectory for the queries
      │  └─ queries.txt ..... Queries for pose estimation
      ├─ ios_* ..... Individual recordings in Capture format
      ├─ hl_map
      ├─ hl_query
      ├─ hl_*
      ├─ spot_map
      ├─ spot_query
      └─ spot_*
```

Supp. Figure 25. **Data structure.** We present the folder structure for the data of one location above.

2.3.3. Design Museum Collection



Supp. Figure 26. **Qualitative results of the renderings vs the raw images recorded in the Design Museum.** The images show the very good alignment between the rendered images and the recorded images by the different devices: iOS, HoloLens, and Spot. This is needed for accurate GT to evaluate against.

3. Data structure

The data structure is provided in Figure 25.

4. Drone data processing and alignment

For the acquisition of drone data a DJI Mini 4 Pro is used, providing monocular video with a resolution of 3840×2160 pixel at 30Hz. Additionally, per-frame synchronized GNSS (longitude, latitude) and metric altitude readings can be accessed. Note that the drone does not provide IMU measurements or odometry. Similar to the Spot data, drone data preprocessing is not implemented in the LaMAR pipeline and a separate data pipeline is therefore used to convert raw drone footage. To this end, we first calibrate the camera intrinsics using Kalibr [3] and then compute a scaleless trajectory using monocular Deep-Patch Visual Odometry [6]. This trajectory is then scaled to roughly metric scale as described below and then transformed into the capture format expected by the LaMAR [5] pipeline.

Outdoor Environments. For outdoor environments like ARCHE Grande, GNSS data is used to resolve the problems of scale-ambiguity, scale drift and pose drift of the monocular VO trajectory. We first transform the GNSS' longitude and latitude coordinates into the metric projected coordinate system for Switzerland and combine them with the altitude readings to generate a metric trajectory. We then find the transformation between odometry in camera coordinate system and the metric trajectory by scaling and aligning the first N frames of the trajectory using Procrustes analysis. We transform the metric trajectory into the iOS coordinate system and substitute the translational part of the odometry with it, thus scaling and eliminating drift from the trajectory, while keeping VO orientations. It is important to note that while this approach introduces local coarseness in the trajectory, it ensures global consistency. Further refinement of the trajectory is subsequently carried out through bundle adjustment and pose-graph optimization during the LaMAR alignment process.

Indoor Environments. In indoor environments without GPS reception, such as HYDRO, we are limited to using only metric altitude readings to scale the VO trajectory. However, this approach has two main limitations. Firstly, in long trajectories with inevitable drift, global consistency cannot be guaranteed. Secondly, altitude measurements are provided in 0.1-meter increments which introduces scale inaccuracies. To address the first issue, we restrict ourselves to trajectories that exhibit only negligible drift. For instance, trajectories containing degenerate motions, such as fast or pure rotations, are split up at such motions, mitigating their impact. To handle scale inaccuracies, we compute the median scale of the trajectory, averaging out errors. Additionally, we only include trajectories that demonstrate measurable changes in altitude, ensuring reliable scaling information.

Table 1. **Drone-on-X localization results.** We report recall at 5deg, 0.5m and 10deg, 1m for drone-on-X localization using NetVLAD–SuperPoint–LightGlue.

Drone-on-X	HL	iOS	Spot	NV
ARCHE D2	72.8%/78.9%	76.2%/80.3%	N/A	85.0%/87.8%
ARCHE Grande	73.5%/77.7%	82.7%/84.6%	19.8%/35.2%	74.7%/83.3%

Supp. Figure 27. **Qualitative results fon ARCHE D2 data from the Drone aligned with the GT from the NavVis scanner.**

Results. We show some initial results of the alignment in Figure 27. We also report some results for drone-on-X localization using NetVLAD [1] image retrieval coupled with SuperPoint [2] feature extraction and LightGlue [4] matcher in Table 1. These results support the take-home messages from the main paper. The drone performs best in iOS and NavVis maps due to the similarities in camera (rolling shutter, RGB) and viewpoint. Second best is the location in HL maps which do share a similar viewpoint, but the sensors are different. Finally, the localization in Spot maps is very poor due to very limited viewpoint overlap. In the ARCHE D2 location, due to movement restrictions and sensor setups for Spot and Drone, we were so far unable to get data with significant visual overlap.

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 9

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 9

[3] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified Temporal and Spatial Calibration for Multi-Sensor Systems. In *IEEE International Conference on Intelligent Robots and Systems*, 2013. 9

[4] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. *International Conference on Computer Vision*, 2023. 9

[5] 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. 9

[6] Zachary Teed, Lahav Lipson, and Jia Deng. Deep Patch Visual Odometry. In *Advances in Neural Information Processing Systems*, 2023. 9