

Seeing in the Dark: Benchmarking Egocentric 3D Vision with the Oxford Day-and-Night Dataset (Supplementary)

[HuggingFace: active-vision-lab/oxford-day-and-night](https://huggingface.co/active-vision-lab/oxford-day-and-night)

A Full Dataset Statistics

We collected our dataset across five locations in Oxford by walking while wearing ARIA glasses. The data collection took place over the course of one month. During this period, two collectors wore ARIA glasses and walked randomly within each collection site. In total, the walking trajectory spans 30 kilometers, includes 7 hours of walking, and covers an area of 40,000 m².

Detailed dataset statistics are provided in Tab. 6. Notably, our dataset offers a well-balanced distribution of day and night recordings, with an approximately 1:1 ratio. The covered areas and walking trajectories are visualized in Figs. 6 and 7.

Table 6: **Dataset Statistics.** We present a summary of the number of frames in the recorded videos, the NVS data variant (obtained by subsampling the video by 5×), and the visual relocalization data variant (with additional spatial subsampling and splitting into database, daytime queries, and nighttime queries). We also report the recording durations, trajectory lengths, and mapped area sizes.

Scene	# Video Fr	# NVS Img	# Visual Reloc Img			Duration (hh:mm)			Trajectory Len (m)			Area (m ²)
	D & N	D & N	DB	Day Q	Night Q	Day	Night	D & N	Day	Night	D & N	D & N
Bodleian Lib.	205405	41081	2542	1310	2908	01:32	01:18	02:50	7170	5617	12787	25939
H.B. Allen Cen.	29340	5868	305	158	449	00:13	00:10	00:24	975	765	1740	1271
Keeble College	112205	22441	1020	526	1511	00:46	00:46	01:33	3574	3400	6974	5709
Obs. Quarter	87210	17442	821	424	1342	00:34	00:38	01:12	2853	3050	5903	5950
Robotics Inst.	57590	11518	778	401	987	00:25	00:22	00:47	1249	1030	2279	600
Total	491750	98350	5466	2819	7197	03:32	03:16	06:48	15822	13862	29685	39469

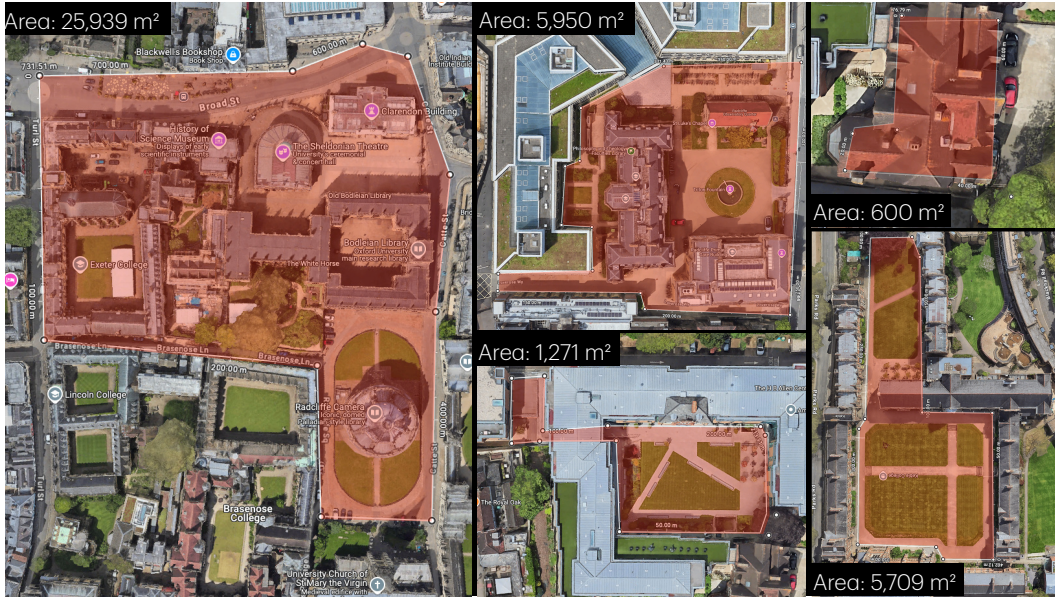


Figure 6: Our dataset covers 40,000 m² area.

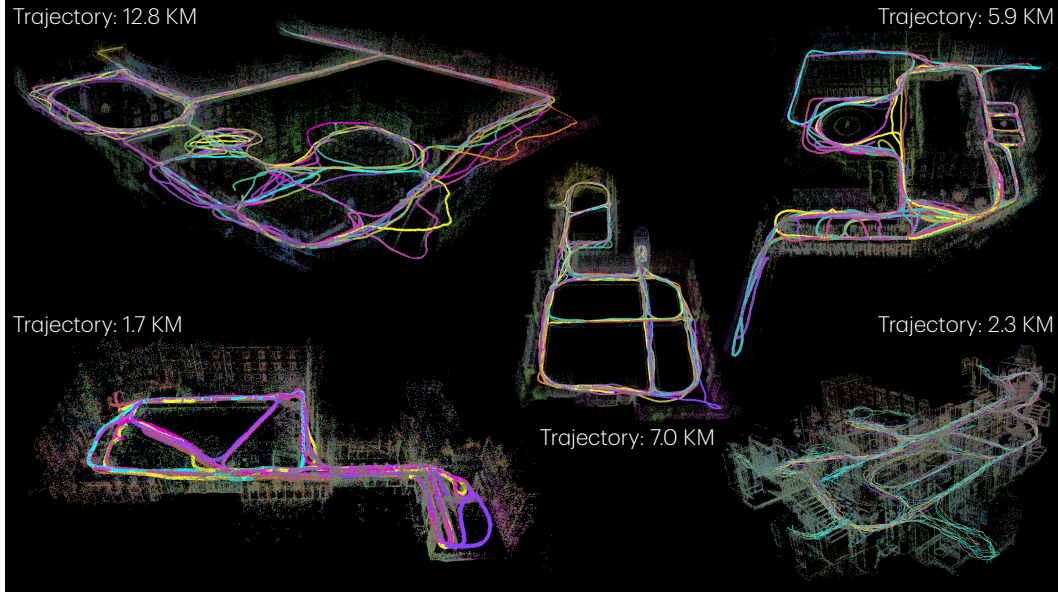


Figure 7: Our dataset spans 30 kilometers of walking trajectory.

B Image Variants

ARIA glasses are equipped with fisheye lenses, resulting in fisheye distortion in the original recordings. To facilitate the use of our dataset, we undistort these images using two different pinhole camera configurations. We provide both the original fisheye images and the undistorted versions. The three image variants are visualized in Fig. 8.



Figure 8: We provide three image types: the original fisheye, a *Max FOV* undistorted version with wider coverage and black borders (with a valid pixel mask provided), and an *All Valid* version with no black borders but a smaller field of view for easier use.

C Additional Details on Visual Relocalization Dataset

Spatial Filtering. We provide additional details about our visual relocalization dataset. In Algorithm 1, we present the pseudo-code for the spatial filtering algorithm used to eliminate redundant images when generating the database, daytime query, and nighttime query splits. For outdoor scenes (Bodleian Library, H.B. Allen Centre, Keble College, Observatory Quarter), we use thresholds of $\theta_{\text{pos}} = 1.5$ meters and $\theta_{\text{ori}} = 20^\circ$; for the indoor Robotics Institute scene, we adopt stricter thresholds of $\theta_{\text{pos}} = 0.5$ meters and $\theta_{\text{ori}} = 20^\circ$ to reflect its smaller scale. Notably, even after applying strong spatial filtering, our dataset includes 7,197 nighttime query images, 37 times more than the 191 nighttime queries in the Aachen Day-Night dataset [5]. Full statistics are summarized in Tab. 6.

Algorithm 1: Spatial Filtering of Camera Poses

Require: Image list I with poses (p_i, R_i) , thresholds $\theta_{\text{pos}}, \theta_{\text{ori}}$

Ensure: Filtered image list S

```

1: Shuffle  $I$ ; initialize  $S \leftarrow []$ , cache  $C \leftarrow []$ 
2: for each image  $i$  in  $I$  do
3:    $N \leftarrow \{(p_j, R_j) \in C \mid \|p_i - p_j\| < \theta_{\text{pos}}\}$ 
4:   if  $N = \emptyset$  or  $\forall (p_j, R_j) \in N, \angle(R_i, R_j) > \theta_{\text{ori}}$  then
5:     Append  $i$  to  $S$ , append  $(p_i, R_i)$  to  $C$ 
6:   end if
7: end for
8: return  $S$ 

```

Coverage of Nighttime Queries Across Distance and Rotation Thresholds. To further illustrate the challenge posed by our benchmark, Fig. 9 plots the percentage of nighttime queries that have at least one daytime database image within a specified spatial and angular threshold. Our dataset spans a wide spectrum of difficulty levels, including a particularly challenging subset: nighttime queries that are more than 5 meters and 50° away from any corresponding database image. These difficult cases account for approximately 10% of the nighttime queries. This diversity enables a more fine-grained evaluation of relocalization methods, allowing the community to assess performance across both easy and hard cases.

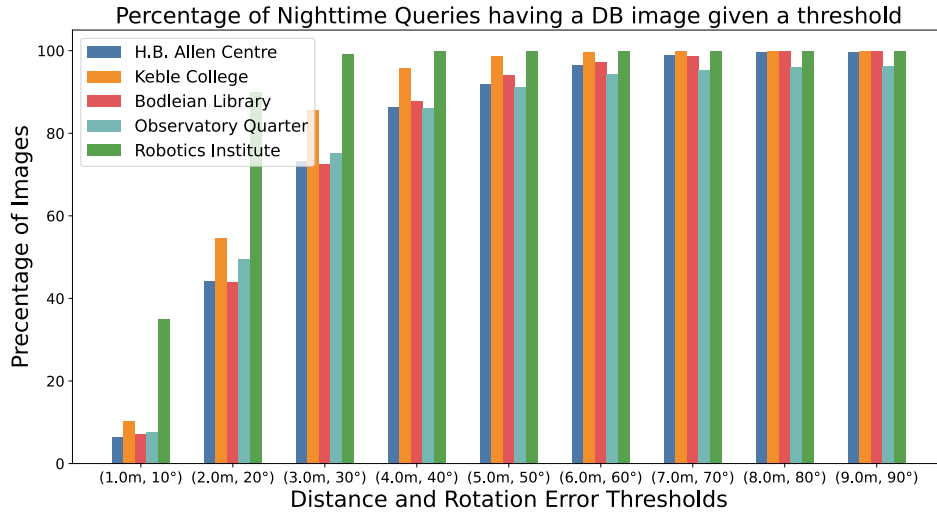
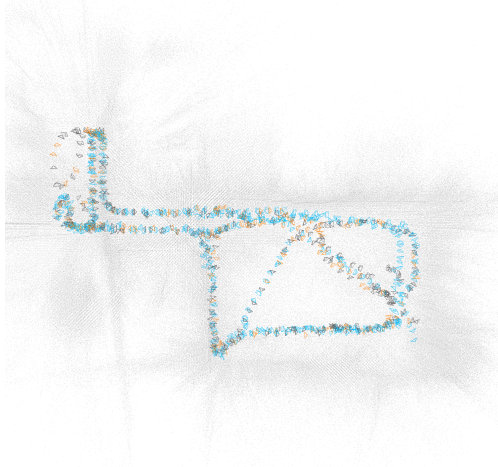
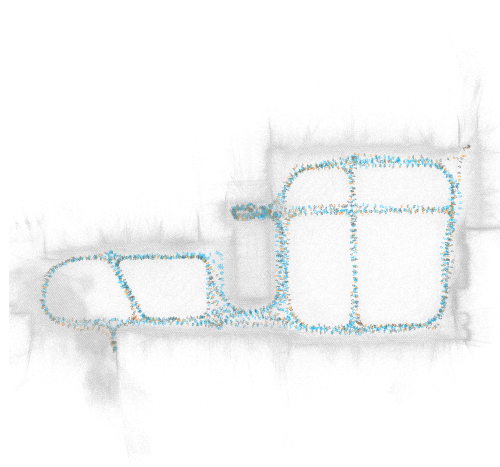


Figure 9: The percentage of nighttime queries that have a database image given a spatial and orientation threshold.

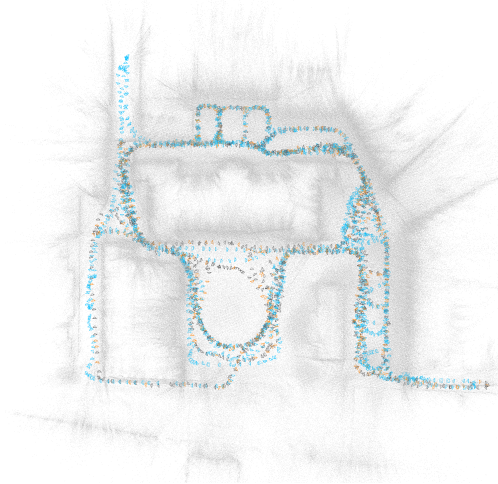
Database Creation, COLMAP, and HLoc. We structure our relocalization dataset using simple image lists, where each split (database, daytime queries, and nighttime queries) corresponds to a text file containing the image filenames relative to the image directory. To facilitate seamless integration with the HLoc toolbox [69], we also provide a COLMAP model for the database images, generated using ARIA MPS output poses. Specifically, for each database image, we project the 3D point cloud of the scene onto the image plane using the corresponding ground-truth camera pose. We then apply a series of filtering steps to remove invalid projections: depth filtering, image boundary checks, and z-buffer visibility checks. From the valid set of projections, we randomly sample 3,000 2D-3D correspondences per image. Using this information, we construct the `images.bin`, `cameras.bin`, and `points3D.bin` files following COLMAP standard format. Note that our COLMAP model does not incorporate explicit occlusion reasoning. As a result, we do not recommend using it directly for PnP-RANSAC without additional filtering or refinement. However, this limitation does not affect integration with the HLoc toolbox, as it does not rely on the database point cloud. We provide the visualization of the distribution of database, daytime, and nighttime camera poses in each scene in Fig. 10.



(a) H.B. Allen Centre



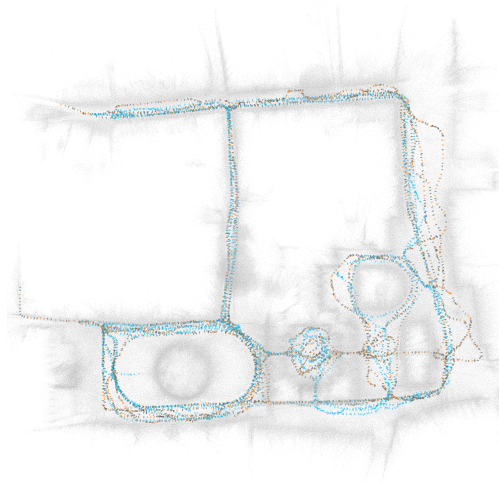
(b) Keble College



(c) Observatory Quarter



(d) Robotics Institute



(e) Bodleian Library

Figure 10: **Camera poses for visual relocalization in each scene.** The cameras of database images are in **black**; the cameras of day query images are in **orange** and the cameras of night query images are in **blue**.

D Additional Experiments on Visual Relocalization Dataset

We also report image retrieval evaluation results using RoMA [61] within the HLoc pipeline [48] across all five scenes, as shown in Tab. 4. The results confirm that the primary bottleneck in the HLoc pipeline lies in the image retrieval stage rather than the image matcher itself. Even when using the most advanced retrieval method, MegaLoc [68], a noticeable performance gap remains compared to retrieval based on ground-truth poses, indicating clear room for future improvements in retrieval.

Table 7: **Accuracy of RoMA on the visual relocalization dataset with using HLoc pipeline and different image retrieval methods.** We report the percentage of correctly localized query images within thresholds of (0.25m, 2°), (0.5m, 5°), and (1m, 10°).

	Bodleian Library	H.B. Allen Centre	Keble College	Observatory Quarter	Robotics Institute
RoMA + NetVLAD 50 [57]	25.24 / 30.98 / 35.18	57.91 / 74.39 / 79.06	14.96 / 22.63 / 30.91	58.94 / 66.92 / 70.79	72.34 / 75.38 / 76.60
RoMA + DIR 50 [65, 66]	33.46 / 39.10 / 42.30	55.46 / 72.16 / 81.51	16.48 / 23.36 / 28.33	56.33 / 65.28 / 69.90	74.27 / 78.52 / 79.84
RoMA + OpenIBL 50 [67]	43.95 / 51.24 / 54.50	60.36 / 73.50 / 78.62	18.66 / 27.47 / 35.67	59.09 / 66.32 / 70.19	71.73 / 75.18 / 76.19
RoMA + MegaLoc 50 [68]	70.25 / 79.09 / 82.22	66.37 / 81.51 / 87.31	31.50 / 42.22 / 51.82	72.06 / 80.92 / 84.50	78.22 / 82.27 / 83.69
RoMA + GT Pose 20	80.57 / 89.58 / 92.88	68.82 / 81.51 / 85.75	41.50 / 57.78 / 71.01	80.55 / 87.48 / 90.98	84.50 / 89.36 / 90.48

E Additional Results on NVS Dataset

Image Quality. We provide additional NVS evaluation results in Tab. 8 and Fig. 11, which complement the findings presented in Tab. 5 and Fig. 5. Specifically, Table 8 highlights that both 3DGS in-the-wild methods exhibit limited NVS performance on our dataset, as indicated by high LPIPS values. Note that PSNR and SSIM values do not capture this performance degradation.

Table 8: **3DGS In-the-Wild Image Quality.** We report image rendering quality in PSNR (\uparrow) / LPIPS (\downarrow) / SSIM (\uparrow). This table complements Tab. 5 and Fig. 5 by providing additional SSIM scores.

Method	Bodleian Library	H.B. Allen Centre	Keble College	Observatory Quarter	Robotics Institute
Splatfacto-W [46]	25.98 / 0.60 / 0.79	25.65 / 0.59 / 0.81	27.96 / 0.59 / 0.78	25.83 / 0.63 / 0.78	22.73 / 0.61 / 0.81
Gaussian-Wild [47]	28.38 / 0.56 / 0.86	24.94 / 0.59 / 0.86	30.92 / 0.56 / 0.84	28.57 / 0.60 / 0.86	25.05 / 0.57 / 0.88

Geometry. Figure 11 visualizes the centers of Gaussian primitives after Splatfacto-W training. During this process, the initialized point cloud is culled to a reasonable density in the *H.B. Allen Centre* and *Observatory Quarter*. In contrast, the same culling procedure results in degenerate representations in the *Bodleian Library* and *Keble College* scenes, possibly due to the larger spatial extent of the *Bodleian Library* and the more extreme lighting variations present in *Keble College*.

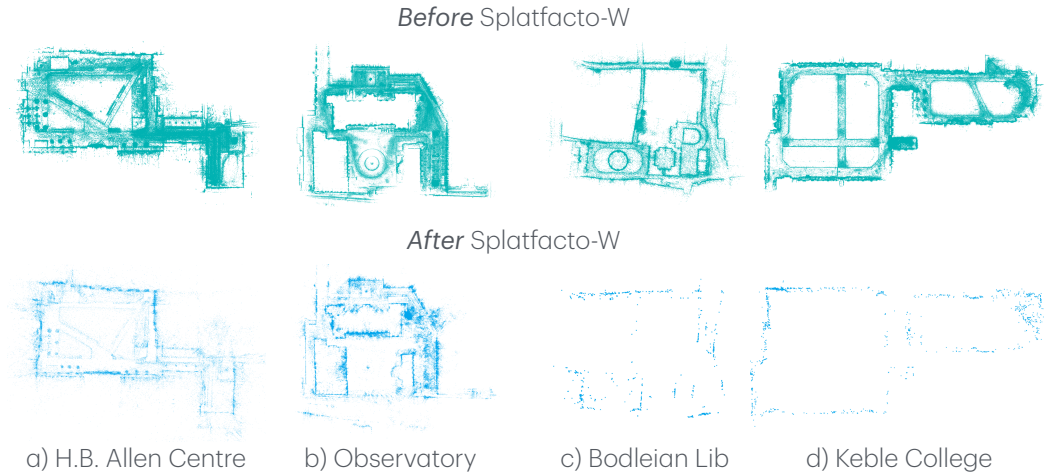


Figure 11: **Visualisation of 3D Geometry.** In c) and d), less than 2000 Gaussian primitives remain after the culling process during training. This may be due to limited capability in handling large-scale scenes and dramatic light variations, resulting in a degenerated case for 3DGS rendering.

Overall, our experiments demonstrate that current 3DGS in-the-wild methods continue to face significant challenges in large-scale scenes with dramatic lighting variations.