Benchmarking Localization, 3D Reconstruction and Radiance Fields for Navigation Across Day and Night

Jiahao Wang[1], Jianeng Wang[1], Yifu Tao[1], Zirui Wang[2], Victor Adrian Prisacariu[2], Maurice Fallon[1]

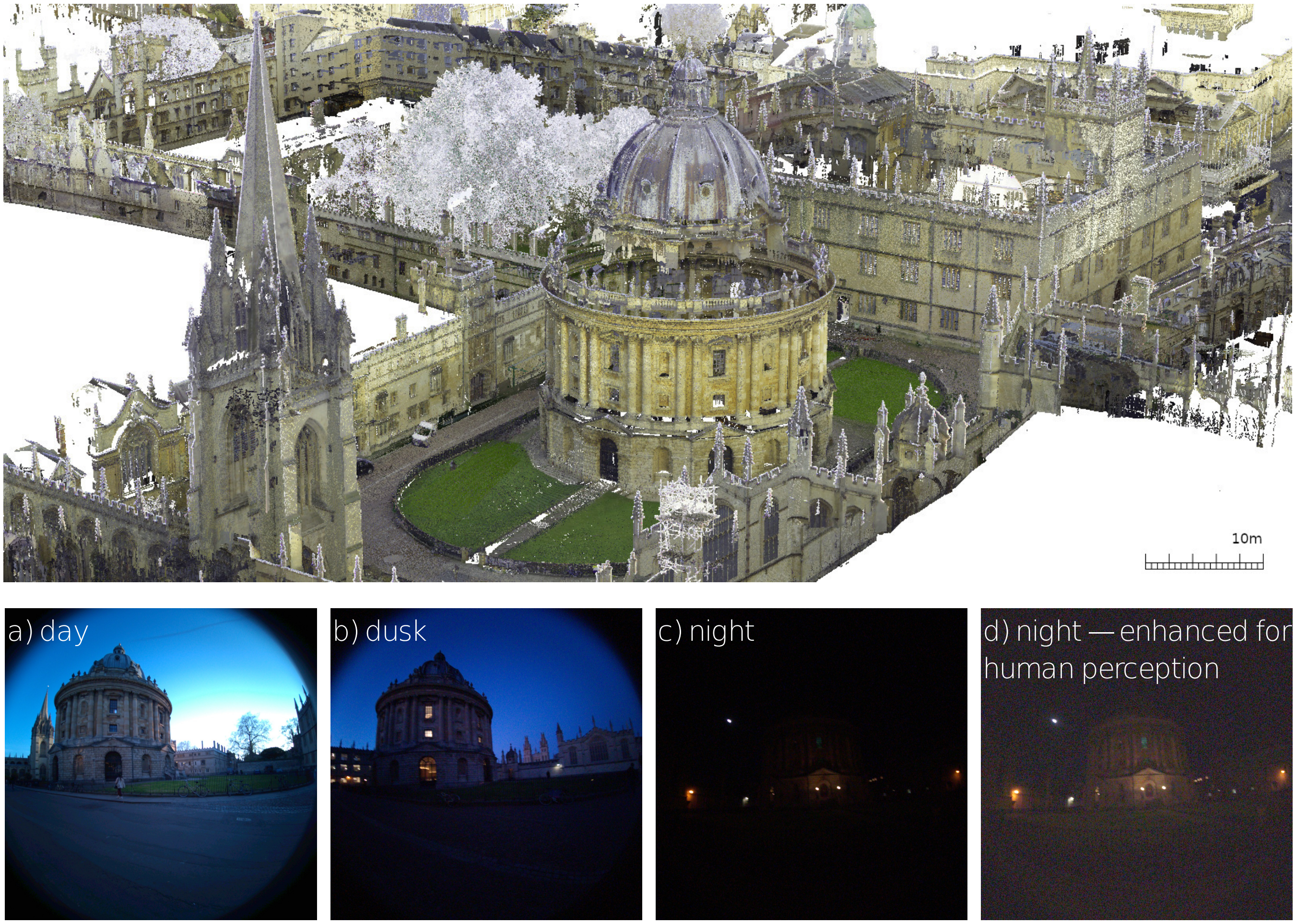

Fig. 1: Top: millimeter-accurate 3D ground truth of Bodleian Library from Oxford Spires dataset. Bottom: egocentric video data captured across different lighting conditions throughout the day from Oxford Day and Night dataset.

***Abstract*— For robots operating across day and night, the ability to localize and model the environment under varying illumination is essential. We present two complementary datasets for developing and benchmarking perception algorithms, including localization, 3D reconstruction, and novel-view synthesis in large-scale indoor and outdoor environments. The first, Oxford Spires Dataset, provides multi-sensor recordings around historical Oxford landmarks, paired with millimeter-accurate 3D ground-truth maps enabling precise trajectory estimation. The second, Oxford Day and Night Dataset, captures egocentric video in the same areas under diverse lighting conditions, from daylight to nighttime. Together, these datasets offer a unique platform for advancing robust perception methods capable of handling challenging illumination changes.**

I. Introduction

Robust localization and 3D reconstruction under varying illumination are fundamental for robot navigation in both indoor and outdoor environments. To avoid obstacles and plan effectively, a robot must estimate its position while modeling the surrounding 3D structure. Recent advances in radiance field methods, such as Neural Radiance Fields (NeRF) [1] and 3D Gaussian Splatting (3DGS) [2], further enable photorealistic rendering that can support learning-based navigation.

Despite progress in SLAM and reconstruction, existing datasets often lack the combination of large-scale coverage, accurate 3D ground truth, and illumination diversity. Outdoor 3D ground truth is especially rare, as survey-grade reference

[1] These authors are with Oxford Robotics Institute, Univ. of Oxford, UK. {jiahaowang,jianeng,yifu,mfallon}@robots.ox.ac.uk
[2] These authors are with the Active Vision Lab, University of Oxford, UK. {ryan,victor}@robots.ox.ac.uk

TABLE I: ATE result (using RMS) versus the provided ground truth for several open source algorithms (using OSD). Best results are maked with blue tints (darker is better). SC-LIO-SAM fails on some sequences. COLMAP gives incomplete results on some sequences.

Site	Sec	Len	VILENS-SLAM	Fast-LIO-SLAM	SC-LIO-SAM	ImMesh	Fast-LIVO2	HBA	COLMAP
Blenheim Palace	01	490	0.47	0.18	6.74	0.27	0.14	0.21	0.08
	02	390	0.16	0.12	4.41	0.36	0.22	0.08	0.05
	05	390	1.05	0.28	✗	0.22	0.26	0.14	0.26
Keble College	02	290	0.06	0.25	1.26	0.08	0.95	0.11	0.05
	03	280	0.14	0.11	4.02	0.14	0.06	0.12	0.05
	04	780	0.16	0.49	✗	3.67	0.09	0.12	0.07
	05	710	0.11	0.29	✗	0.13	0.11	0.13	0.09
Observatory Quarter	01	400	0.06	0.17	0.23	0.20	0.04	0.05	0.07
	02	390	0.09	0.24	0.14	0.27	0.07	0.08	0.08

models are costly and difficult to obtain. This limits rigorous benchmarking of SLAM and mapping, and constrains evaluation of radiance field methods, which require precise poses and geometry. Most available datasets either provide large-scale coverage without illumination variation, or egocentric recordings without reliable geometric ground truth.

To address these gaps, we introduce two complementary datasets. The *Oxford Spires Dataset (OSD)*[1] [3] provides multi-sensor recordings around historical Oxford landmarks, paired with millimeter-accurate 3D ground truth for benchmarking localization and reconstruction (Fig. 1**Top**). The *Oxford Day-and-Night Dataset (OXDAN)*[2] [4] captures egocentric video across times of day to evaluate robustness to illumination changes (Fig. 1**a**–**d**). Together, these datasets enable research in SLAM, 3D reconstruction, radiance field learning, and visual relocalization under open-world conditions.

II. Datasets

A. Oxford Spires Dataset - Frontier Device

The Oxford Spires Dataset is a large-scale, multi-sensor resource covering both indoor and outdoor environments with dynamic motion and diverse lighting. It spans multiple historical landmarks in Oxford, each averaging one hectare in area, with sequence lengths typically exceeding 400 m.

Data were collected using a custom handheld perception unit, *Frontier*, equipped with a 64-beam LiDAR, a high-frequency IMU, and three wide-angle global-shutter cameras (Fig. 2**i**). All sensors are hardware-synchronized to ensure precise temporal alignment. The cameras are arranged forward, left, and right to provide a panoramic field of view with substantial overlap with LiDAR, facilitating robust multi-sensor fusion.

Each site is paired with a millimeter-accurate ground-truth scan obtained from a survey-grade Terrestrial LiDAR Scanner (TLS). These scans enable rigorous evaluation of 3D reconstruction and novel-view synthesis. By registering *Frontier* LiDAR data to the TLS model, centimeter-accurate ground-truth trajectories can also be derived for SLAM evaluation [5].

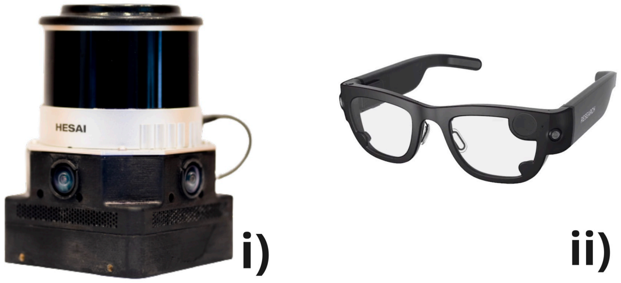

Fig. 2: These two Oxford-based datasets use these multi-sensor perception units for data recording **i)** *Frontier* and **ii)** Meta Aria glasses.

B. Oxford Day and Night Dataset - Meta Aria

The Oxford Day-and-Night Dataset complements the sensor-rich Frontier platform with a large-scale, vision-only resource captured from an egocentric perspective. Recording sites overlap with those of the Oxford Spires Dataset, enabling reuse of the same ground-truth scans. Its primary purpose is to benchmark visual relocalization and novel-view synthesis (NVS) algorithms under the extreme challenge of day-to-night illumination changes.

Data were collected using Meta Aria glasses [6], a research platform well suited for large-scale egocentric recording (Fig. 2**ii**). The glasses integrate a high-resolution RGB camera, two global-shutter grayscale cameras for motion tracking, and dual high-frequency IMUs. Sensor data are processed by a cloud-based, multi-session SLAM service that robustly aligns trajectories recorded at different times—from bright daylight to complete darkness—into a unified coordinate frame.

These trajectories are further registered to the survey-grade TLS scans from the Oxford Spires Dataset, providing an accurate reference for evaluating visual relocalization and NVS methods.

III. Benchmarks

In this section, we introduce four benchmarks—SLAM & SfM, 3D reconstruction, visual relocalization, and novel view synthesis—to demonstrate the capabilities that can be evaluated when using our dataset. These benchmarks

[1] website: https://dynamic.robots.ox.ac.uk/datasets/oxford-spires

[2] website: https://oxdan.active.vision/

TABLE II: Quantitative evaluation of the 3D reconstructions from VILENS-SLAM, OpenMVS and Nerfacto - using the Oxford Spires Dataset (OSD). We indicate the best results with a dark blue background.

Site	SEC.	Method	Accuracy↓	Completeness↓		5cm			10cm	
					Precision	Recall	F-score	Precision	Recall	F-score
Blenheim Palace	05	VILENS-SLAM	0.070	0.506	0.670	0.392	0.495	0.867	0.661	0.750
		OpenMVS	0.126	1.045	0.451	0.251	0.323	0.574	0.381	0.458
		Nerfacto	0.302	0.676	0.232	0.094	0.134	0.388	0.257	0.309
Keble College	04	VILENS-SLAM	0.067	0.342	0.527	0.527	0.527	0.816	0.779	0.797
		OpenMVS	0.050	0.409	0.766	0.606	0.677	0.918	0.718	0.806
		Nerfacto	0.137	0.150	0.418	0.484	0.449	0.654	0.709	0.680
Observatory Quarter	01	VILENS-SLAM	0.047	0.233	0.708	0.536	0.610	0.909	0.806	0.854
		OpenMVS	0.048	0.622	0.745	0.470	0.577	0.902	0.618	0.734
		Nerfacto	0.197	0.398	0.415	0.395	0.405	0.587	0.598	0.592

highlight the dataset's versatility and its potential to support a wide range of research directions.

A. SLAM & SfM Benchmark (OSD)

This benchmark evaluates the pose estimation accuracy of LiDAR-based SLAM and vision-based SfM systems.

1) Benchmarked Methods: We evaluate several representative approaches: online LiDAR-based methods (VILENS-SLAM [7], Fast-LIO-SLAM [8], SC-LIO-SAM [9], ImMesh [10]), offline LiDAR bundle adjustment (HBA) [11]), and a vision-only SfM pipeline (COLMAP [12]).

2) Evaluation Metrics: Predicted trajectories are compared against ground truth trajectories derived from survey-grade TLS scans by registering LiDAR point clouds to the TLS map using Iterative Closest Point (ICP) [13]. We report Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) using the *evo* package [14]. ATE measures global consistency after SE(3) Umeyama alignment, while RPE captures local drift and is recommended for evaluating odometry-style systems. The results are reported in Table I.

B. 3D Reconstruction Benchmark (OSD)

This benchmark evaluates reconstruction completeness and accuracy across multi-sensor SLAM, multi-view stereo, and radiance field methods.

1) Benchmarked Methods: We compare three representative approaches: VILENS-SLAM [7], OpenMVS [15], and Nerfacto [16]. All produce 3D point clouds. For Nerfacto, the cloud is extracted by estimating expected depth and color along training rays from the radiance field and projecting these into 3D space.

2) Evaluation Metrics: We adopt the *F-score*, the harmonic mean of precision and recall, to jointly measure reconstruction accuracy and completeness. A reconstructed point is treated as a true positive (TP) if it lies within 5 cm or 10 cm of a ground-truth point, and as a false positive (FP) otherwise. False negatives (FN) arise when ground-truth points are missing in the reconstruction, while true negatives (TN) correspond to valid gaps. From these definitions we compute precision, recall, and F-score, and additionally report point-to-point distances as a direct measure of geometric accuracy. Results are presented in Table II.

C. Novel View Synthesis Benchmark (OSD & OXDAN)

This benchmark evaluates the ability of neural rendering methods to generate photorealistic novel views under controlled conditions (OSD) and in-the-wild illumination changes (OXDAN).

1) Benchmarked Methods: On OSD, we test Nerfacto [16] and Splatfacto [30], an implementation of 3D Gaussian Splatting [2] with performance comparable to the original. To assess model scalability, we also include larger-capacity variants: Nerfacto-big (with expanded hash grids, proposal networks, and ray samples) and Splatfacto-big (with reduced densification thresholds, producing more Gaussians). On OXDAN, we evaluate in-the-wild NVS systems Splatfacto-W [28] and Gaussian-Wild [29].

TABLE III: Quantitative evaluation of Novel View Synthesis (OSD). Results for Nerfacto-big and Splatfacto-big are reported in the original paper [3]. Best results are highlighted in blue with different tints. Test images are drawn from the training trajectory (In-Sequence) as well as from a separate trajectory with viewpoints far from training poses (Out-of-Sequence).

Sequence	Method		In-Sequence			Out-of-Sequence	
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Blenheim Palace	Nerfacto	18.42	0.716	0.506	17.09	0.682	0.537
	Splatfacto	19.34	0.726	0.589	16.02	0.668	0.659
Keble College	Nerfacto	21.10	0.731	0.397	20.29	0.748	0.368
	Splatfacto	20.47	0.651	0.514	19.92	0.658	0.500
Observatory Quarter	Nerfacto	23.40	0.807	0.336	21.25	0.786	0.370
	Splatfacto	22.76	0.791	0.373	19.47	0.736	0.445

TABLE IV: Visual Relocalization Results on *Day* and *Night* Queries (OXDAN). We report the percentage of query images correctly localized within three thresholds: (0.25m, 2°), (0.5m, 5°) and (1m, 10°). Results are shown for both feature-matching (FM) and scene coordinate regression (SCR) approaches. For FM approaches, the top 50 images retrieved using NetVLAD [17] are used for matching.

		***Day* Queries**		***Night* Queries**	
		Keble College	**Observatory Quarter**	**Keble College**	**Observatory Quarter**
FM	SIFT [18]	84.98 / 88.78 / 91.06	89.86 / 92.69 / 92.92	0.40 / 0.79 / 1.39	2.38 / 3.35 / 4.55
	SP+SG [19]	94.68 / 97.34 / 98.10	94.81 / 95.99 / 95.99	10.66 / 13.57 / 17.27	48.14 / 54.40 / 58.05
	SP+LG [20]	92.78 / 96.20 / 97.15	94.34 / 95.75 / 95.99	9.99 / 14.16 / 18.27	47.91 / 53.50 / 57.68
	DISK+LG [21]	85.74 / 89.54 / 91.25	92.45 / 95.05 / 95.28	0.53 / 0.79 / 1.06	16.77 / 20.42 / 22.95
	LoFTR [22]	94.30 / 96.96 / 97.91	94.81 / 95.28 / 95.99	10.39 / 13.63 / 17.01	50.00 / 57.00 / 60.13
	RoMA [23]	91.83 / 96.20 / 97.15	91.27 / 93.87 / 93.87	14.96 / 22.63 / 30.91	58.94 / 66.92 / 70.79
	MASt3R [24]	94.68 / 97.91 / 98.86	89.39 / 92.92 / 94.58	12.24 / 16.08 / 19.66	48.66 / 54.47 / 59.91
SCR	ACE [25]	0.57 / 3.80 / 22.24	0.24 / 8.02 / 25.24	0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00
	GLACE [26]	0.19 / 4.18 / 35.93	0.24 / 6.13 / 33.02	0.00 / 0.00 / 0.99	0.00 / 0.00 / 0.00
	R-SCoRe [27]	60.46 / 75.10 / 85.74	45.52 / 58.02 / 71.23	0.20 / 0.99 / 1.92	3.06 / 7.75 / 13.34

TABLE V: 3DGS In-the-Wild Results (OSD). We report image rendering and geometry quality using the following metrics: PSNR (↑) / LPIPS (↓) / point-to-point distance (↓). The 3DGS geometry is derived by extracting the centers of all Gaussian primitives, with point-to-point distance (meter) computed against the ground truth laser-scanned point cloud. Symbol ”-” denotes the system produces a degenerated point cloud (less than 2000 gaussians after training).

Method	**Bodleian Library**	**H.B. Allen Centre**	**Keble College**	**Observatory Quarter**	**Robotics Institute**
Splatfacto-W [28]	25.98 / 0.60 / -	25.65 / 0.59 / 0.75	27.96 / 0.59 / -	25.83 / 0.63 / 0.36	22.73 / 0.61 / 0.42
Gaussian-Wild [29]	28.38 / 0.56 / 1.44	24.94 / 0.59 / 1.48	30.92 / 0.56 / 0.69	28.57 / 0.60 / 0.69	25.05 / 0.57 / 0.76

2) Evaluation Metrics: For OSD, we report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [31], and Learned Perceptual Image Patch Similarity (LPIPS) [32]. For OXDAN, we additionally evaluate geometry by computing point-to-point distances against TLS scans.

Results in Table V show Splatfacto-W outperforming Gaussian-Wild on the H.B. Allen Centre scene but underperforming on the other four. However, both methods struggle overall, as indicated by LPIPS scores, due to the dataset's large scale and extreme lighting variations from daylight to poorly illuminated night conditions.

*D. Visual Relocalization Benchmark (*OXDAN*)*

This benchmark evaluates the ability of relocalization methods to recover accurate poses from egocentric images, particularly under challenging day-to-night variations.

1) Benchmarked Methods: We benchmark both feature matching (FM) and scene coordinate regression (SCR) approaches on OXDAN. For FM, we use the HLoc pipeline, retrieving top images with NetVLAD and estimating poses via PnP-RANSAC. We evaluate four sparse methods: SIFT, SuperPoint (SP) with SuperGlue (SG) or LightGlue (LG), and DISK with LightGlue, as well as three dense methods: LoFTR, RoMA, and MASt3R. For SCR, we test ACE, GLACE, and R-SCoRe, which regress dense 2D–3D correspondences.

2) Evaluation Metrics: Performance is measured on both daytime and nighttime queries, reporting the percentage of images localized within thresholds of (0.25 m, 2°), (0.5 m, 5°), and (1 m, 10°). Overall, FM approaches outperform SCR, which frequently fails under low-light conditions. FM methods perform strongly in daytime but degrade at night, while RoMA achieves the best overall robustness across lighting conditions.

IV. Conclusion

We have presented two complementary datasets—Oxford Spires (OSD) and Oxford Day-and-Night (OXDAN)—that together enable comprehensive benchmarking of localization, 3D reconstruction, visual relocalization, and neural radiance field methods under both controlled and in-the-wild illumination. By pairing sensor-rich ground truth with egocentric recordings across day-to-night transitions, these datasets provide a unique platform for developing and evaluating perception algorithms that must remain robust in realistic navigation scenarios. We hope they will serve as a valuable resource for advancing open-world robotic perception and long-term autonomy.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, July 2023.

[3] Y. Tao, M. Á. Muñoz-Bañón, L. Zhang, J. Wang, L. F. T. Fu, and M. Fallon, "The oxford spires dataset: Benchmarking large-scale lidar-visual localisation, reconstruction and radiance field methods," *Intl. J. of Robot. Res.*, 2025.

[4] Z. Wang, W. Bian, X. Li, Y. Tao, J. Wang, M. Fallon, and V. A. Prisacariu, "Seeing in the dark: Benchmarking egocentric 3d vision with the oxford day-and-night dataset," *arXiv preprint arXiv:2506.04224*, 2025.

[5] 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 *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 4353–4360.

[6] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[7] M. Ramezani, G. Tinchev, E. Iuganov, and M. Fallon, "Online lidar-slam for legged robots with robust registration and deep-learned loop closure," in *IEEE Int. Conf. Robot. Autom.* IEEE, 2020, pp. 4158–4164.

[8] W. Xu and F. Zhang, "Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3317–3324, 2021.

[9] G. Kim, S. Yun, J. Kim, and A. Kim, "Sc-lidar-slam: A front-end agnostic versatile lidar slam system," in *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, 2022, pp. 1–6.

[10] J. Lin, C. Yuan, Y. Cai, H. Li, Y. Ren, Y. Zou, X. Hong, and F. Zhang, "Immesh: An immediate lidar localization and meshing framework," *IEEE Trans. Robotics*, vol. 39, no. 6, pp. 4312–4331, 2023.

[11] X. Liu, Z. Liu, F. Kong, and F. Zhang, "Large-scale lidar consistent mapping using hierarchical lidar bundle adjustment," *IEEE Robot. Autom. Lett.*, vol. 8, no. 3, pp. 1523–1530, 2023.

[12] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[14] M. Grupp, "evo: Python package for the evaluation of odometry and slam." 2017.

[15] D. Cernea, "OpenMVS: Multi-view stereo reconstruction library," 2020.

[16] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. Mcallister, J. Kerr, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *SIGGRAPH*, 2023, pp. 1–12.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[19] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *CVPR*, 2020.

[20] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *CVPR*, 2023.

[21] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," in *NeurIPS*, 2020.

[22] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021.

[23] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *CVPR*, 2024.

[24] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *ECCV*, 2024.

[25] E. Brachmann, T. Cavallari, and V. A. Prisacariu, "Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses," in *CVPR*, 2023.

[26] F. Wang, X. Jiang, S. Galliani, C. Vogel, and M. Pollefeys, "Glace: Global local accelerated coordinate encoding," in *CVPR*, 2024.

[27] X. Jiang, F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "R-score: Revisiting scene coordinate regression for robust large-scale visual localization," in *CVPR*, 2025.

[28] C. Xu, J. Kerr, and A. Kanazawa, "Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections," *arXiv preprint arXiv:2407.12306*, 2024.

[29] D. Zhang, C. Wang, W. Wang, P. Li, M. Qin, and H. Wang, "Gaussian in the wild: 3d gaussian splatting for unconstrained image collections," in *ECCV*, 2024.

[30] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, and A. Kanazawa, "gsplat: An open-source library for Gaussian splatting," *Journal of Machine Learning Research*, vol. 26, no. 34, pp. 1–17, 2025.

[31] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.