Supplementary Material

In this appendix, we first present additional implementation details and experimental settings in Sec. A and Sec. B, which were omitted from the main paper due to page limit. We then report additional analyses in Sec. C. Finally, we show more reconstruction results of our method in Sec. D.

A. Implementation details

Retrieval module. We propose a lightweight module for efficient scene frame retrieval to support the keyframe registration. The retrieval module directly reuses I2P's decoder blocks as its backbone, followed by a linear projection and an average-pooling layer. Specifically, it uses the first two blocks from both the supporting and keyframe decoders, for scene frames and keyframes (awaiting registration), respectively. It takes as input image features of one keyframe and all the scene frames in the buffering set, predicting correlation scores between the keyframe and each buffering frame. Notably, the correlation scores share similar behavior with the mean confidence of the I2P model's final prediction and offer unique advantages over the cosine similarity between image features of two frames. These correlation scores account for both visual similarity and provide suitable baselines for 3D reconstruction.

The module inherits the weights of the first two layers of the decoder in I2P model. During training, only the weights of the linear projection are updated using an L1 loss:

$$\mathcal{L}_{Retr} = \sum_{i=1}^{R} |S_i' - Mean(C_i')|,$$
$$S_i' = Sigmoid(S_i),$$
$$C_i' = (C_i - 1)/C_i,$$

where R is the number of input supporting frames, S_i is the predicted correlation score between supporting frame i and the keyframe, C_i is the predicted confidence from the complete I2P model. Both S_i and C_i are normalized to [0,1] before calculating the loss.

Multi-keyframe co-registration. In practice, our scene decoder in the L2W model adopts the same architecture as the keyframe decoder in the I2P model, allowing for the simultaneous input and registration of multiple keyframes. In the decoding stage, scene frames and keyframes exchange information bidirectionally: each scene frame queries features from all keyframes, and each keyframe interacts with all scene frames. Compared to single-keyframe registration, this extension significantly reduces computational overhead by registering multiple keyframes with a single pass of the scene decoder. Furthermore, incorporating information from additional keyframes enhances the refinement of scene frame features, leading to more accurate reconstruction for all input frames.

Training details. To construct the training data, we utilize all iPhone and DSLR frames registered by COLMAP [47] from the training splits of ScanNet++[71]. Additionally, we include all frames from the first 450 scenes of the Aria Synthetic Environments (ASE)[3] dataset and 41 categories from CO3D-v2 [41], with each category containing up to 50 randomly sampled scene sequences. We introduce two ways to extract video clips for training. For ScanNet++ and ASE, we adopt uniform sampling with strides of 3 and 2, respectively. For CO3D-v2, frames are randomly sampled within temporal segments covering half the length of each video. In total, we extract approximately 850K clips. During each epoch of training, we randomly sample 4000, 2000, and 2000 clips from the ScanNet++, ASE, and CO3D-v2 datasets, respectively. All training images are resized and then center-cropped to 224×224 pixels. Standard data augmentation techniques [64] are applied.

To train our I2P model, we extend the training process of DUSt3R from two views to multiple views. Specifically, our I2P model takes as input a video clip of length 11, and designates the middle frame as the keyframe. We train the I2P model for 100 epochs, which takes about 6 hours. After that, we train the retrieval module built on the I2P model. During training, we freeze all other modules and use L1 loss to supervise the correlation score against the mean confidence of the I2P model's final predictions. This module requires 50 epochs of training, which takes about 2 hours.

To train the L2W model, we use clips of length 12, with the first six images selected as scene frames, and the last six images designated as keyframes to register. The model is trained for 200 epochs in total, and the training process takes approximately 16 hours. When training with ground truth pointmaps as input, we set invalid points to (0,0,0). A confidence-aware loss without scale normalization is applied, ensuring that the predicted point maps retain consistent scale with the input scene frames.

Our training is conducted on 8 NVIDIA 4090D GPUs, each with 24GB of memory and a batch size of 4 per GPU.

B. Details for experimental settings

Calculation of the evaluation metrics. To evaluate reconstruction quality, we use accuracy and completeness as our metrics. They are calculated by:

$$Accuracy = \frac{1}{P} \sum_{i=1}^{P} min_j(D(x_i, y_j)),$$

$$Completeness = \frac{1}{Q} \sum_{j=1}^{Q} min_i(D(x_i, y_j)).$$

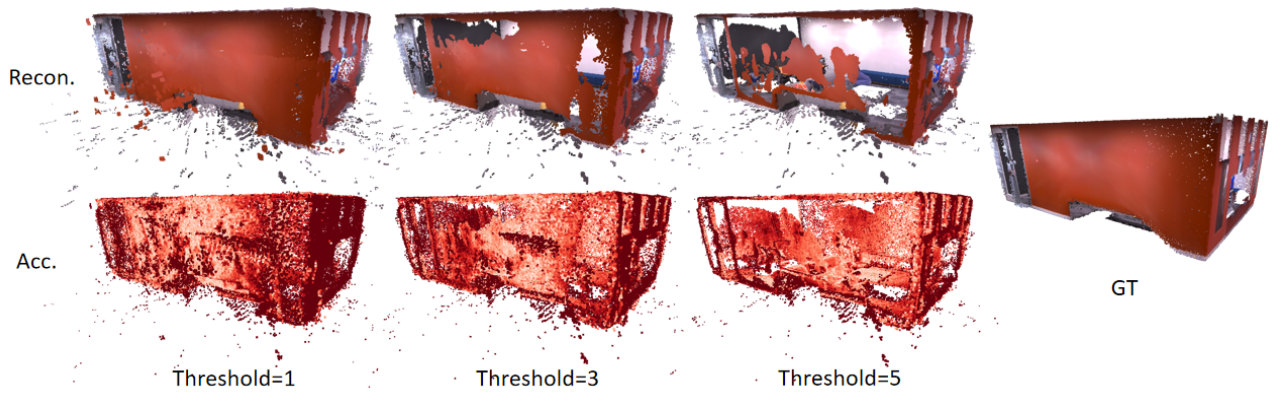

Figure 6. The reconstruction results and the corresponding accuracy heatmaps of MASt3R [28] on Office 3 from Replica [54] dataset under different confidence thresholds. Lighter colors indicate higher accuracy.

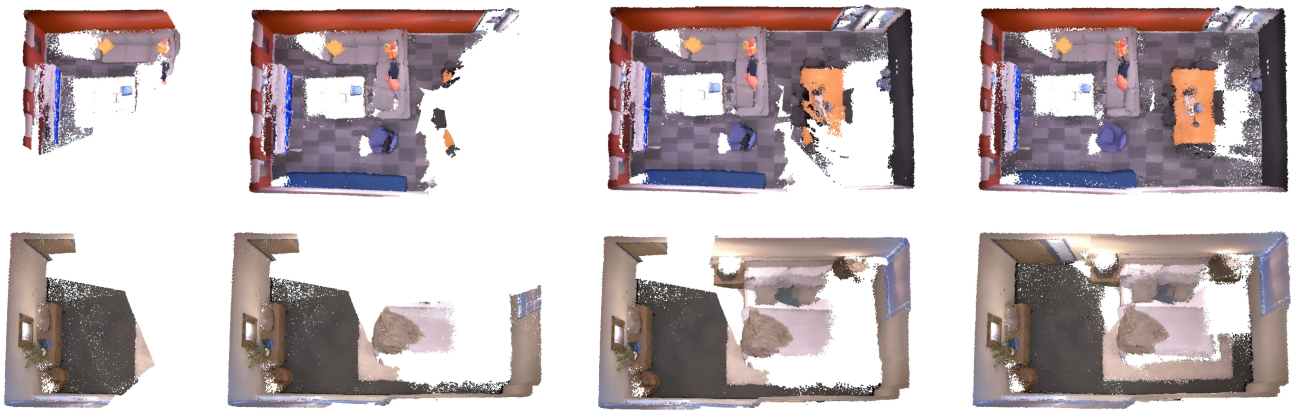

Figure 7. Visualization of the incremental reconstruction process of our method on the Office 3 and Room 1 of Replica [54] dataset. Our method achieves low drift without any global-optimization stage.

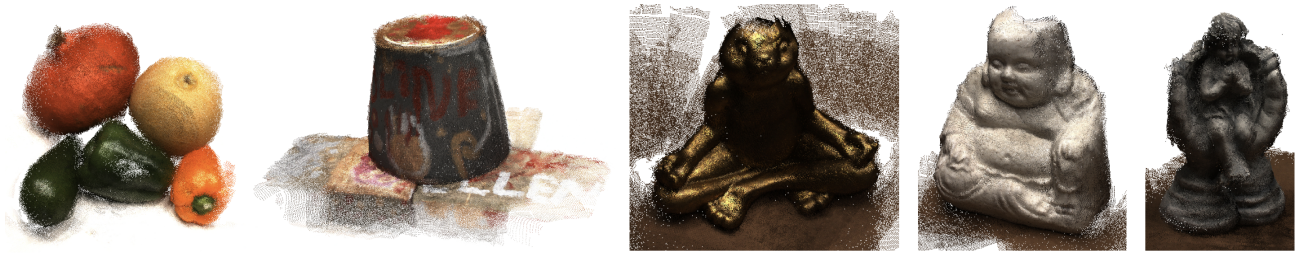

Figure 8. Reconstruction results on unorganized image collections from DTU [1] dataset.

P and Q are the numbers of points in the reconstructed point cloud and GT point cloud respectively. $D(\cdot)$ represents Euclidean distance, and x_i and y_j represent iterating each point from the reconstructed and GT point cloud.

To measure the efficiency, we report FPS (frames per second), which is calculated by:

$$FPS = F/time,$$

where $time$ is the total time used to reconstruct the scene, and F is the number of frames from the video.

We evaluate the camera pose accuracy using absolute trajectory error (ATE-RMSE), which is formulated by:

$$ATE\text{-}RMSE = \sqrt{\frac{1}{F}\sum_{i=1}^{F} D(T_i^{gt}, T_i^{perd})^2},$$

Method	Chess	Fire	Heads	Office	Pumpkin	RedKitchen	Stairs	Average	FPS
DUSt3R [64] (w/PnP)	5.09	4.88	2.52	12.07	**10.64**	10.35	10.55	8.02	<1
MASt3R [28] (w/PnP)	4.32	**2.92**	**1.47**	12.37	11.82	7.98	**3.04**	**6.28**	≪1
NICER-SLAM [79]*	**3.28**	6.85	4.16	**10.84**	20.00	**3.94**	10.81	8.55	<1
DROID-SLAM [56]*	**3.36**	**2.40**	**1.43**	**9.19**	16.46	**4.94**	**1.85**	**5.66**	∼20
Spann3R [61]	9.18	6.69	7.10	21.56	12.83	14.06	10.43	11.70	>50
SLAM3R-NoConf (Ours)	6.29	5.33	4.47	12.42	11.74	9.53	9.30	8.44	∼25
SLAM3R (Ours)	6.20	5.30	4.56	12.40	**11.71**	9.47	9.20	8.41	∼25

Table 6. Camera pose estimation results on the 7Scenes [51] dataset reported using the ATE-RMSE (cm) metric. The average numbers are computed over all test scenes. * denotes the results reported in NICER-SLAM.

Method	Room 0	Room 1	Room 2	Office 0	Office 1	Office 2	Office 3	Office 4	Average	FPS
DUSt3R [64] (w/PnP)	4.00	4.49	7.62	4.88	4.04	3.90	2.84	6.30	4.76	<1
MASt3R [28] (w/PnP)	**1.07**	**0.99**	**0.87**	**0.90**	4.90	**1.21**	1.77	**1.63**	**1.67**	≪1
NICER-SLAM [79]*	1.36	1.60	1.14	2.12	**3.23**	2.12	**1.42**	2.01	1.88	<1
GO-SLAM [75]	-	-	-	-	-	-	-	-	0.39	∼8
DIM-SLAM [29]	0.48	0.78	0.35	0.67	**0.37**	0.36	**0.33**	**0.36**	0.46	∼3
DROID-SLAM [56]*	**0.34**	**0.13**	**0.27**	**0.25**	0.42	**0.32**	0.52	0.40	**0.33**	∼20
Spann3R [61]	29.76	34.78	26.08	34.50	22.65	34.47	42.24	37.84	32.79	>50
SLAM3R-NoConf (Ours)	4.54	5.89	5.73	11.17	6.32	6.15	4.99	8.05	6.61	∼24
SLAM3R (Ours)	4.56	5.88	5.72	11.17	6.32	6.15	4.95	8.09	6.61	∼24

Table 7. Camera pose estimation results on the Replica [54] dataset reported using the ATE-RMSE (cm) metric.

where T^{perd} and T^{gt} are the camera center positions of the predicted and GT camera trajectories.

Full video as input on Replica [54]. On the Replica dataset, we reconstruct the entire scene geometry using all video frames. With the stride of the sliding window set to 1, all frames will be used as a keyframe once. For each window, frames are sampled around the keyframe, with $Skip = 20$ frames per supporting frame, to ensure reasonable camera motion (disparity). We co-register $Co = 10$ keyframes at each time, which share the same $K = 10$ scene frames as a reference. These scene frames are selected through a two-step process. First, we calculate the correlation score between all frames in the buffering set and the Co keyframes. Then, we select K frames from the buffering set that show the highest total correlation score with these keyframes. After every $R = 20$ registered keyframes, we update the buffering set by retaining the keyframes with the highest reconstruction scores, where reconstruction score of a frame is the product of its mean confidence predicted by I2P and L2W model. The insertion/update follows the reservoir sampling probability described in the main paper.

Sampled frames as input on 7 Scenes [51]. Following Spann3R [61], the frames in each test sequence are sampled with a stride of 20, and we only reconstruct the points from the sampled frames. To handle sampled-frame-only input, we adapt our reconstruction pipeline for full-video input by setting $Skip = 1$, $Co = 2$, $K = 5$, and $R = 1$ in practice.

Experiments on DUSt3R [64] and MASt3R [28]. The global optimization with complete graph setting in DUSt3R and MASt3R requires substantial GPU memory. Consequently, to evaluate the global reconstruction quality of these two methods on the Replica dataset, we uniformly sample 1/20 of the images. DUSt3R is tested using the weight-224 model with a resolution of 224×224, the same as our input resolution, while MASt3R is tested using the weight-512 model with resolutions of 512×384 and 512×288 as inputs for reconstructing the 7 Scenes [51] and Replica [54] datasets, respectively. Note that a resolution of 224×224 results in less overlap between adjacent frames, making reconstruction inherently more challenging.

During the evaluation, we observed that MASt3R occasionally generates floating points with high confidence scores, which are difficult to filter using confidence thresholds and significantly degrade accuracy. An example of this issue is shown in Figure 6. In contrast, our confidence scores are more effective and successfully reduce erroneous points. The results of SLAM3R reported on 7 Scenes and Replica datasets use a fixed confidence threshold of 3.

C. Additional comparisons and analyses

More numerical results. We report more quantitative comparisons of reconstruction results on ScanNet [11],

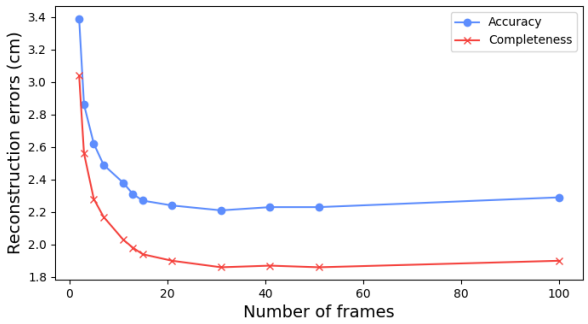

Figure 9. Inner-window keyframe reconstruction results from various window lengths.

Tanks and Temples [26], and ETH3D [50] datasets. We sampled three scenes from each dataset, and report the results in Table 8. SLAM3R outperforms Spann3R in most cases and demonstrates performance either comparable to or better than DUSt3R. These results further verify our method's effectiveness.

ScanNet	scene0011_00	scene0015_00	scene0019_00	Average
DUSt3R [64]	**5.56 / 3.76**	**5.04 / 4.10**	4.52 / 4.74	**5.04 / 4.20**
Spann3R [61]	13.09 / 11.37	8.51 / 7.79	7.97 / 9.66	9.86 / 9.61
SLAM3R (Ours)	5.86 / 3.98	5.98 / 5.97	**4.27 / 4.34**	5.37 / 4.76
Tanks and Temples	Ignitius	Truck	Caterpillar	Average
DUSt3R [64]	3.55/ 1.22	9.31 / **4.85**	12.67 / 5.25	8.51 / **3.77**
Spann3R [61]	5.51 / 1.10	6.40 / 12.61	**11.50**/ 5.74	7.80 / 6.48
SLAM3R (Ours)	**3.30 / 0.94**	**5.35**/ 5.59	12.26 / **5.05**	**6.97** / 3.86
ETH3D	plant_scene_1	table_3	sofa_1	Average
DUSt3R [64]	2.98 / 2.48	3.13 / **1.30**	**2.05** / 3.67	2.72 / 2.48
Spann3R [61]	2.54 / 4.25	3.03 / 2.08	2.10 / 4.55	2.56 / 3.62
SLAM3R (Ours)	**2.36 / 1.98**	**2.75** / 1.34	2.13 / **1.90**	**2.41 / 1.74**

Table 8. Reconstruction errors (accuracy / completeness) on ScanNet [11], Tanks and Temples [26], and ETH3D [50] datasets.

Diminishing return of window length. In the main paper, we report the I2P reconstruction results with different window lengths. Here, we further analyze the diminishing returns, which indicate that the window length should not be too large. As Figure 9 shows, the accuracy and completeness of the keyframe reconstruction improve rapidly at first as input frames increase, but then gradually decline. This is because larger windows result in less and less overlapping. Additionally, the inference time becomes significantly slower as length increases. Consequently, we set the window size to 11 in our main experiments, balancing the reconstruction quality and runtime efficiency.

Effect of scene frame numbers on registration. We conduct experiments on the Replica [54] dataset to investigate how the number of scene frames selected as a global reference affects the registration quality of keyframes. As reported in Table 9, the accuracy of full-scene registration initially improves as the maximum number of input scene frames increases but eventually declines beyond a certain threshold. Retrieving too few scene frames from the buffering set risks missing suitable frames and causing keyframe registration to get stuck in local minimums. Conversely, selecting too many scene frames can introduce irrelevant ones that add noise and hinder registration.

# Scene frames	Acc.	Comp.	FPS
1	4.18	2.61	~398
5	3.99	2.79	~247
10	**3.57**	2.62	~152
20	**3.57**	2.60	~86
30	3.59	**2.58**	~61
40	4.15	3.05	~46
50	4.27	3.15	~37

Table 9. Reconstruction results on Replica [54] dataset, with various maximum number of scene frames selected for keyframe registration. The FPS of the L2W model aligning 10 keyframes at once with different numbers of input scene frames is also reported.

To balance reconstruction accuracy and runtime efficiency, we set the number of retrieved scene frames to 5 and 10 on 7 Scenes [51] and Replica [54] dataset, which achieves consistent and reliable performance.

Camera pose estimation. The detailed results are presented in Table 6 and Table 7. For DUSt3R [64] and MASt3R [28], we evaluate the camera poses derived via the PnP-RANSAC solver with their predicted pointmaps (after global alignment) and GT intrinsic parameters. When evaluating Spann3R [61] on the Replica [54] dataset, only one-twentieth of the frames are used, as it fails to give reasonable results with all frames input.

We outperform the concurrent work Spann3R [61], demonstrating the effectiveness of our hierarchical design with multi-view input and global retrieval. Among classical SLAM systems, the pose errors of GO-SLAM [75] and DROID-SLAM [56] are lower than those of NICER-SLAM. However, their reconstruction accuracy and completeness are worse. This discrepancy between pose and reconstruction errors indicates that effective end-to-end 3D reconstruction is possible and promising without first obtaining precise camera poses.

D. More visual results

Visualization of incremental reconstruction. Figure 7 visualizes the process of our incremental reconstruction on two scenes from Replica [54]. Our method achieves effective alignment at loops while experiencing minimal cumulative drift, without offline global optimization step.

Reconstruction on DTU [1] dataset. The results are shown in Figure 8. Note that our method does not require any camera parameters, and produces dense point cloud reconstructions end-to-end in real-time.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 2, 11, 14

[2] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022. 7, 8

[3] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. *arXiv preprint arXiv:2403.13064*, 2024. 6, 10

[4] Tim Bailey and Hugh Durrant-Whyte. Simultaneous localization and mapping (slam): Part ii. *IEEE robotics & automation magazine*, 13(3):108–117, 2006. 2

[5] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam—learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018. 1, 2

[6] G Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. 8

[7] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 1, 2

[8] Yiyang Chen, Siyan Dong, Xulong Wang, Lulu Cai, Youyi Zheng, and Yanchao Yang. Sg-nerf: Neural surface reconstruction with scene graph optimization. *arXiv preprint arXiv:2407.12667*, 2024. 2

[9] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang-Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H Hsu. Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9400–9406. IEEE, 2023. 1, 2

[10] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. Deepfactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 5 (2):721–728, 2020. 1

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 12, 13

[12] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. 3

[15] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006. 2

[16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 1, 2

[17] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 1, 2

[18] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 2

[19] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 2

[20] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[21] Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photo-realistic mapping for monocular stereo and rgb-d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21584–21593, 2024. 2

[22] Jiahui Huang, Shi-Sheng Huang, Haoxuan Song, and Shi-Min Hu. Di-fusion: Online implicit 3d reconstruction with deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8932–8941, 2021. 1, 2

[23] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 2

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[25] Georg Klein and David Murray. Parallel tracking and mapping on a camera phone. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 83–86. IEEE, 2009. 1, 2

[26] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 7, 8, 13

[27] Lukas Koestler, Nan Yang, Niclas Zeller, and Daniel Cremers. Tandem: Tracking and dense mapping in real-time using deep multi-view stereo. In *Conference on Robot Learning*, pages 34–45. PMLR, 2022. 2

[28] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 3, 6, 7, 11, 12, 13

[29] Heng Li, Xiaodong Gu, Weihao Yuan, Luwei Yang, Zilong Dong, and Ping Tan. Dense rgb slam with neural implicit maps. In *Proceedings of the International Conference on Learning Representations*, 2023. 1, 2, 6, 7, 12

[30] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 2

[31] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. 1, 2

[32] Lorenzo Liso, Erik Sandström, Vladimir Yugay, Luc Van Gool, and Martin R Oswald. Loopy-slam: Dense neural slam with loop closures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20363–20373, 2024. 2

[33] Shaohui Liu, Yifan Yu, Rémi Pautrat, Marc Pollefeys, and Viktor Larsson. 3d line mapping revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21445–21455, 2023. 1, 2

[34] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 2

[35] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 7, 8

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[37] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 1, 2

[38] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1, 2

[39] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011. 1, 2

[40] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 1

[41] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 6, 10

[42] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023. 1, 2

[43] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001. 7, 8

[44] Erik Sandström, Yue Li, Luc Van Gool, and Martin R Oswald. Point-slam: Dense neural point cloud-based slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18433–18444, 2023. 2

[45] Erik Sandström, Kevin Ta, Luc Van Gool, and Martin R Oswald. Uncle-slam: Uncertainty learning for dense neural slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4537–4548, 2023. 2

[46] Erik Sandström, Keisuke Tateno, Michael Oechsle, Michael Niemeyer, Luc Van Gool, Martin R Oswald, and Federico Tombari. Splat-slam: Globally optimized rgb-only slam with 3d gaussians. *arXiv preprint arXiv:2405.16544*, 2024. 1, 2

[47] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 10

[48] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 2

[49] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 7, 8

[50] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 7, 8, 13

[51] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d

images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 6, 7, 8, 12, 13

[52] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibarated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2

[53] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 1, 2

[54] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6, 7, 8, 11, 12, 13

[55] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6229–6238, 2021. 1, 2

[56] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 1, 2, 6, 7, 12, 13

[57] Fabio Tosi, Youmin Zhang, Ziren Gong, Erik Sandström, Stefano Mattoccia, Martin R Oswald, and Matteo Poggi. How nerfs and 3d gaussian splatting are reshaping slam: a survey. *arXiv preprint arXiv:2402.13255*, 4, 2024. 1

[58] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 7, 8

[59] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985. 5

[60] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8606–8615, 2022. 1, 2

[61] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 3, 6, 7, 12, 13

[62] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2

[63] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 2

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1, 2, 3, 4, 6, 7, 8, 10, 12, 13

[65] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 2

[66] Changchang Wu. Visualsfm: A visual structure from motion system. *http://www. cs. washington. edu/homes/ccwu/vsfm*, 2011. 1, 2

[67] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 2

[68] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 2

[69] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 7, 8

[70] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 2

[71] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 6, 10

[72] Vladimir Yugay, Yue Li, Theo Gevers, and Martin R Oswald. Gaussian-slam: Photo-realistic dense slam with gaussian splatting. *arXiv preprint arXiv:2312.10070*, 2023. 2

[73] Ganlin Zhang, Erik Sandström, Youmin Zhang, Manthan Patel, Luc Van Gool, and Martin R Oswald. Glorie-slam: Globally optimized rgb-only implicit encoding point cloud slam. *arXiv preprint arXiv:2403.19549*, 2024. 1, 2

[74] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 2

[75] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3727–3737, 2023. 1, 2, 6, 7, 12, 13

[76] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 822–838, 2018. 1, 2

[77] Heng Zhou, Zhetao Guo, Shuhong Liu, Lechen Zhang, Qihao Wang, Yuxiang Ren, and Mingrui Li. Mod-slam:

Monocular dense mapping for unbounded 3d scene reconstruction. *arXiv preprint arXiv:2402.03762*, 2024. 1, 2

[78] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12786–12796, 2022. 1, 2

[79] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. In *2024 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2024. 1, 2, 6, 7, 8, 12