3D Reconstruction with Spatial Memory
Supplementary Material

6. Additional details

Training loss. The confidence loss as in DUSt3R [81] is:

$$\mathcal{L}_{\text{conf}} = \sum_t \sum_{i \in \mathcal{V}} C_t^i \mathcal{L}_{\text{reg}}(i) - \alpha \log C_t^i, \tag{10}$$

where $\mathcal{V}$ is the set of all valid pixels. The confidence C_t^i is an exponential function of the raw output of the network $\hat{C}_t^i$:

$$C_t^i = 1 + \exp(\hat{C}_t^i) \tag{11}$$

In confidence loss $\mathcal{L}_{\text{conf}}$, α controls the total confidence score the model needs to distribute to the loss of each pixel. Since the regions with larger depths usually have a larger loss, the model assigns more confidence weight to regions with smaller depths. This loss shares a similar spirit to other depth representations, e.g., inverse depth, which gives more weight to pixels with smaller depth. However, instead of explicitly encoding the depth, this confidence loss let the model learn the weight function along the training. Our scale loss is defined as:

$$\mathcal{L}_{\text{scale}} = \max(0, \bar{X} - \bar{X}_{\text{gt}}), \tag{12}$$

where $\bar{X}$ and $\bar{X}_{\text{gt}}$ are the average distance of all predicted and ground-truth points to the origin. The scale loss encourages the predicted scale to be smaller than the GT scale to prevent the model from learning trivial solutions.

To tune the hyper-parameter α, we find that the best way is to ensure the overall training loss becomes smaller than 0 after 30% of epochs. A typical sign of choosing the α that is not big enough is the $\mathcal{L}_{\text{scale}}$ becomes quite large along the training. This indicates that some pixels with large depths make the model predict the trivial solutions. We find that $\alpha \geq 0.4$ achieves the best results in our case.

Curriculum training. Given the minimal and maximum sampling interval $T_{\min}$ and $T_{\max}$ between adjacent frames, our curriculum sampling can be written as:

$$T = T_{\min} + \eta_a (T_{\max} - T_{\min}), \tag{13}$$

where η_a is the active ratio of the training ratio η:

$$\eta_a = \begin{cases} \min(1, 2\eta) & \text{if } \eta < 0.75 \\ \max(0.5, 4 - 4\eta) & \text{otherwise} \end{cases} \tag{14}$$

7. Additional analysis

Ablation study on view selection. Since the confidence function in Eq. 11 tends to over-weight patches with higher confidence, we instead use the sigmoid function for view selection of the offline reconstruction. The overall confidence function becomes:

$$C = \frac{C_1 - 1}{C_1} + \frac{C_2 - 1}{C_2}. \tag{15}$$

The difference in performance is illustrated in Tab. 4.

Method	Acc↓		Comp↓		NC↑	
	Mean	Med.	Mean	Med.	Mean	Med.
Ours* (exp)	3.099	1.361	2.247	0.993	0.731	0.835
Ours*	**2.902**	**1.273**	**2.120**	**0.937**	**0.732**	**0.836**

Table 4. **Ablation study on view selection.** Ours* (exp): exponential confidence function for view selection as in DUSt3R [81]. Ours*:sigmoid confidence function for view selection.

datasets	Method	Acc↓		Comp↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.
7scenes	**Dust3R†**	0.0286	0.0123	0.0280	0.0091	0.6681	0.7683
	Dust3R$^{\text{ours}}$	**0.0278**	**0.0117**	0.0247	0.0101	**0.6775**	**0.7842**
	Ours	0.0342	0.0148	**0.0241**	**0.0085**	0.6635	0.7625
7scenes (FV)	**Dust3R†**	0.0279	0.0133	0.0276	0.0108	0.7630	0.8841
	Dust3R$^{\text{ours}}$	0.0242	0.0114	0.0249	0.0106	**0.7785**	**0.9003**
	Ours	**0.0239**	**0.0111**	**0.0247**	**0.0103**	0.7768	0.8985
NRGBD	**Dust3R†**	**0.0544**	0.0251	0.0315	**0.0103**	0.8024	0.9529
	Dust3R$^{\text{ours}}$	0.0644	**0.0246**	0.0396	0.0110	**0.8041**	**0.9623**
	Ours	0.0691	0.0315	**0.0291**	0.0110	0.7775	0.9371
NRGBD (FV)	**Dust3R†**	**0.0591**	0.0266	0.0409	0.0136	0.8305	0.9556
	Dust3R$^{\text{ours}}$	0.0606	**0.0252**	0.0407	0.0143	**0.8439**	**0.9630**
	Ours	0.0611	0.0254	**0.0392**	**0.0135**	0.8330	0.9593

Table 5. **Ablation study on Dust3R$^{\text{ours}}$ on indoor scene.**

datasets	Method	Acc↓		Comp↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.
DTU	**Dust3R†**	**2.296**	**1.297**	2.158	**1.002**	**0.747**	**0.848**
	Dust3R$^{\text{ours}}$	3.386	1.469	2.228	1.017	0.734	0.837
	Ours*	2.902	1.273	**2.120**	0.937	0.732	0.836
DTU (FV)	**Dust3R†**	**2.511**	**1.484**	**2.661**	**1.230**	**0.788**	**0.883**
	Dust3R$^{\text{ours}}$	3.875	1.869	2.916	1.438	0.777	0.874
	Ours* (FV)	3.055	1.600	2.878	1.345	0.781	0.878

Table 6. **Ablation study on Dust3R$^{\text{ours}}$ on DTU.**

Ablation study on DUSt3R in Spann3R. Since our model inherits from the network architecture of DUSt3R [81], we can directly compare the performance of the ViT encoder with two decoders in our model, denoted as DUSt3R$^{\text{ours}}$, with the original DUSt3R. As shown in Tab. 5 and Tab. 6, even though we re-purpose the two decoders, DUSt3R$^{\text{ours}}$ still shows on-par median accuracy and completion and consistent better normal consistency compared to DUSt3R†

Scene	Method	Acc↓		Comp↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.
chess03	**Dust3R**†	0.0270	0.0093	**0.0180**	0.0055	0.6351	0.7144
	Ours	**0.0237**	**0.0072**	0.0193	**0.0052**	**0.6505**	**0.7389**
chess05	**Dust3R**†	0.0335	0.0141	0.0178	0.0080	0.6352	0.7156
	Ours	**0.0229**	**0.0073**	**0.0142**	**0.0058**	**0.6413**	**0.7249**
pumpkin01	**Dust3R**†	0.0337	0.0133	0.0292	0.0123	**0.7302**	**0.8509**
	Ours	**0.0271**	**0.0131**	**0.0205**	**0.0087**	0.7068	0.8258
pumpkin07	**Dust3R**†	0.0193	**0.0055**	**0.0132**	**0.0052**	0.6793	0.7860
	Ours	**0.0178**	0.0062	0.0164	0.0060	**0.6832**	**0.7929**
stairs01	**Dust3R**†	**0.0636**	**0.0357**	0.1023	0.0193	0.6475	0.7476
	Ours	0.0739	0.0421	**0.0672**	**0.0151**	**0.6507**	**0.7496**
stairs04	**Dust3R**†	0.0475	0.0212	0.0900	0.0174	0.6446	0.7335
	Ours	**0.0390**	**0.0160**	**0.0357**	**0.0069**	**0.6588**	**0.7589**
fire03	**Dust3R**†	0.0112	0.0044	0.0096	0.0042	**0.6539**	**0.7474**
	Ours	**0.0089**	**0.0042**	**0.0086**	**0.0039**	0.6523	0.7454
fire04	**Dust3R**†	0.0104	0.0037	0.0111	0.0037	0.6515	0.7408
	Ours	**0.0086**	**0.0034**	**0.0098**	**0.0036**	**0.6556**	**0.7472**
office02	**Dust3R**†	0.0462	**0.0179**	0.0381	0.0145	0.6819	0.7957
	Ours	**0.0403**	0.0187	**0.0223**	**0.0124**	**0.6843**	**0.7969**
office06	**Dust3R**†	**0.0257**	**0.0152**	**0.0218**	**0.0094**	**0.7195**	**0.8477**
	Ours	0.0879	0.0414	0.0420	0.0154	0.6731	0.7823
office07	**Dust3R**†	0.0270	0.0132	**0.0224**	0.0126	**0.6864**	**0.7944**
	Ours	**0.0269**	**0.0127**	0.0232	**0.0101**	0.6740	0.7772
office09	**Dust3R**†	**0.0351**	**0.0165**	**0.0281**	**0.0102**	**0.6777**	**0.7854**
	Ours	0.0791	0.0279	0.0541	0.0192	0.6579	0.7560
redkit03	**Dust3R**†	**0.0250**	**0.0112**	0.0183	0.0087	**0.6983**	**0.8186**
	Ours	0.0367	0.0203	**0.0158**	**0.0075**	0.6765	0.7849
redkit04	**Dust3R**†	**0.0184**	**0.0069**	0.0235	**0.0059**	**0.6570**	**0.7509**
	Ours	0.0242	0.0083	**0.0179**	0.0061	0.6532	0.7467
redkit06	**Dust3R**†	**0.0240**	0.0127	**0.0170**	**0.0075**	**0.6533**	**0.7442**
	Ours	0.0285	**0.0120**	0.0214	0.0087	0.6404	0.7229
redkit12	**Dust3R**†	**0.0191**	**0.0068**	**0.0161**	**0.0074**	**0.6423**	**0.7271**
	Ours	0.0257	0.0091	0.0178	0.0076	0.6364	0.7211
redkit14	**Dust3R**†	0.0216	0.0087	0.0197	0.0080	0.6332	0.7106
	Ours	**0.0187**	**0.0086**	**0.0171**	**0.0070**	**0.6427**	**0.7264**
heads01	**Dust3R**†	**0.0256**	**0.0056**	**0.0082**	**0.0037**	0.6983	0.8180
	Ours	0.0267	0.0082	0.0098	0.0043	**0.7056**	**0.8288**
Avg.	**Dust3R**†	**0.0286**	**0.0123**	0.0280	0.0091	**0.6681**	**0.7683**
	Ours	0.0342	0.0148	**0.0241**	**0.0085**	0.6635	0.7625

Table 7. **Per-scene results on 7scenes dataset.**

on indoor scene reconstruction. This opens up the possibility of combining optimization-based techniques in DUSt3R with Spann3R within one set of model parameters. Additionally, the inferior results on DTU datasets might be due to 1) Our training set only consists of a small fraction of object-centric scenes. 2) DUSt3R uses an internal pair selection model, which can potentially boost the performance of the object-centric scenes. In contrast, we use a simple strategy of random sampling.

Per-scene performance. We show a per-scene breakdown of quantitative results in Tab. 7 and Tab. 8. Our method achieves competitive per-scene results compared to DUSt3R. However, in some challenging scenes, our model might produce more outliers compared to DUSt3R, which leads to a higher accuracy score. Fig 10 shows an example on the NRGBD dataset, where the scene contains a mirror. This leads our model to produce more outliers and eventually leads to twice higher accuracy compared to DUSt3R.

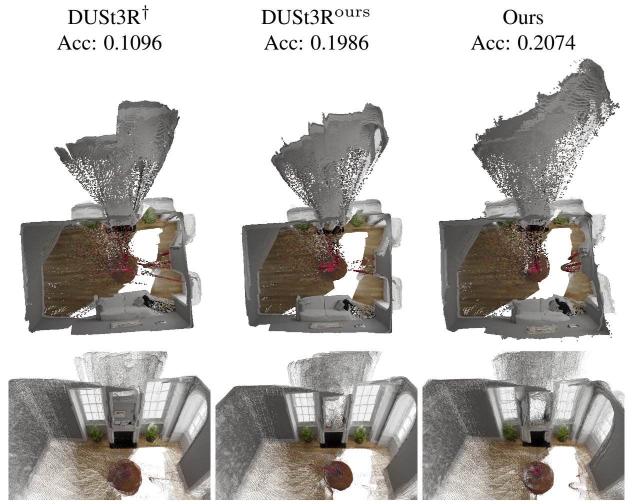

Figure 10. **Qualitative example of outlier scene on NRGBD.** Due to the presence of the mirror, only DUSt3R† reconstructs the geometry of the mirror and produces fewer floaters. We hypothesize this is due to more synthetic training data used in DUSt3R.

Scene	Method	Acc↓		Comp↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.
SC	**Dust3R**†	**0.0731**	0.0370	0.0296	0.0107	**0.7404**	**0.9018**
	Ours	0.0740	**0.0359**	**0.0249**	**0.0106**	0.7234	0.8779
CK	**Dust3R**†	**0.0553**	**0.0252**	**0.0242**	**0.0113**	**0.8167**	**0.9706**
	Ours	0.0916	0.0356	0.0310	0.0139	0.7811	0.9460
GWR	**Dust3R**†	**0.1097**	**0.0348**	**0.0342**	**0.0170**	**0.8198**	**0.9625**
	Ours	0.2074	0.0646	0.0499	0.0224	0.7628	0.9262
MA	**Dust3R**†	**0.0220**	**0.0154**	**0.0158**	0.0090	**0.8126**	**0.9728**
	Ours	0.0250	0.0173	0.0160	**0.0077**	0.8089	0.9677
GR	**Dust3R**†	**0.0330**	**0.0232**	0.0554	**0.0086**	**0.8003**	**0.9534**
	Ours	0.0486	0.0341	**0.0529**	**0.0101**	0.7816	0.9423
Kit.	**Dust3R**†	0.0965	0.0438	0.0656	0.0177	**0.8157**	**0.9732**
	Ours	**0.0649**	**0.0359**	**0.0333**	**0.0132**	0.8140	0.9683
WR	**Dust3R**†	**0.0300**	**0.0170**	**0.0119**	**0.0071**	**0.7866**	**0.9352**
	Ours	0.0426	0.0270	0.0169	0.0081	0.7500	0.9022
BR	**Dust3R**†	0.0476	**0.0211**	0.0343	**0.0061**	0.7460	0.9162
	Ours	**0.0472**	0.0215	**0.0255**	0.0067	**0.7613**	**0.9312**
TG	**Dust3R**†	0.0228	**0.0084**	0.0130	**0.0054**	**0.8838**	**0.9911**
	Ours	**0.0207**	0.0119	**0.0120**	0.0065	0.8150	0.9719
Avg.	**Dust3R**†	**0.0544**	**0.0251**	0.0315	**0.0103**	**0.8024**	**0.9529**
	Ours	0.0691	0.0315	**0.0291**	0.0110	0.7775	0.9371

Table 8. **Per-scene results on NRGBD dataset.**

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 (IJCV)*, 120:153–168, 2016. 5, 6

[2] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building rome in a day. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 72–79, 2009. 1, 2

[3] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 29–42, 2010. 1, 2

[4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 8

[5] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of Learning and Motivation*, pages 89–195. Elsevier, 1968. 2, 3

[6] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022. 5, 6

[7] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 3, 8

[8] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 19697–19705, 2023. 3

[9] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[10] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 1, 2

[11] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4160–4169, 2023. 3

[12] 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 (CVPR)*, pages 2560–2568, 2018. 1

[13] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6684–6692, 2017. 2

[14] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5044–5053, 2023.

[15] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. *arXiv preprint arXiv:2404.14351*, 2024. 2

[16] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 640–658. Springer, 2022. 2, 3, 4

[17] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11781–11794, 2021.

[18] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3151–3161, 2024. 3

[19] David Crandall, Andrew Owens, Noah Snavely, and Dan Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3008, 2011. 1, 2

[20] 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 (CVPR)*, pages 5828–5839, 2017. 5

[21] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1052–1067, 2007. 1, 2

[22] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 224–236, 2018. 1, 2

[23] Eric Dexheimer and Andrew J Davison. Learning a depth covariance function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13122–13131, 2023. 2

[24] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisser-

man. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 10061–10072, 2023. 3

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3, 6

[26] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15324–15333, 2021. 2, 6

[27] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 834–849, 2014. 2

[28] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(3):611–625, 2017. 2

[29] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2009. 1, 2

[30] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 873–881, 2015. 1, 2

[31] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 1, 2

[32] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[33] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH*, pages 1–11, 2024. 3

[34] Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21584–21593, 2024. 3

[35] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 12949–12958, 2021. 3

[36] Wonbong Jang and Lourdes Agapito. Nvist: In the wild new view synthesis from a single image with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10181–10193, 2024. 3

[37] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 3

[38] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024. 2

[39] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21357–21366, 2024. 3

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

[41] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *ISMAR*, pages 1–10, 2007. 1, 2

[42] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 952–961, 2023. 3

[43] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018. 5

[44] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 5987–5997, 2021. 2

[45] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 17627–17638, 2023. 2

[46] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999. 1, 2

[47] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, 2004. 1, 2

[48] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18039–18048, 2024. 3

[49] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 5

[50] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 3, 8

[51] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *EMNLP*, pages 1400–1409, 2016. 2, 3

[52] 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. 2

[53] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011. 2

[54] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2320–2327, 2011. 1, 2

[55] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9226–9235, 2019. 3

[56] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[57] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 6

[58] 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 International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021. 5

[59] Jerome Revaud, Yohann Cabon, Romain Brégier, JongMin Lee, and Philippe Weinzaepfel. Sacreg: Scene-agnostic coordinate regression for visual localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 688–698, 2024. 2

[60] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2564–2571, 2011. 1, 2

[61] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, 2020. 1, 2

[62] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9339–9347, 2019. 5

[63] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–19, 2022. 2, 6

[64] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 2, 3

[65] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518, 2016. 1, 2

[66] 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 (CVPR)*, pages 3260–3269, 2017. 8

[67] 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 (CVPR)*, pages 2930–2937, 2013. 5, 6

[68] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics (TOG)*, 25(3):835–846, 2006. 1, 2

[69] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, 2024. 3

[70] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 573–580, 2012. 8

[71] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6229–6238, 2021. 3

[72] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2440–2448, 2015. 2, 3

[73] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, 2021. 1

[74] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2446–2454, 2020. 5

[75] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 801–809, 2015. 1, 2

[76] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 3

[77] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, pages 298–372, 2000. 1, 2

[78] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13293–13302, 2023. 3, 5

[79] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21686–21697, 2024. 1

[80] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 27171–27183, 2021. 3

[81] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2, 3, 4, 5, 6, 8, 9

[82] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 17969–17980, 2023. 6

[83] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 2, 3

[84] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 61–75, 2014. 1, 2

[85] Changchang Wu. Towards linear-time incremental structure from motion. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 127–134, 2013. 1, 2

[86] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. Multicore bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2011. 1, 2

[87] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20406–20417, 2024. 3

[88] Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, and Feng Zhao. Frozenrecon: Pose-free 3d scene reconstruction with frozen depth models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9276–9286, 2023. 5, 6

[89] 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, 6

[90] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1790–1799, 2020. 5

[91] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4805–4815, 2021. 3

[92] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 5

[93] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 467–483, 2016. 2

[94] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 1, 2

[95] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12786–12796, 2022. 3, 5