Appendix

1. Training Datasets

We trained our model on 32 datasets that covers a diverse range of scene types, including static and dynamic environments, as well as indoor, outdoor, and object-centric scenarios. A complete list of these datasets is provided in Tab. 1.

The original MapFree [1] and DL3DV [21] datasets do not include dense depth maps. We performed multi-view stereo (MVS) reconstruction [28] using the provided camera parameters to generate dense depth maps. This results in complete annotations for these datasets for training. RealEstate10K [53], CoP3D [30], and MVImgNet [48] also do not provide dense depth maps. For these three datasets, we only use the provided camera parameters to supervise the camera prediction. For RealEstate10K, we only include a subset of 2325 training scenes for training.

EDEN [18], IRS [38], Synscapes [42], SmartPortraits [17], and HOI4D [22] are treated as single views. To train on single-view datasets with a specified context length, we construct sequences by stacking independent views to the desired context length, and importantly always reset the state to $\boldsymbol{s}_0$ after each view. This allows us to jointly train using both multi-view and single-view data within the same batch. Although both EDEN [18] and SmartPortraits [17] provide camera poses, EDEN [18] lacks clear documentation of camera conventions, and SmartPortraits [17] offers camera poses that are not synchronized with RGBD frames. Therefore, we treat both as single-view datasets.

For PointOdyssey [52], we filter scenes with incorrect depth annotations (mostly scenes with fogs, like `cab_h_bench_ego2`) and scenes with unrealistic motion and material (like `Ani`). For BEDLAM [4], we remove scenes with panorama backgrounds.

2. More Implementation Details

Sequence Sampling Details. Our training dataset comprises a combination of video sequences and unordered photo collections. For video sequences, we subsample frames at intervals randomly selected between 1 and k, where k is set for each dataset based on its frame rate and camera motion. Within each sequence, either variable or fixed intervals are used, each accounting for approximately half of the samples. For photo collections, we use similar methods as in DUSt3R [39] and compute the overlap ratios between images to guide the frame sampling. Additionally, when the scene from a video is largely static, we shuffle the frames and treat them as a photo collection to increase data diversity. When the sequences contain major dynamic objects (like sequences from BEDLAM [4] and PointOdyssey [52] datasets), we only treat them as videos and feed frames into the model in temporal order using a fixed interval.

When the data is metric scale, frames (excluding the first frame) in a sequence are randomly masked with a 20% probability and replaced by their corresponding raymap inputs, using ground truth intrinsics and poses. Note that raymap mode is activated only when data are in metric scale, as our model learns metric-scale 3D scene priors. When the 3D annotation is at an unknown scale, raymap querying is disabled to avoid scale inconsistency with the scene content captured in the state.

More Architecture Details. Similar to DUSt3R [39], we reduce training costs by first training the model on 224×224 image resolution with linear heads, and then increasing the resolution and setting the longer side of the images to 512 pixels. Specifically, in the first two stages of training, $\mathrm{Head}_{\mathrm{self}}$ and $\mathrm{Head}_{\mathrm{world}}$ are implemented as linear layers. In the final two stages, $\mathrm{Head}_{\mathrm{self}}$ and $\mathrm{Head}_{\mathrm{world}}$ are switched to DPT [25] architecture. Compared to $\mathrm{Head}_{\mathrm{self}}$, $\mathrm{Head}_{\mathrm{world}}$ incorporates an additional modulation function, which modulates $\boldsymbol{F}'_t$ using the pose token $\boldsymbol{z}'_t$ within the Layer Normalization layers. This modulation design is inspired by LRM [11] and aims to integrate pose information to achieve implicit rigid transformations. Specifically, within $\mathrm{Head}_{\mathrm{world}}$, we first use two self-attention blocks modulated by the pose token $\boldsymbol{z}'_t$ to generate the pose-modulated tokens, which is then fed as input to either the linear or DPT architecture to generate the final pointmap output $\hat{\boldsymbol{X}}_t^{\mathrm{world}}$. The dimension of $\boldsymbol{z}'_t$ is 768, and $\mathrm{Head}_{\mathrm{pose}}$ is a 2-layer MLP whose hidden size is 768. We apply Rotary Positional Embedding (ROPE) [32] to the query and key feature before each attention operation.

More Training Details. In the first stage of training, we use the following datasets: ARKit, ARKit-HighRes, ScanNet, ScanNet++, TartanAir, Waymo, MapFree, BlendedMVS, HyperSim, MegaDepth, Unreal4K, DL3DV, CO3Dv2, WildRGBD, and VirtualKITTI2. In the second stage, we incorporate the rest of datasets. In the final stage (long context training), we exclude single-view datasets (EDEN, IRS, Synscapes, 3D Ken Burns, SmartPortraits, UrbanSyn, and HOI4D) and train only on multi-view datasets, as the goal of the final stage training is to enhance scene-level reasoning within a sequence. Unlike DUSt3R, which applies color

Dataset Name	Scene Type	Metric?	Real?	Dynamic?	Camera only?	Single View?
ARKitScenes [2]	Indoor	Yes	Real	Static	No	No
ARKitScenes-HighRes [2]	Indoor	Yes	Real	Static	No	No
ScanNet [9]	Indoor	Yes	Real	Static	No	No
ScanNet++ [47]	Indoor	Yes	Real	Static	No	No
TartanAir [40]	Mixed	Yes	Synthetic	Dynamic	No	No
Waymo [33]	Outdoor	Yes	Real	Dynamic	No	No
MapFree [1]	Outdoor	Yes	Real	Static	No	No
BlendedMVS [46]	Mixed	No	Synthetic	Static	No	No
HyperSim [27]	Indoor	Yes	Synthetic	Static	No	No
MegaDepth [20]	Outdoor	No	Real	Static	No	No
Unreal4K [36]	Mixed	Yes	Synthetic	Static	No	No
DL3DV [21]	Mixed	No	Real	Static	No	No
CO3Dv2 [26]	Object-Centric	No	Real	Static	No	No
WildRGBD [44]	Object-Centric	Yes	Real	Static	No	No
VirtualKITTI2 [6]	Outdoor	Yes	Synthetic	Dynamic	No	No
Matterport3D [7]	Indoor	Yes	Real	Static	No	No
BEDLAM [4]	Mixed	Yes	Synthetic	Dynamic	No	No
Dynamic Replica [14]	Indoor	Yes	Synthetic	Dynamic	No	No
PointOdyssey [52]	Mixed	Yes	Synthetic	Dynamic	No	No
Spring [23]	Mixed	Yes	Synthetic	Dynamic	No	No
MVS-Synth [13]	Outdoor	Yes	Synthetic	Dynamic	No	No
UASOL [3]	Outdoor	Yes	Real	Static	No	No
OmniObject3D [43]	Object-Centric	Yes	Synthetic	Static	No	No
RealEstate10K [53]	Indoor	No	Real	Static	Yes	No
MVImgNet [48]	Object-Centric	No	Real	Static	Yes	No
CoP3D [30]	Object-Centric	No	Real	Dynamic	Yes	No
EDEN [18]	Outdoor	Yes	Synthetic	Static	No	Yes
IRS [38]	Indoor	Yes	Synthetic	Static	No	Yes
Synscapes [42]	Outdoor	Yes	Synthetic	Dynamic	No	Yes
3D Ken Burns [24]	Mixed	No	Synthetic	Static	No	Yes
SmartPortraits [17]	Indoor	Yes	Real	Dynamic	No	Yes
UrbanSyn [10]	Outdoor	Yes	Synthetic	Dynamic	No	Yes
HOI4D [22]	Indoor	Yes	Real	Dynamic	No	Yes

Table 1. **Training Datasets.** We provide more details of our training datasets. We classify a dataset as dynamic if annotations exist for moving objects like humans. If there is only camera parameters (intrinsics and extrinsics) available, we mark them as "camera only". If the dataset only contains depth and intrinsics for single views, we mark them as "single view".

jittering to each image independently, we perform sequence-level color jittering by applying the same color jitter across all frames in a sequence.

3. More Comparisons

Video Depth Estimation. We expand the video depth comparison in the main paper and compare with a wider range of baseline methods, including single-frame depth techniques (Marigold [15] and Depth-Anything-V2 [45]), video depth approaches (NVDS [41], ChronoDepth [29], and DepthCrafter [12]), and joint depth-and-pose methods such as Robust-CVD [16], CasualSAM [50], DUSt3R [39], MASt3R [19], MonST3R [49], and Spann3R [37]. The results are shown in Tab. 2.

Camera Pose Estimation Similar to video depth estimation, we include a diverse set of baselines for camera pose estimation. Learning-based visual odometry methods, such as DROID-SLAM [34], DPVO [35], and LEAP-VO [8], require ground truth camera intrinsics as input. Optimization-based methods, including Particle-SfM [51], Robust-CVD [16], CasualSAM [50], DUSt3R-GA [39], MASt3R-GA [19], and MonST3R-GA [49], generally operate more slowly com-

Alignment	Method	Optim.	Onl.	Sintel Abs Rel ↓	Sintel δ<1.25 ↑	BONN Abs Rel ↓	BONN δ<1.25 ↑	KITTI Abs Rel ↓	KITTI δ <1.25 ↑	FPS
Per-sequence scale & shift	Marigold [15]		✓	0.532	51.5	0.091	93.1	0.149	79.6	<0.1
	Depth-Anything-V2 [45]		✓	0.367	55.4	0.106	92.1	0.140	80.4	3.13
	NVDS [41]		✓	0.408	48.3	0.167	76.6	0.253	58.8	-
	ChronoDepth [29]		✓	0.687	48.6	0.100	91.1	0.167	75.9	1.89
	DepthCrafter [12]		✓	**0.292**	**69.7**	0.075	**97.1**	0.110	88.1	0.97
	Robust-CVD [16]		✓	0.703	47.8	-	-	-	-	-
	CasualSAM [50]	✓		0.387	54.7	0.169	73.7	0.246	62.2	-
	DUSt3R-GA [39]	✓		0.531	51.2	0.156	83.1	0.135	81.8	0.76
	MASt3R-GA [19]	✓		0.327	59.4	0.167	78.5	0.137	83.6	0.31
	MonST3R-GA [49]	✓		0.333	59.0	**0.066**	96.4	0.157	73.8	0.35
	Spann3R [37]		✓	0.508	50.8	0.157	82.1	0.207	73.0	13.55
	Ours		✓	0.454	55.7	0.074	94.5	**0.106**	**88.7**	16.58
Per-sequence scale	DUSt3R-GA [39]	✓		0.656	45.2	0.155	83.3	0.144	81.3	0.76
	MASt3R-GA [19]	✓		0.641	43.9	0.252	70.1	0.183	74.5	0.31
	MonST3R-GA [49]	✓		**0.378**	**55.8**	**0.067**	**96.3**	0.168	74.4	0.35
	Spann3R [37]		✓	0.622	42.6	0.144	81.3	0.198	73.7	13.55
	Ours		✓	0.421	47.9	0.078	93.7	**0.118**	**88.1**	16.58
Metric scale	MASt3R-GA [19]	✓		**1.022**	14.3	0.272	70.6	0.467	15.2	0.31
	Ours		✓	1.029	**23.8**	**0.103**	**88.5**	**0.122**	**85.5**	16.58

Table 2. **Video Depth Evaluation**. We report scale&shift-invariant depth, scale-invariant depth and metric depth accuracy on Sintel, Bonn, and KITTI datasets. Methods requiring global alignment are marked "GA", while "Optim." and "Onl." indicate optimization-based and online methods, respectively. We also report the FPS on KITTI dataset using 512× 144 image resolution for all methods, except Spann3R which only supports 224×224 inputs.

Method	Optim.	Onl.	Sintel ATE ↓	Sintel RPE trans ↓	Sintel RPE rot ↓	TUM-dynamics ATE ↓	TUM-dynamics RPE trans ↓	TUM-dynamics RPE rot ↓	ScanNet ATE ↓	ScanNet RPE trans ↓	ScanNet RPE rot ↓
DROID-SLAM [34]		✓	0.175	0.084	1.912	-	-	-	-	-	-
DPVO [35]		✓	0.115	0.072	1.975	-	-	-	-	-	-
LEAP-VO [8]		✓	**0.089**	**0.066**	**1.250**	0.068	0.008	1.686	0.070	0.018	0.535
Particle-SfM [51]	✓		0.129	0.031	**0.535**	-	-	-	0.136	0.023	0.836
Robust-CVD [16]	✓		0.360	0.154	3.443	0.153	0.026	3.528	0.227	0.064	7.374
CasualSAM [50]	✓		0.141	**0.035**	0.615	0.071	**0.010**	1.712	0.158	0.034	1.618
DUSt3R-GA [39]	✓		0.417	0.250	5.796	0.083	0.017	3.567	0.081	0.028	0.784
MASt3R-GA [19]	✓		0.185	0.060	1.496	**0.038**	0.012	**0.448**	0.078	0.020	**0.475**
MonST3R-GA [49]	✓		**0.111**	0.044	0.869	0.098	0.019	0.935	**0.077**	**0.018**	0.529
DUSt3R [39]		✓	0.290	0.132	7.869	0.140	0.106	3.286	0.246	0.108	8.210
Spann3R [37]		✓	0.329	0.110	4.471	0.056	0.021	0.591	**0.096**	0.023	0.661
Ours		✓	**0.213**	**0.066**	**0.621**	**0.046**	**0.015**	**0.473**	0.099	**0.022**	**0.600**

Table 3. **Evaluation on Camera Pose Estimation** on Sintel [5], TUM-dynamic [31], and ScanNet [9] datasets. Note that unlike the the rest of the methods, the three methods in the first section require ground truth camera intrinsics as input.

pared to online methods like Spann3R [37] and our proposed approach. To assess performance in an online setting, we also evaluate DUSt3R without global alignment. The results are presented in Tab. 3.

References

[1] 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 *ECCV*, 2022. 1, 2

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2

[3] Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escolano, and Miguel Cazorla. Uasol, a large-scale high-resolution outdoor stereo dataset. *Scientific data*, 6(1): 162, 2019. 2

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1, 2

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012. 3

[6] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 2

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2

[8] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19844–19853, 2024. 2, 3

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 3

[10] Jose L. Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A. Iglesias-Guitian, and Antonio M. López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes, 2023. 2

[11] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1

[12] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 2, 3

[13] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[14] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023. 2

[15] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[16] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 2, 3

[17] Anastasiia Kornilova, Marsel Faizullin, Konstantin Pakulev, Andrey Sadkov, Denis Kukushkin, Azat Akhmetyanov, Timur Akhtyamov, Hekmat Taherinejad, and Gonzalo Ferrer. Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21318–21329, 2022. 1, 2

[18] Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. Eden: Multimodal synthetic dataset of enclosed garden scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1579–1589, 2021. 1, 2

[19] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 3

[20] 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*, pages 2041–2050, 2018. 2

[21] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 1, 2

[22] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 1, 2

[23] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2023. 2

[24] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019. 2

[25] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 1

[26] 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 *International Conference on Computer Vision*, 2021. 2

[27] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 2

[28] Johannes Lutz Schönberger and Jan-Michael Frahm.

Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[29] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. 2, 3

[30] Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, and David Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. *CVPR*, 2023. 1, 2

[31] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 3

[32] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

[33] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[34] 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. 2, 3

[35] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 2023. 2, 3

[36] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8942–8952, 2021. 2

[37] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 3

[38] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1, 2

[39] 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

[40] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 2

[41] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9466–9476, 2023. 2, 3

[42] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 1, 2

[43] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[44] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024. 2

[45] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2, 3

[46] 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. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[47] 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. 2

[48] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 1, 2

[49] 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, 3

[50] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 2, 3

[51] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pages 523–542. Springer, 2022. 2, 3

[52] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 1, 2

[53] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 1, 2