Any4D: Towards Unified Feed-Forward Metric 4D Reconstruction

Jay Karhade Nikhil Keetha Tanisha Gupta Yuchen Zhang
Akash Sharma Sebastian Scherer Deva Ramanan
Carnegie Mellon University

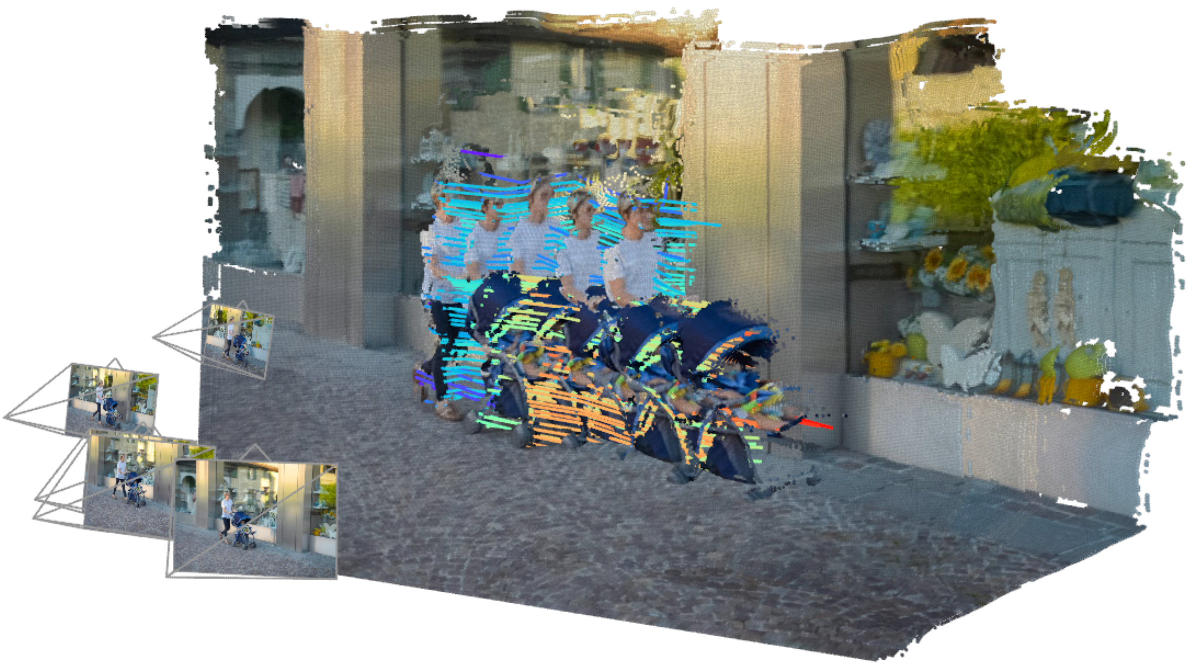

Figure 1. Any4D is a feed-forward model capable of producing dense 4D reconstructions of dynamic scenes. Any4D is flexible in accommodating diverse sensor inputs to improve performance, can produce predictions upto 72% faster than existing methods. Note that while Any4D produces both dense 3D tracking vectors, the figure above only visualizes the sparse motion tracks.

Abstract

We present Any4D, a framework for feed-forward metric-scale dense 4D reconstruction. Compared to other recent methods for feedforward 4D reconstruction from monocular RGB videos, Any4D is multimodal due to its focus on diverse camera setups, allowing it to process additional modalities and sensors when available, such as RGB-D frames, IMU-based egomotion and Doppler measurements from Radar. Moreoever, Any4D can directly generate dense feedforward predictions for N frames, in contrast to prior work that typically focuses on either 2-view dense scene flow or sparse 3D point tracking. One of the innovations that allow for such flexible input modalities is a modular approach to representing the 4D scene; specifically, 4D predictions are encoded using a variety of egocentric factors (such as depthmaps and camera intrinsics) represented in local camera coordinates, as well as allocentric factors (such as camera extrinsics and scene flow) represented in global world coordinates. We show that Any4D achieves superior performance over existing methods across diverse sensor setups - both in terms of accuracy and compute efficiency, opening up avenues for real-time deployment in downstream robotics applications.

1. Introduction

Reconstructing the 4D (3D + t) world from sensor observations has been a long-standing goal of computer vision. Such a technology can unlock transformative capabilities

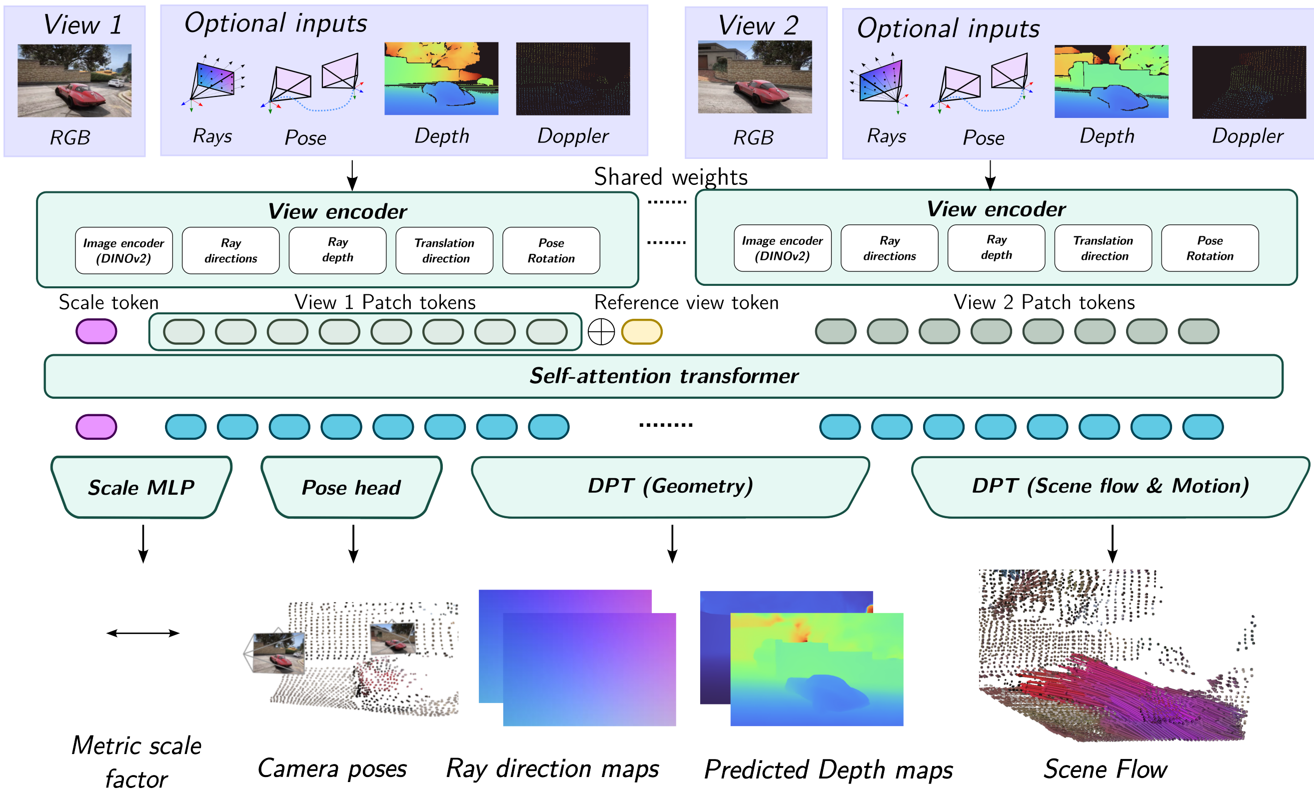

Figure 2. Any4D produces dense 4D metric-scale 4D reconstruction of a scene in the form of factorised outputs consisting of ego-centric factors such scale factor, depth maps, ray directions and allo-centric factors such as normalized forward scene flow and camera poses. Note that while we show only 2 views for simplicity, Any4D can take arbitrary number of views at test time.

across a wide range of downstream tasks. In generative AI, for instance, 4D reconstruction can improve dynamic video synthesis, video editing and understanding, as well as the creation of interactive dynamic assets such as VR avatars. In robotics, 4D reconstruction can improve predictive control (MPC) for robot navigation and manipulation[16].

While there has been tremendous progress in recent history on 4D reconstruction [9, 17, 20, 26], dynamic reconstruction remains challenging for many reasons. First, 4D reconstruction is severely under-constrained, requiring environment simplifying assumptions such as rigid motion, smoothness priors, or a mostly-static world. Second, because 4D reconstruction and tracking is such a challenging problem, marked progress has been achieved by treating dynamic attribute prediction as independent sub tasks (i.e., 2D/3D tracking [5, 6, 15, 23, 25], video-consistent depth estimation[7, 10, 21, 26], scene flow estimation [11, 14, 18, 22] in dynamic scenes). Third, this focus on sub-tasks has led to fragmented datasets and benchmarks that lack consistent and coherent 4D annotations.

In this work, we focus on 4D reconstruction for the motivating task of autonomous robots. This in turn motivates us to seek a solution with following desiderata: a) **metric scale outputs**: while most existing 4D reconstruction methods produce outputs in a normalized coordinate frame, physical agents undeniably operate in the metric-scale physical world; b) **multimodal inputs**: Many robotic platforms make use of additional sensors[2, 4**?**], but most prior work fails to exploit such diverse configurations. c) **efficiency**: much prior work often makes use of iterative optimization-based methods that maybe too slow for real-time deployment.

We present Any4D, a unified framework for feed-forward metric-scale, multimodal, and dense 4D reconstruction from arbitrary image sequences, whilst being *up to 72% faster* than baselines. Concretely, we propose the following 3 contributions:

- **Dense metric-scale 4D reconstruction**: Any4D predicts the dense geometry and motion of the scene in metric coordinates, unlike existing methods that can reconstruct only up-to-scale or sparse tracks. To do so, we exploit our factored 4D representation and train on diverse datasets with partial annotations, including metric-scale 3D reconstruction datasets without motion annotations as well as non-metric datasets with motion annotations.
- **Multi-modal conditioning**: When available, Any4D im-

Method	Scale	PStudio			Dynamic Replica			Drive Track		
	abs_rel ↓	EPE ↓	Inliers ↑	Outliers ↓	EPE ↓	Inliers ↑	Outliers ↓	EPE ↓	Inliers ↑	Outliers ↓
Monst3R + CoTracker3	-	0.6561	0.2641	0.6789	0.8108	0.3577	0.5208	11.0635	0.1363	0.7163
VGGT + CoTracker3	-	0.4405	0.3630	0.3124	0.7465	0.4336	0.4918	6.2421	0.5281	0.3356
SpaTrackv2	-	0.3215	0.5957	0.1765	0.6879	0.6062	0.2313	4.7475	0.6239	0.2812
Any4D Image-Only	1.5%	0.4129	0.5298	0.3320	0.0803	0.9513	0.0282	3.9670	0.7008	0.1800
Any4D Images + Geometry + Doppler	-	0.3352	0.6092	0.2675	0.07219	0.9576	0.0206	3.598	0.7031	0.1458
Any4D-SpaTrackv2 Image-Only	1.5%	0.3398	0.7202	0.1549	0.7009	0.5791	0.24110	3.683	0.738	0.1534
Any4D-SpaTrackv2 Images + Geometry + Doppler	-	0.1911	0.7846	0.1156	0.6908	0.5651	0.238	3.435	0.752	0.1367

Table 1. Any4D achieves state-of-the-art performance on all methods, while also predicting geometry and motion in metric scale unlike other baselines which produce upto-scale tracking and reconstruction.

proves 4D reconstruction by exploiting modalities like depth, poses, and Doppler from additional sensors.
- **Efficient feed-forward reconstruction**: Any4D infers both geometry and motion from images in a single feed-forward pass, while also being a strong front-end model that improves joint-optimization based methods.

2. Any4D

Any4D is a framework for producing dense metric-scale 4D reconstruction in a feed-forward manner exploiting multi-modal sensor inputs usually forming the sensing package in robots, including RGB cameras, and optionally depth, IMU based depth, and doppler measurements from radar. Any4D takes as input, a set of RGB images $\mathbf{I} \triangleq \{I_i\}_{i=1}^N$ and optional auxiliary multi-modal sensor inputs denoted as $\mathbf{O} \triangleq (O_i)_{i=1}^N$. Any4D is a function that maps these inputs to a factored output representation:

$$\text{Any4D}\big(\mathbf{I}, \mathbf{O}\big) = (\tilde{s}, \{\tilde{R}_i, \tilde{D}_i, \tilde{T}_i, \tilde{F}_i\}_{i=1}^N), \tag{1}$$

where specifically the optional inputs $\mathbf{O}$ can contain information such as depth maps, calibrated camera intrinsics, camera poses from IMUs and measured doppler velocity from RADAR. The output on the other hand represented with $\sim$ is a factored representation of the 4D scene, consisting of a global metric scaling factor $\tilde{s} \in \mathbb{R}$, egocentric quantities predicted in the camera coordinate frame, namely
- Predicted Ray directions per view $\tilde{R}_i \in \mathbb{R}^{3\times H\times W}$
- Normalized predicted depths along the rays per view $\tilde{D}_i \in \mathbb{R}^{1\times H\times W}$.

and *allocentric* quantities predicted in a world coordinate frame chosen as the first view camera coordinate frame,
- Normalized forward scene flow from the first view to all other views, denoted as $\tilde{F}_i \in \mathbb{R}^{3\times H\times W}$.
- Camera pose of each view in the coordinate system of the first view $\tilde{T}_i \triangleq [p_i, q_i] \in \mathbb{R}^7$ represented using a translation vector and quaternion.

From this factorized representation, one can recover the predicted metric-scale geometry $\tilde{\mathbf{G}}_i$ and scene motion as $\tilde{M}_i$ as:

$$\tilde{G}_i = \tilde{s} \cdot \tilde{R}_i \cdot \tilde{D}_i, \tilde{M}_i = s \cdot \tilde{F}_i \quad \in \mathbb{R}^{3\times H\times W} \tag{2}$$

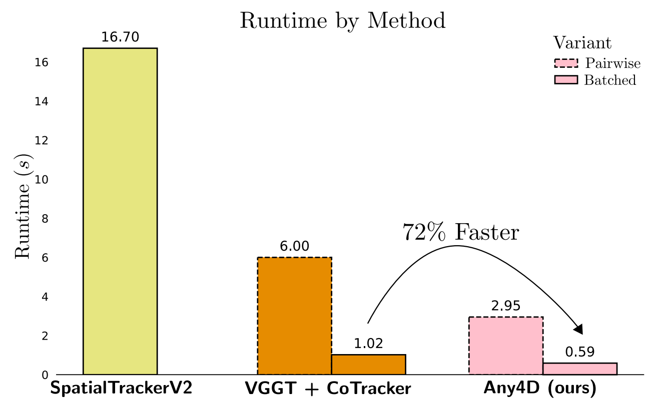

Figure 3. Unlike prior methods, Any4D can produce comparable results with just one single feed-forward pass. Note that in the batched setting, Any4D processes multiple input views at once demonstrating inference generalization to multiple input views.

3. Benchmarking

Benchmarks: There is a lack of standard and unified benchmarks for evaluating 4D reconstruction in existing literature. On one hand, there are datasets for benchmarking dynamic depth, while on the other hand there exist benchmarks for egocentric 3D point-tracking and egocentric scene-flow [1, 8, 12, 13]. Unfortunately, there exists no standard benchmark for allocentric tracking or reconstruction. Hence, we take insipration from concurrent work [3] and repurpose existing datasets to form an allocentric 4D reconstruction benchmark. Unlike [3] we choose to drop ADT due to the lack of motion in this dataset. We use Parallel-Studio and Drive-Track datasets, and add Dynamic Replica, which contains camera motion along with 3D point tracking labels. The final benchmark contains 192 total sequences uniformly spread across the 3 datasets, and consist of 50-64 frames sampled from each of the sequences.

Baselines and Metrics: We compare Any4D against state-of-art dynamic 3D reconstruction reconstruction methods with state of the art 2D point tracking algorithms, including Monst3R [26](pairwise feedforward + post-optimization) VGGT [19](N-view feedforward) with CoTracker3[5] as 2D tracker, and SpatialTrackerv2 [24].

Figure 4. Qualitative Visualization of Any4D estimating 3D geometry and point tracking on diverse scenes

We evaluate these methods using *predicted 3D points after motion* in allocentric coordinates on end-point-error, inliers within 10 percent relative error, and outliers greater than 15 percent relative error.

Results As we see from Table 1. , Any4D achieves stronger results on all datasets compared to the feed-forward baselines Monst3R and VGGT with CoTrackerv3 and can achieve over 20% higher inlier performance and 10-15% lesser outliers across all datasets. Furthermore, while SpaTrackv2 which is a recurrent joint-optimization method indeed beats the feed-forward only model, Any4D shows a strong boost when coupled with the recurrent joint optimization, beating SpaTrackv2 by nearly 2x on EPE and upto 5-10% inlier rates. Any4D performance is further boosted when conditioned with geometry and doppler resulting in 10% lower EPE and outliers.

4. Conclusion

In this paper, we presented Any4D, a unified model that enables metric 4D reconstruction of dynamic scenes from both monocular and multimodal setups. We chose a factorized output representation of 4D scenes, which allows for using diverse data with partial supervision for auxiliary sub-tasks in addition to the target task of dense scene flow estimation. Finally, due to the feed-forward nature of Any4D, we showed that during inference one can obtain dynamic scene estimates an order of magnitude faster than existing methods. Any4D generalizes to many different scenarios and can be improved with more availability of large-scale 4D data. We believe Any4D can serve as a foundation model prior, enabling real-time 4D scene reconstruction for applications such as Generative AI, AR/VR and Robotics.

References

[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 3

[2] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024. 2

[3] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J Black, Trevor Darrell, and Angjoo Kanazawa. St4rtrack: Simultaneous 4d reconstruction and tracking in the world. *arXiv preprint arXiv:2504.13152*, 2025. 3

[4] Tianshu Huang, Akarsh Prabhakara, Chuhan Chen, Jay Karhade, Deva Ramanan, Matthew O'Toole, and Anthony Rowe. Towards foundational models for single-chip radar. *arXiv preprint arXiv:2509.12482*, 2025. 2

[5] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *arxiv*, 2024. 2, 3

[6] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proc. ECCV*, 2024. 2

[7] 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

[8] Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, João Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. Tapvid-3d: A benchmark for tracking any point in 3d, 2024. 3

[9] Jiahui Lei, Yijia Weng, Adam W Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6165–6177, 2025. 2

[10] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024. 2

[11] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019. 2

[12] 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. 3

[13] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[14] Himangi Mittal, Brian Okorn, and David Held. Just go with the flow: Self-supervised scene flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11177–11185, 2020. 2

[15] Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. Delta: Dense efficient long-range 3d tracking for any video. *arXiv preprint arXiv:2410.24211*, 2024. 2

[16] Dantong Niu, Yuvan Sharma, Haoru Xue, Giscard Biamby, Junyi Zhang, Ziteng Ji, Trevor Darrell, and Roei Herzig. Pretraining auto-regressive robotic models with 4d representations. *arXiv preprint arXiv:2502.13142*, 2025. 2

[17] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4gm: Large 4d gaussian reconstruction model. In *Advances in Neural Information Processing Systems*, 2024. 2

[18] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[19] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[20] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *International Conference on Computer Vision (ICCV)*, 2025. 2

[21] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024. 2

[22] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: Cost volume on point clouds for (self-) supervised scene flow estimation. In *European Conference on Computer Vision*, pages 88–107. Springer, 2020. 2

[23] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 2

[24] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Iurii Makarov, Bingyi Kang, Xin Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. In *ICCV*, 2025. 3

[25] Bowei Zhang, Lei Ke, Adam W Harley, and Katerina Fragkiadaki. Tapip3d: Tracking any point in persistent 3d geometry. *arXiv preprint arXiv:2504.14717*, 2025. 2

[26] 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