MoRE: 3D Visual Geometry Reconstruction Meets Mixture-of-Experts

Supplementary Material

1. Additional Visualization

We present additional visualizations of the reconstructed pointmaps in Fig. 1, including results from both real-world (first-row) and synthetic inputs (second-row). MoRE produces highly accurate and realistic reconstructions in a single forward pass, demonstrating its potential to further benefit downstream 3D applications. We also provide a visualization of the expert preference distribution across different types of datasets in Fig. 2.

2. Additional Comparison

Pointmap Comparison. We provide additional pointmap comparisons in Fig. 4, where our method reconstructs more accurate pointmaps and captures more realistic structural details.

Normal Comparison. We also present normal prediction comparisons with previous methods in Fig. 5, showing that our approach produces sharper and more accurate surface normals.

Video Depth Comparison. We also provide video depth estimation comparison in Tab. 4. It can be seen that MoRE can also achieve accurate depth estimation for dynamic scenarios.

Model Comparison We further compare MoRE against two categories of baselines in Tab. 1 to demonstrate its effectiveness: 1) **Dense models with equivalent active parameters** (FLARE, VGGT); and 2) **A dense model with equivalent total parameters** (Aether). Results indicate that MoRE: 1) outperforms efficiency-matched models without sacrificing inference speed; and 2) surpasses the capacity-matched dense model in both accuracy and speed. Notably, directly scaling the dense architecture to 5B resulted in demanding training costs and optimization difficulties. In contrast, our method offers a scalable approach to increase model capacity without inducing additional computational overhead.

Model	#Params	Sintel ATE ↓	TUM-D ATE ↓	Bonn Abs Rel ↓	Kitti Abs Rel ↓	FPS (Kitti)	GPUs
FLARE	1.4B	0.207	0.026	0.142	0.357	1	64 A800-80G
VGGT	1.2B	0.167	0.012	0.052	0.052	23	64 A100-80G
Aether	5.5B	0.189	0.092	0.308	0.054	3	80 A100-80G
Ours	1.6B (5.3B)	**0.101**	**0.010**	**0.042**	**0.046**	21	48 H20-96G

Table 1. Comparison with efficiency/capacity-matched baselines.

Ablation Study. Furthermore, Fig. 3 present additional ablation results on confidence-based depth refinement, demonstrating that the proposed design effectively enhances depth accuracy. In addition, Fig. 6 and Tab. 3 present ablation results on dense semantic fusion for normal prediction, demonstrating that the proposed fusion effectively contributes to sharper and more precise normal estimations.

Model	DTU			NYUv2		RealEstate10K		
	Acc. ↓	Comp. ↓	N.C. ↑	Abs Rel ↓	$\delta < 1.25$ ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑
VGGT (baseline)	1.338	1.896	0.676	0.056	0.951	99.97	93.13	77.62
w/o $\mathcal{L}_p$, $\mathcal{L}_d$, w/o MoE	1.336	1.884	0.678	0.058	0.950	99.92	93.21	77.85
w/o $\mathcal{L}_p$, w/o MoE	1.327	1.767	0.679	0.053	0.954	99.93	93.52	79.86
w/o MoE	1.297	1.625	0.682	0.054	0.953	99.94	94.27	85.14
Ours	**1.011**	**1.482**	**0.695**	**0.051**	**0.957**	**99.98**	**95.47**	**86.28**

Table 2. Ablation study on the key components of our model.

To further validate the key components, we conducted training ablations where **all variants** were fine-tuned on the **exact same training dataset (including internal)** for the same epochs. We denote the losses as: $\mathcal{L}_p = \mathcal{L}_{\text{pts\_local}} \& \mathcal{L}_{\text{pts\_n}}$, $\mathcal{L}_d = \mathcal{L}_{\text{depth}}$. Tab. 2 presents the ablation study of our key components. Notably, the first variant corresponds to the *fine-tuned VGGT* baseline, which reveals a critical insight: **without strict data proportion curation, the dense baseline fails to adapt to the new data distribution,** leading to performance degradation. In contrast, adding our proposed designs yields consistent gains.

3. Additional Implementation Details

We implement our MoE framework based on DeepSpeed-MoE [6]. The hidden dimension of each MoE layer is set to 1024, with 8 experts and a top-k routing of $k = 2$. Overall, the model architecture comprises 48 MoE layers, corresponding to the 24 global-attention and 24 frame-attention blocks in the alternating-attention structure of VGGT [9]. The model is trained on 48 NVIDIA H20 GPUs for approximately two weeks. The normal-prediction head, which incorporates dense semantic fusion, requires an additional three days of fine-tuning on high-quality datasets. It is noteworthy that the proposed **dense semantic fusion** and **confidence-based depth refinement** modules can be integrated into **any multi-view method** to enhance the sharpness and detail of geometric predictions.

4. Limitation and Future Work

Although MoRE achieves superior performance in geometric prediction, it still has limitations and can be further improved. First, accurately reconstructing thin structures remains challenging due to the noisy training data. Second, while our confidence-based depth refinement alleviates part of the noise, multi-view methods still fall short of monocular methods in producing sharp and detailed results in monocular depth estimation. We attribute this gap

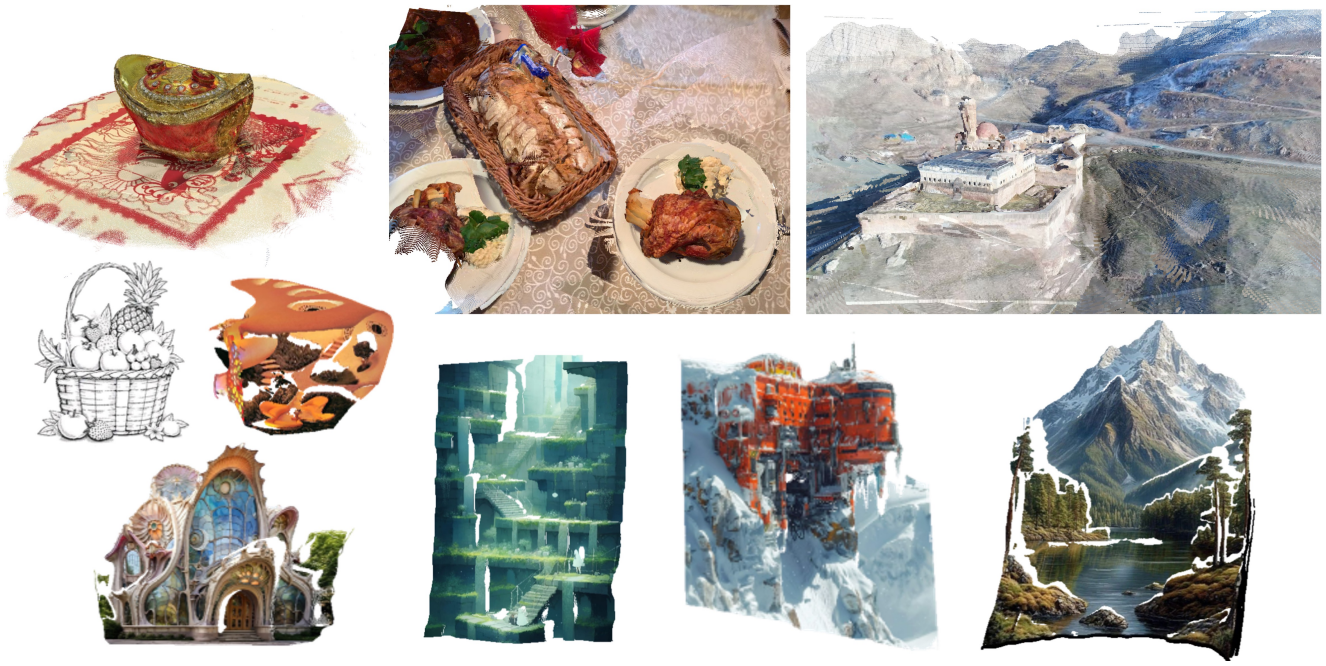

Figure 1. Additional pointmap visualization. MoRE achieves highly accurate and realistic reconstructions for both real-world (first-row) and synthetic inputs (second-row).

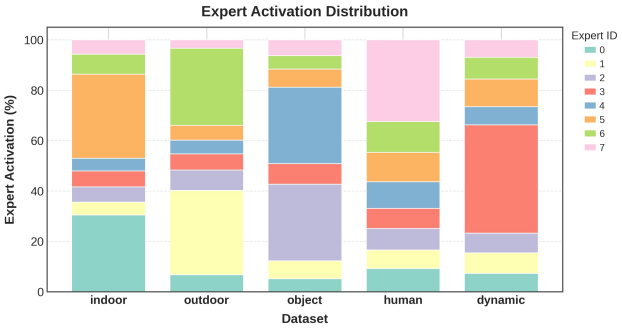

Figure 2. Distribution of expert preference across different types of datasets.

to the quality of depth supervision and the need for a more compact feature representation to better preserve fine-scale details across views. Finally, extending the model with a metric-scale prediction head represents a promising direction for enhancing its applicability in industrial and real-world scenarios.

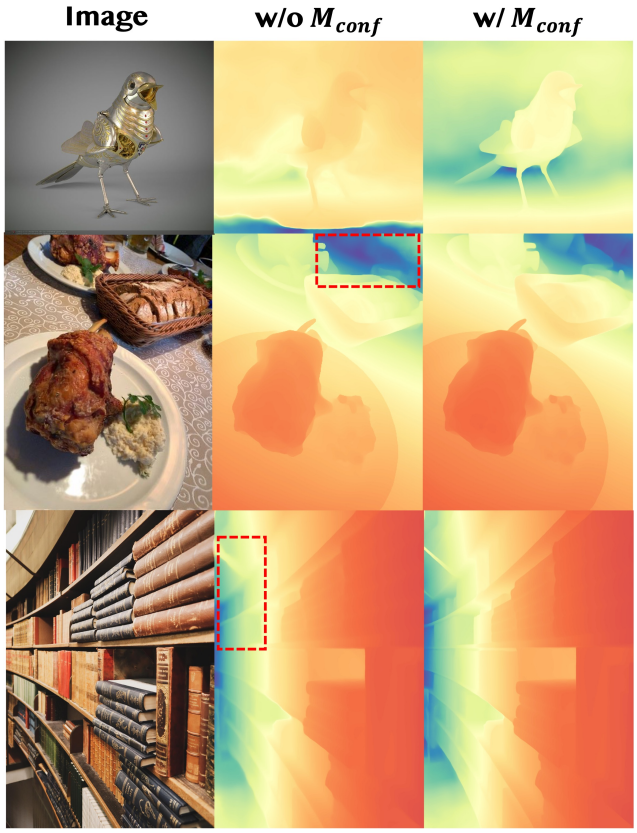

Figure 3. Additional ablation for confidence-based depth refinement. Please zoom in for better details.

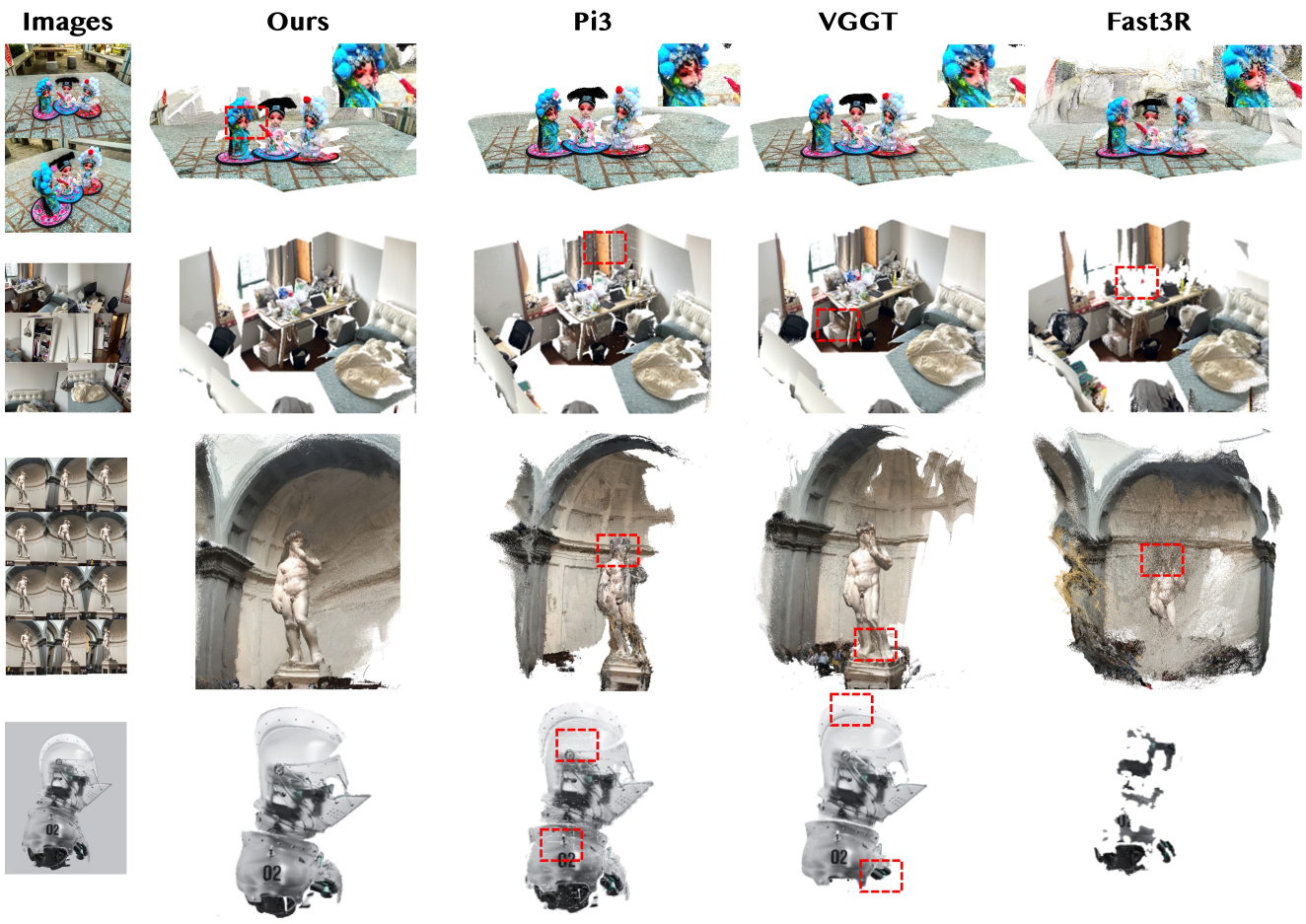

Figure 4. Additional pointmap comparison. Our method can produce more accurate pointmaps and capture more realistic structural details.

	NYUv2 [7]		ScanNet [14]		IBims [3]	
	Mean ↓	$\delta_{11.25^\circ}$ ↑	Mean ↓	$\delta_{11.25^\circ}$ ↑	Mean ↓	$\delta_{11.25^\circ}$ ↑
w/o f_s	15.6	63.0	16.5	63.8	15.5	71.2
Ours	**15.1**	**63.5**	**16.1**	**64.4**	**15.0**	**72.6**

Table 3. Normal quantitative metrics for ablation. We demonstrate that the dense semantic feature fusion (f_s) contributes to more accurate normal predictions.

Method	Sintel [1]		Bonn [5]		KITTI [2]	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
DUSt3R [11]	0.570	0.493	0.152	0.835	0.135	0.818
MASt3R [4]	0.480	0.517	0.189	0.771	0.115	0.849
MonST3R [15]	0.402	0.526	0.070	0.958	0.098	0.883
Fast3R [13]	0.518	0.486	0.196	0.768	0.139	0.808
MVDUSt3R [8]	0.619	0.332	0.482	0.357	0.401	0.355
CUT3R [10]	0.534	0.558	0.075	0.943	0.111	0.883
FLARE [16]	0.791	0.358	0.142	0.797	0.357	0.579
VGGT [9]	0.230	0.678	0.052	0.969	0.052	0.968
Pi3 [12]	0.210	0.726	0.043	0.975	0.037	0.985
Ours	0.205	0.728	0.042	0.975	0.046	0.978

Table 4. **Video Depth Estimation on Sintel [1], Bonn [5] and KITTI [2].**We present the absolute relative error (Abs Rel) and threshold accuracy ($\delta < 1.25$) as the evaluation metrics with each cell colored to indicate the best and the second .

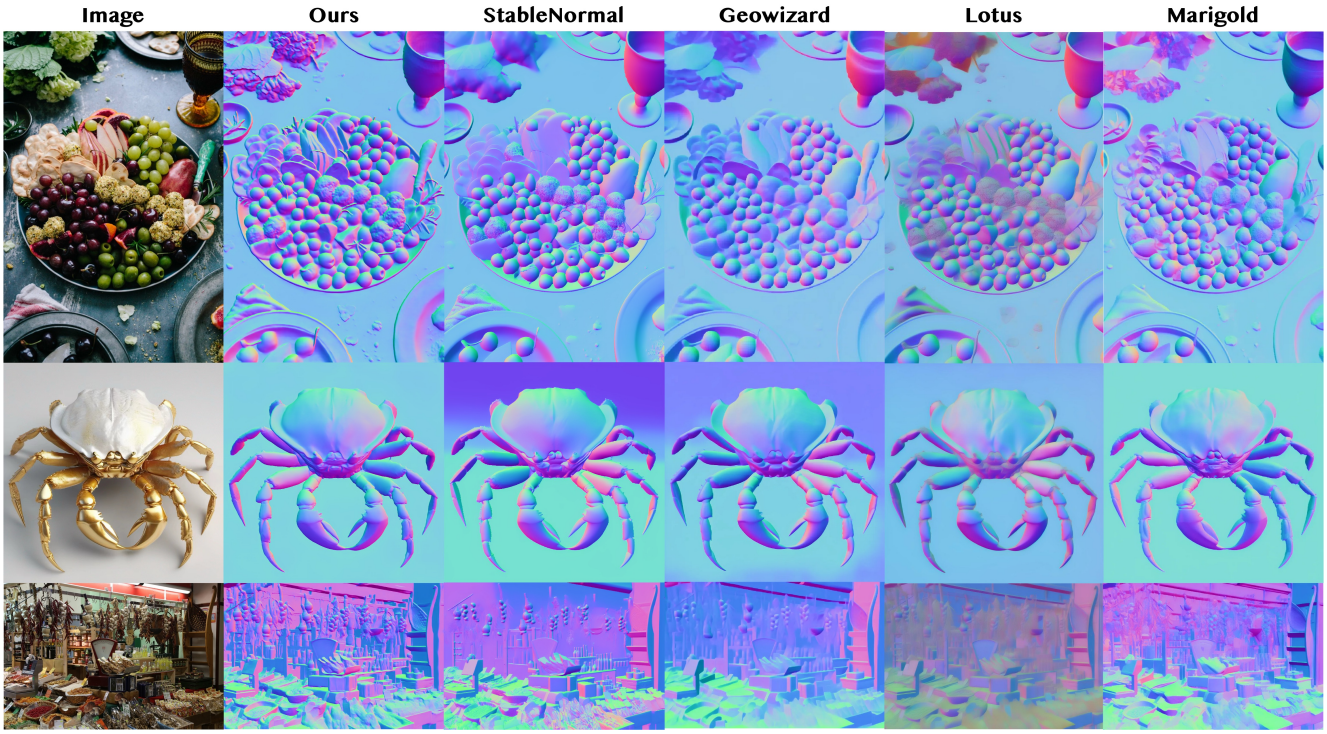

Figure 5. Normal comparison. Our method produces sharper and more accurate surface normals.

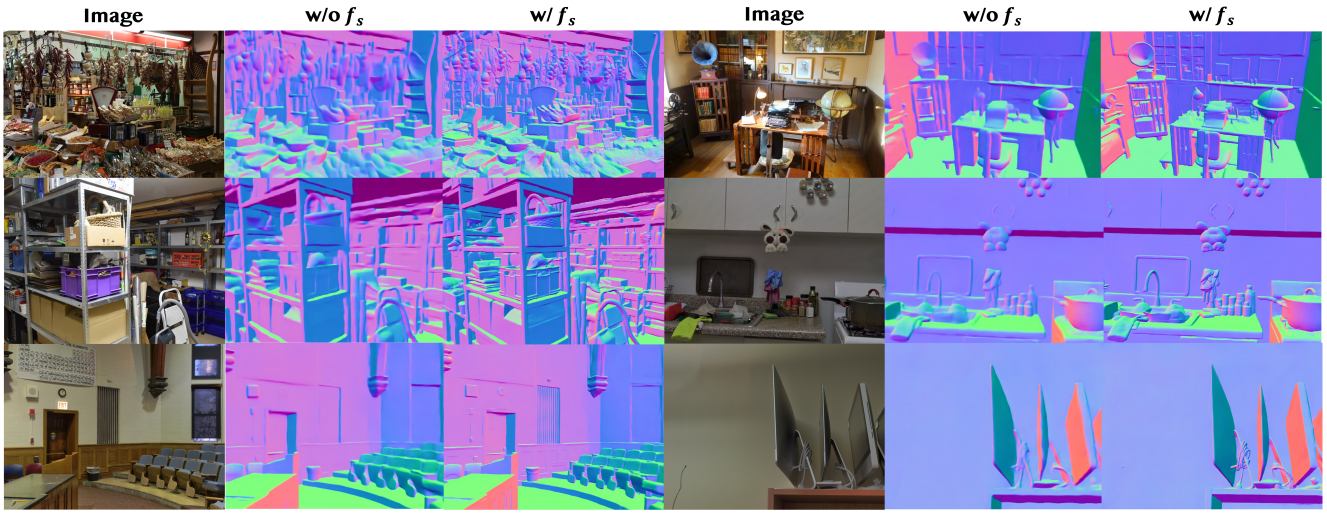

Figure 6. Additional ablation for dense semantic feature fusion. Please zoom in for better details.

References

[1] Aljaz Bozic, Pablo R. Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular RGB scene reconstruction using transformers. In *NeurIPS*, pages 1403–1414, 2021. 3

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013. 3

[3] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. *arXiv preprint arXiv:1805.01328*, 2018. 3

[4] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, pages 71–91, 2024. 3

[5] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IROS*, pages 7855–7862, 2019. 3

[6] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *ICML*, pages 18332–18346, 2022. 1

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760, 2012. 3

[8] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, pages 5283–5293, 2025. 3

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 3

[10] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 3

[11] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024. 3

[12] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 3

[13] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *CVPR*, 2025. 3

[14] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, pages 12–22, 2023. 3

[15] 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. In *ICLR*, 2025. 3

[16] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *CVPR*, 2025. 3