Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass

Supplementary Material

A. Model Scaling Effect

We investigate the effect of scaling model size by trying three model sizes for the Fusion Transformer: ViT-base, ViT-large, and ViT-huge, according to the settings in the original ViT paper [10]. The results are shown in Figure 9. This experiment demonstrates that larger model size continually benefits 3D tasks including camera pose estimations and 3D reconstruction. Note that the Fusion Transformer size used in the main text for all experiments is a ViT-base.

B. Data Scaling Effect

We study the effect of scaling the data using 4 different scales of data, 12.5%, 25%, 50%, and 100%, to train the model. The results are shown in Figure 10. The training settings for all models are kept the same except for how much data they have access to. The results demonstrate that Fast3R continually benefits from more data, suggesting Fast3R could achieve better results in the future given more data.

C. Gaussian Splatting

We qualitatively demonstrate the potential of using Fast3R's output for downstream novel view synthesis tasks. A visualization of the Gaussian Splatting generated by adopting the pipeline of InstantSplat [15] is shown in Figure 11.

D. Bundle Adjustment (via Gaussian Splatting)

While not necessary, using bundle adjustment at inference time can also improve Fast3R's performance. We show an example of bundle adjustment using Gaussian Splatting (GS-BA).

Specifically, we use InstantSplat [15] to optimize a set of Gaussians per scene, using initializations from a point cloud, and update the locations and poses in order to minimize reprojection error. We show an example of the Gaussian reconstruction in Figure 11 shows an example reconstruction on CO3D.

We can compare against ground-truth trajectories from COLMAP. We found that GS-BA significantly reduces both the pose and translation error. Table 6 quantifies this, showing over a 2.5x reduction in translation error and a 4x reduction in rotational error on the "Family" scene from Tanks and Temples, which we found to be representative. We show a visualization of the original reconstruction and the poses pre- and post-bundle-adjustment. There are only 8 scenes in the evaluation set in InstantSplat.

Method	RPE Rotation (↓)	RPE Translation (↓)
Fast3R	27.9	7.64
Fast3R w/ GS-BA	**11.0**	**1.80**

Table 6. **Pose estimation can further improve with Bundle Adjustment.** We show an example on the "Family" scene from Tanks and Temples, using InstantSplat [15].

Methods	ScanNet		ETH3D		DTU		T&T	
	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑
COLMAP-DENSE	38.0	22.5	89.8	23.2	20.8	69.3	25.7	76.4
DUSt3R 224	5.86	50.84	4.71	61.74	**2.76**	**77.32**	5.54	56.38
DUSt3R 512	**4.93**	**60.20**	**2.91**	**76.91**	3.52	69.33	**3.17**	**76.68**
Fast3R	6.27	50.27	4.68	62.68	3.92	62.60	4.43	63.95

Table 7. **Multi-view depth evaluation.** DUSt3R and Fast3R perform on par, while significantly outperforming COLMAP-DENSE.

E. Multi-view Depth Evaluation

We compare Fast3R (using the local pointmap prediction) with DUSt3R and COLMAP on multi-view depth estimation tasks and show results in Table 7.

F. More Visualizations

We show more visualizations of Fast3R's performance on indoor scenes in Figure 15. Fast3R learns the regularity of indoor rooms (square-like shapes) and demonstrates "loop closure" capabilities.

F.1. 4D Reconstruction: Qualitative Results

Because Fast3R can handle multiple frames naturally, one may wonder how well Fast3R can handle *dynamic* scenes. We qualitatively test Fast3R's 4D reconstruction ability, showing examples of dynamic aligned pointmaps at multiple time steps in Figure 16. Fast3R can be trained to achieve such results by finetuning a 16 static views checkpoint on the PointOdyssey [73] and TartanAir [61] datasets, consisting of 110 dynamic and 150 static scenes, respectively. We freeze the ViT encoder, use 224x224 image resolution, and swap in a newly-initialized global DPT head. We fine-tune the model with 15 epochs with a frame length of 16, batch size per GPU of 1, and use the same learning rate schedule as Fast3R. The process takes 45 hours to finetune on 2 Nvidia Quadro RTX A6000 GPUs.

We see that our approach produces qualitatively reasonable reconstructions with minimal changes. MonST3R [69] is a concurrent work also tackling dynamic scene recon-

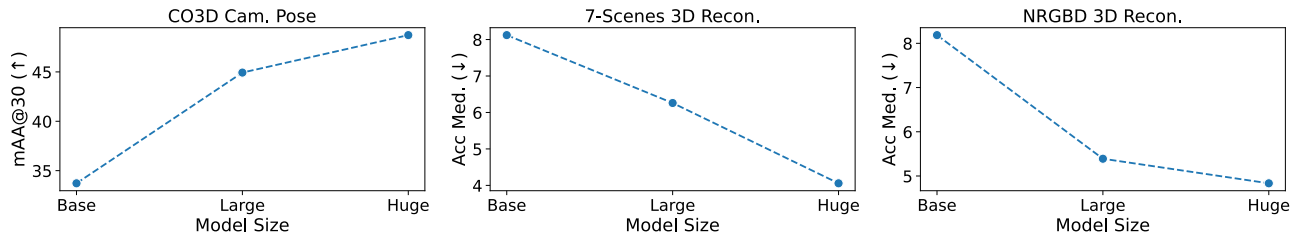

Figure 9. **Model scaling effect.** Increasing the size of the Fusion Transformer leads to better camera pose estimation (↑) and 3D reconstruction (↓). All models are trained for 60k steps (equivalent to 60 epochs; the main paper uses 100 epochs).

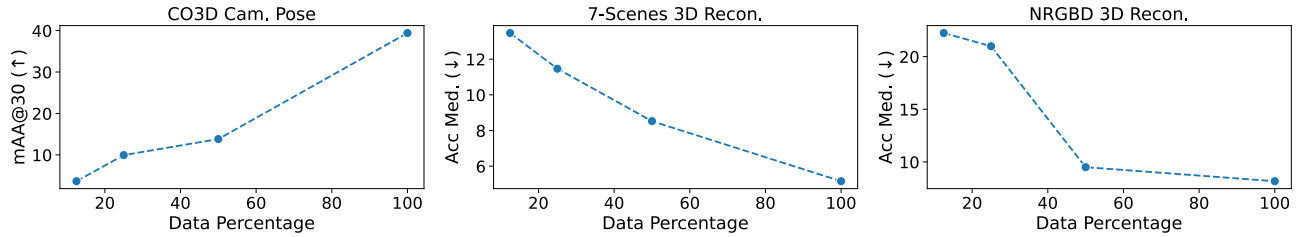

Figure 10. **Data scaling effect.** More training data leads to better camera pose estimation (↑) and 3D reconstruction (↓). All models are trained for 60k steps (equivalent to 60 epochs; the main paper uses 100 epochs).

Figure 11. **Visualization of Gaussians from unseen poses.** The frames are ordered temporally along the direction of the arrows. The middle frames show poses very different from those used for reconstruction, as is evidenced by the large areas with no Gausisans. The scene is fit from 7 images from CO3D.

struction that builds atop DUSt3R. However, like DUSt3R, it assumes a pairwise architecture and also uses a separate model to predict optical flow. We show that the same Fast3R architecture trained end-to-end with the same many-view pointmap regression (just swapping the data to dynamic scenes), can also work for 4D reconstruction. Importantly, our method remains significantly faster, opening the potential for real-time applications.

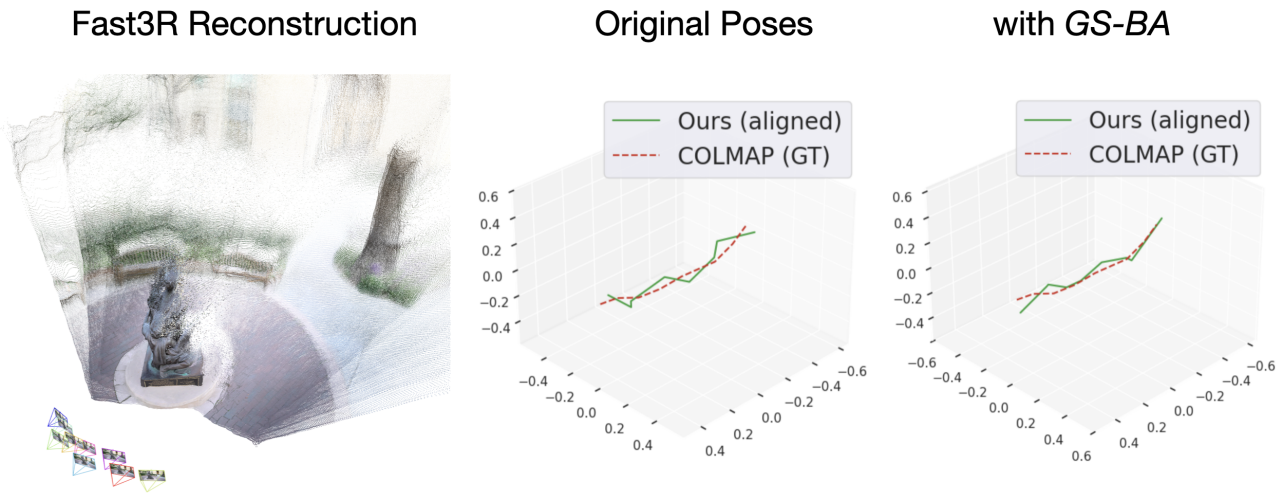

Figure 12. **Bundle adjustment further improves pose.** Left: reconstruction from Fast3R. Middle: Original poses pre-GS-BA. Right: Poses after GS-BA.

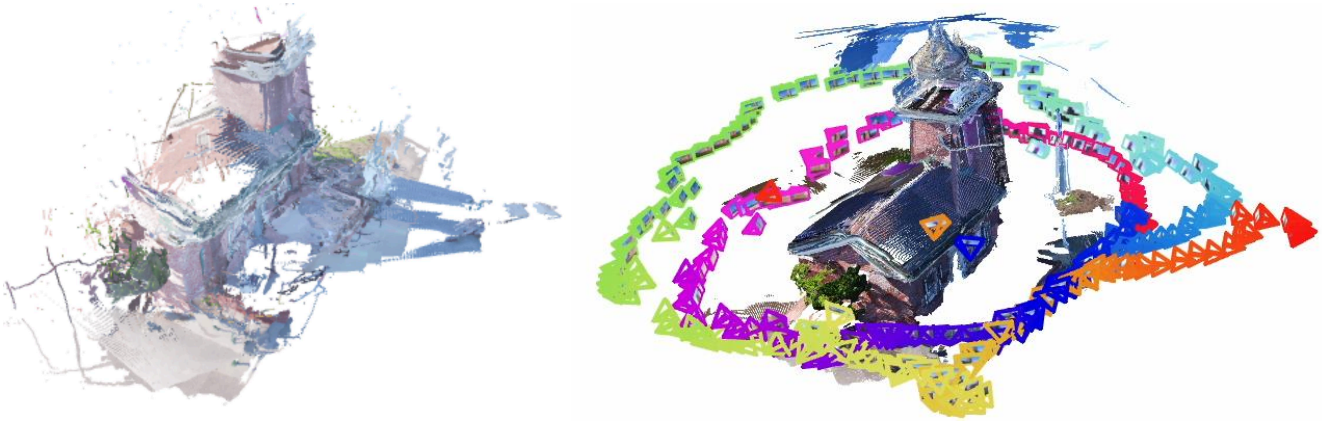

Figure 13. **Large-scale reconstruction: Spann3R vs. Fast3R on the Lighthouse scene from Tanks & Temples dataset.**

Figure 14. **Effect of using local vs. global pointmap.** Global point maps provide good anchors for locations of points while local point maps use those anchors (by aligning using ICP on the anchor points to the global point map) to provide more accurate point locations. Best viewed when zoomed in.

Figure 15. **Visualizations of results on NRGBD scenes.** Fast3R learns the regularity of indoor rooms (square-like shapes) and demonstrates loop closure capabilities.

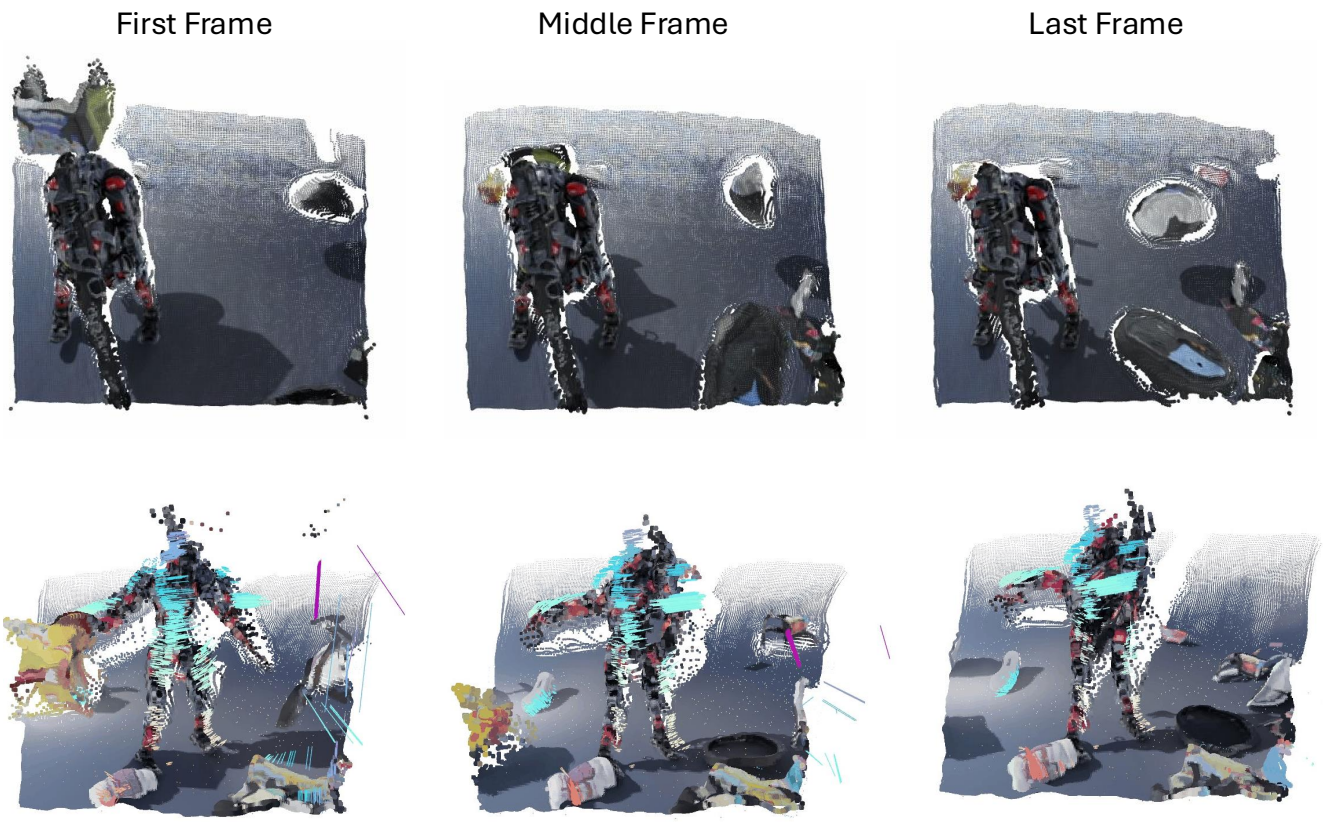

Figure 16. **Qualitative 4D reconstruction early results on dynamic scenes in PointOdyssey [73].** Results are obtained with one forward pass. The tracks are visualized using ground-truth track annotations.