MERG3R: A Divide-and-Conquer Approach to Large-Scale Neural Visual Geometry

Supplementary Material

In this supplementary material, we first present additional qualitative reconstruction results in Section 7. Section 8 showcases example reconstructions from large-scale scenes, highlighting the superior scalability of our approach. In Section 9, we provide per-scene metrics and trajectory visualizations for the Tanks & Temples dataset. Section 10 compares point-cloud reconstructions from individual subsets with our full reconstruction, illustrating the necessity of partitioning the input images. Finally, Section 12 describes implementation details of our method.

7. Reconstruction Gallery

We present additional qualitative reconstruction results on various scenes from Tanks & Temples [16], 7-Scenes [29], and UrbanScene3D [18] in Fig. 8. MERG3R consistently produces robust and detailed reconstructions across both large indoor and outdoor environments. For visualization clarity, we downsample the point clouds and filter points by confidence. We also demonstrate the performance of MERG3R on dynamical scenes from internet videos in Fig. 9.

8. Results on Large Scale Dataset

To demonstrate our method's scalability beyond 1,000 images, we evaluate it on two long sequences from Zip-NeRF [4] with approximately 1500 images and 1900 images respectively, as shown in Fig. 10. We use the raw Zip-NeRF images directly, which depict complex real-world environments spanning both indoor and outdoor scenes. Compared to the datasets used in the main text, adjacent views in these sequences exhibit substantially less overlap, and the indoor subsets in particular feature challenging spatial layouts with significant geometric complexity. Because π^3 is unable to process such large numbers of images, we uniformly subsample each dataset to 500 images for a fair comparison, and set our method's subset size to 500 images accordingly. As illustrated, when presented with large and complex scenes, π^3 fails to reconstruct the full environment, losing significant structural details and geometric consistency. In the second dataset in particular, π^3 incorrectly merges two distinct rooms, resulting in severe overlapping artifacts. In addition, we tested π^3 with our proposed bundle-adjustment step; however, because the initial 3D prior is already severely degraded, the refinement yields minimal visible improvement.

For large-scale datasets with more than 1,000 images, the simple interleaving scheme may select images that are too distant in DINO feature space, resulting in subsets that contain disjoint views and degrade local reconstruction. To address this, after forming the pseudo-video, we refine the interleaving step by searching forward along the sequence for the next image whose (precomputed) DINO similarity to the previously selected image falls within the range $0.5m$ to $0.95m$, where m is the median similarity score with respect to the last chosen image. Once such an image is found, it becomes the new reference point, and the process repeats. This produces the refined sequence $\tilde{P}$ described in Sec. 3.2. We then apply the sliding-window grouping to form subsets. This procedure avoids selecting images that are overly dissimilar or spatially far apart, ensuring that each subset remains locally coherent.

9. Additional Results

We provide the detailed per-scene comparison on Tanks & Temples for MERG3R and the baselines. Shown in Table 10 and Table 11, we outperformed the base model and the other baselines on almost every scene.

In addition, we present qualitative trajectory visualizations for all scenes in the Tanks and Temples dataset (Fig. 14). Each row corresponds to one scene, and each column corresponds to a model. Across scenes, our method produces trajectories that closely follow the ground truth and consistently match or surpass the accuracy of all baselines. These results show that our improvements hold not only in aggregate metrics but also in individual scenes

10. Subset Reconstruction Results

In Fig. 11, we visualize the 3D reconstruction results for each subset, along with the final merged reconstruction. Each subset captures only a partial portion of the scene, highlighting the need to divide truly large-scale environments into manageable subsets and subsequently merge them to obtain a globally consistent reconstruction.

11. Splitting Robustness Analysis

We further investigate the robustness of the DINO feature similarity used for sequence construction. We conduct a perturbation experiment by randomly inserting outlier frames from other scenes into a sequence and quantitatively analyzing the resulting DINO similarities. Specifically, we plot histograms of similarity scores for valid–valid pairs and valid–distractor pairs. As shown in Figure 13, the two distributions are clearly separated, indicating that DINO similarity reliably distinguishes outlier frames and supports robust pseudo-video rearrangement. Figure 12 illustrates an example sequence with randomly inserted frames alongside its rearranged counterpart, where the outlier frames are

effectively pushed to the end of the sequence.

12. Implementation Details

We report all hyperparameters used in our experiments. For bundle adjustment, we set the initial learning rate to 3×10^{-3} and optimize for 300 iterations using a cosine-annealing schedule. For subset alignment, we retain only points whose confidence exceeds the 70th percentile. For tracking, we perform direct matching across at most five frames, extract up to 4,096 keypoints per image, and apply a reprojection error threshold of 8 pixels during geometric verification.

13. Limitations

Although MERG3R demonstrates strong scalability and improved accuracy on existing benchmarks, several limitations remain. First, the method may degrade when viewpoint changes between input images are extremely drastic. In such cases, feature correspondences become sparse or unreliable, which can lead to fragmented reconstruction or unstable pose estimation. This issue is particularly pronounced in large-scale indoor scenes with wide baselines or severe occlusions. MERG3R typically performs better on outdoor scenes. Second, the DINO similarity–based splitting heuristic assumes that feature similarity provides a reliable proxy for geometric consistency. When scenes contain large textureless regions or overly similar images from different places, DINO embeddings may become less discriminative. This can result in suboptimal clustering, where geometrically related images are split across different groups or unrelated images are merged together. Consequently, local reconstructions may lack sufficient overlap, negatively affecting downstream global alignment.

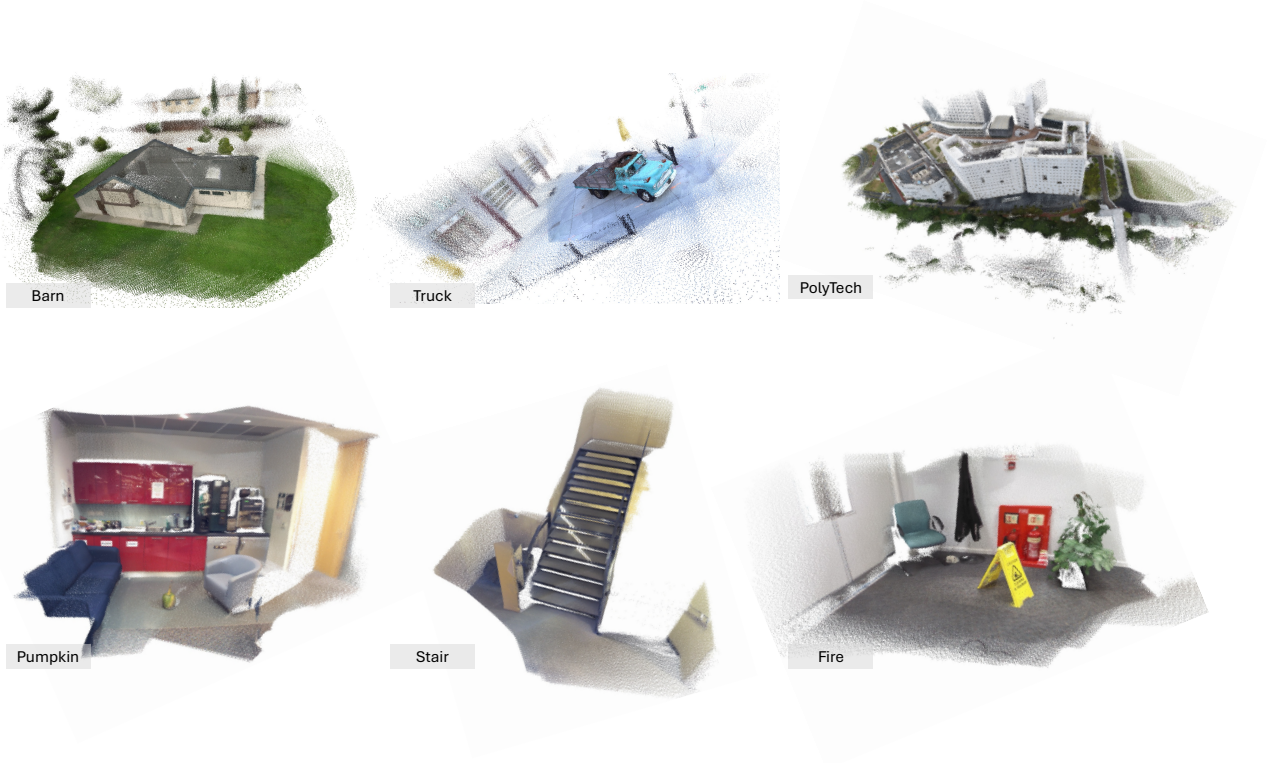


Figure 8. **Qualitative examples of 3D reconstructions on various indoor and outdoor scenes from Tanks & Temples [16], 7-Scenes [29] and UrbanScene3D [18].** MERG3R produces high-quality, detailed reconstructions that preserve fine geometric structure and maintain global consistency.

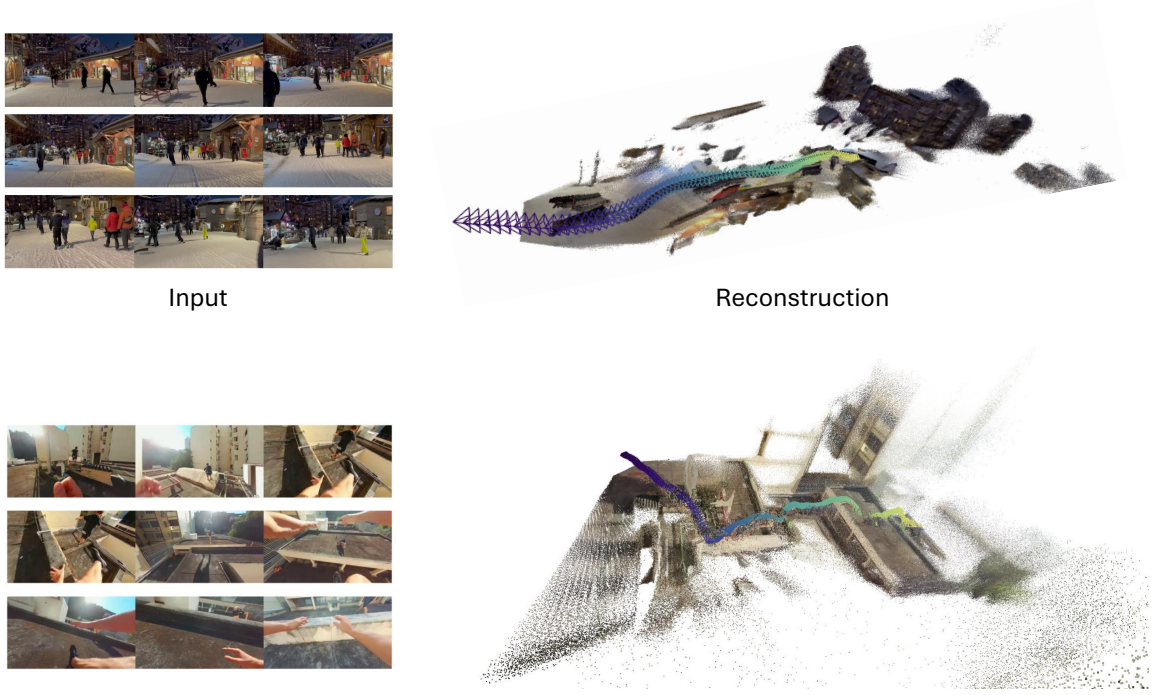


Figure 9. Qualitative examples of 3D reconstructions on dynamical scene from internet videos.

Scene	ATE ↓							RRE↓							RTE↓						
	CUT3R [36]	TT3R [6]	MAST3R-SfM [10]	VGGT* [35]	Pi3 [38]	VGGT* + Ours	Pi3 + Ours	CUT3R [36]	TT3R [6]	MAST3R-SfM [10]	VGGT* [35]	Pi3 [38]	VGGT* + Ours	Pi3 + Ours	CUT3R [36]	TT3R [6]	MAST3R-SfM [10]	VGGT* [35]	Pi3 [38]	VGGT* + Ours	Pi3 + Ours
Barn	1.246	0.827	0.093	0.074	0.048	0.035	**0.030**	0.793	0.773	0.268	0.524	0.234	0.440	**0.179**	0.058	0.064	0.009	0.038	0.015	0.026	**0.006**
Caterpillar	0.974	0.277	0.036	0.051	0.035	**0.021**	0.023	0.987	0.931	0.658	0.299	0.195	0.104	**0.097**	0.069	0.058	0.012	0.028	0.018	0.007	**0.005**
Church	2.288	1.176	0.983	2.934	0.358	3.002	**0.330**	1.684	5.337	1.638	7.378	0.396	8.326	**0.380**	0.124	0.172	0.101	0.274	0.057	0.298	**0.046**
Ignatius	0.663	0.188	0.024	0.043	0.035	**0.020**	0.021	1.043	0.785	0.197	0.210	0.156	0.144	**0.117**	0.091	0.069	**0.008**	0.025	0.020	0.010	**0.008**
Meeting Room	1.819	0.646	0.046	0.071	**0.035**	0.039	0.037	0.987	0.884	**0.163**	0.340	0.201	0.230	0.166	0.102	0.121	**0.008**	0.036	0.021	0.016	0.009
Truck	0.637	0.149	0.032	0.035	0.031	**0.016**	0.018	0.940	0.940	0.204	0.234	0.194	0.135	**0.128**	0.076	0.059	**0.006**	0.023	0.019	**0.006**	**0.006**

Table 10. Per-scene camera pose evaluation results (ATE, RRE, RTE) across methods for the Tanks & Temples dataset.

Scene	RRA@30 ↑							RTA@30↑							AUC@30↑						
	CUT3R [36]	TT3R [6]	MAST3R-SfM [10]	VGGT* [35]	Pi3 [38]	VGGT* + Ours	Pi3 + Ours	CUT3R [36]	TT3R [6]	MAST3R-SfM [10]	VGGT* [35]	Pi3 [38]	VGGT* + Ours	Pi3 + Ours	CUT3R [36]	TT3R [6]	MAST3R-SfM [10]	VGGT* [35]	Pi3 [38]	VGGT* + Ours	Pi3 + Ours
Barn	77.64	**100**	**100**	**100**	**100**	**100**	**100**	88.85	89.71	**99.99**	99.88	99.96	99.93	**99.99**	40.33	63.90	89.23	93.86	**95.82**	95.44	**95.82**
Caterpillar	87.97	**100**	99.48	**100**	**100**	**100**	**100**	89.51	98.66	99.70	99.91	99.77	99.99	**100**	48.15	90.89	92.75	95.36	95.85	96.48	**96.5**
Church	44.33	80.38	79.12	56.82	**100**	58.21	**100**	52.13	69.56	78.01	54.60	95.03	61.49	**95.46**	21.65	46.04	78.01	34.40	86.18	51.79	**87.33**
Ignatius	99.44	**100**	**100**	**100**	**100**	**100**	**100**	1.78	99.38	99.99	99.97	99.97	99.99	**100**	61.09	88.08	**99.99**	95.57	95.98	**96.47**	96.41
Meeting Room	95.09	**100**	**100**	**100**	**100**	**100**	**100**	69.00	92.98	99.95	99.83	**99.96**	99.94	99.95	39.76	68.96	94.73	93.51	**95.72**	95.32	95.27
Truck	99.77	**100**	**100**	**100**	**100**	**100**	**100**	93.76	99.35	**100**	99.99	99.98	**100**	**100**	60.91	88.25	**100**	96.05	95.97	96.48	92.28

Table 11. Per-scene camera pose evaluation results (RRA, RTA, AUC) across methods for the Tanks & Temples dataset.

Figure 10. **Qualitative comparison of 3D reconstructions on large scale dataset.** MERG3R achieves consistently better performance on the Zip-NeRF [4] scenes (Berlin). The input sequence is the original ordering before splitting.

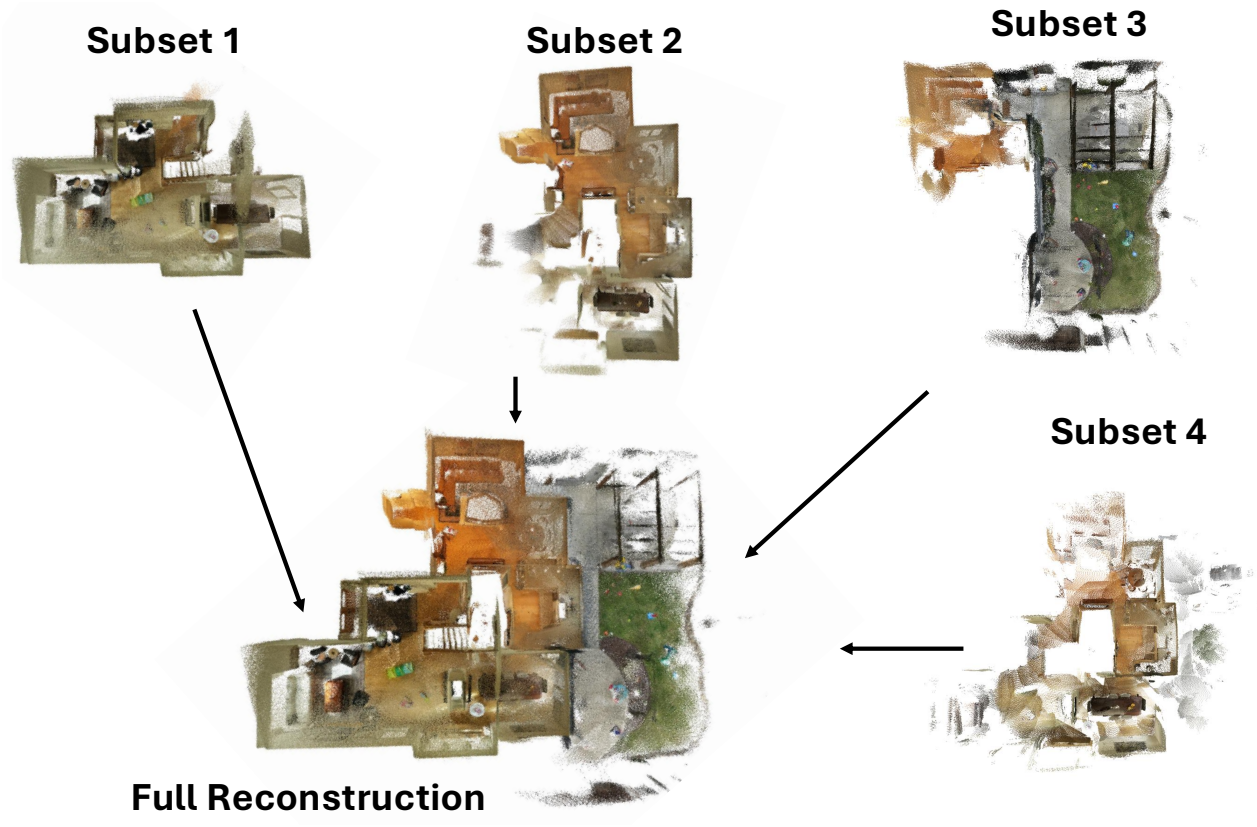


Figure 11. **Visualization results for each subset.** Our full reconstruction faithfully recovers the entire house, while each individual subset captures only partial and fragmented portions of the scene.

Figure 12. A sample sequence from Zip-NeRF dataset with randomly inserted frames from other scenes (above). The reordered sequence (bottom) places all inserted frames at the end.

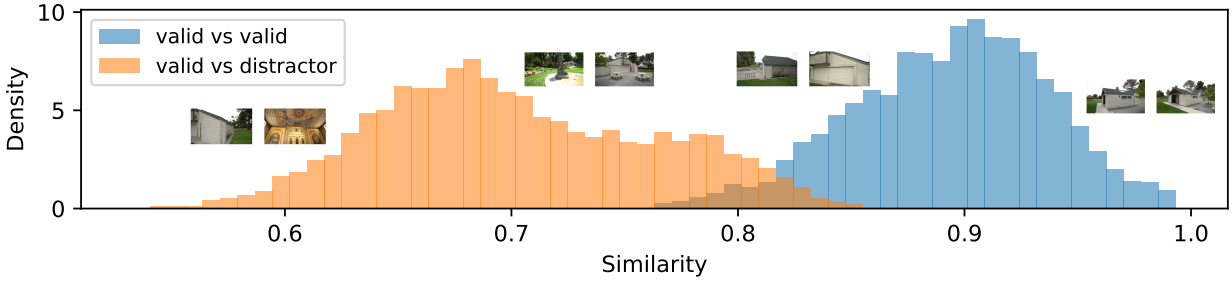
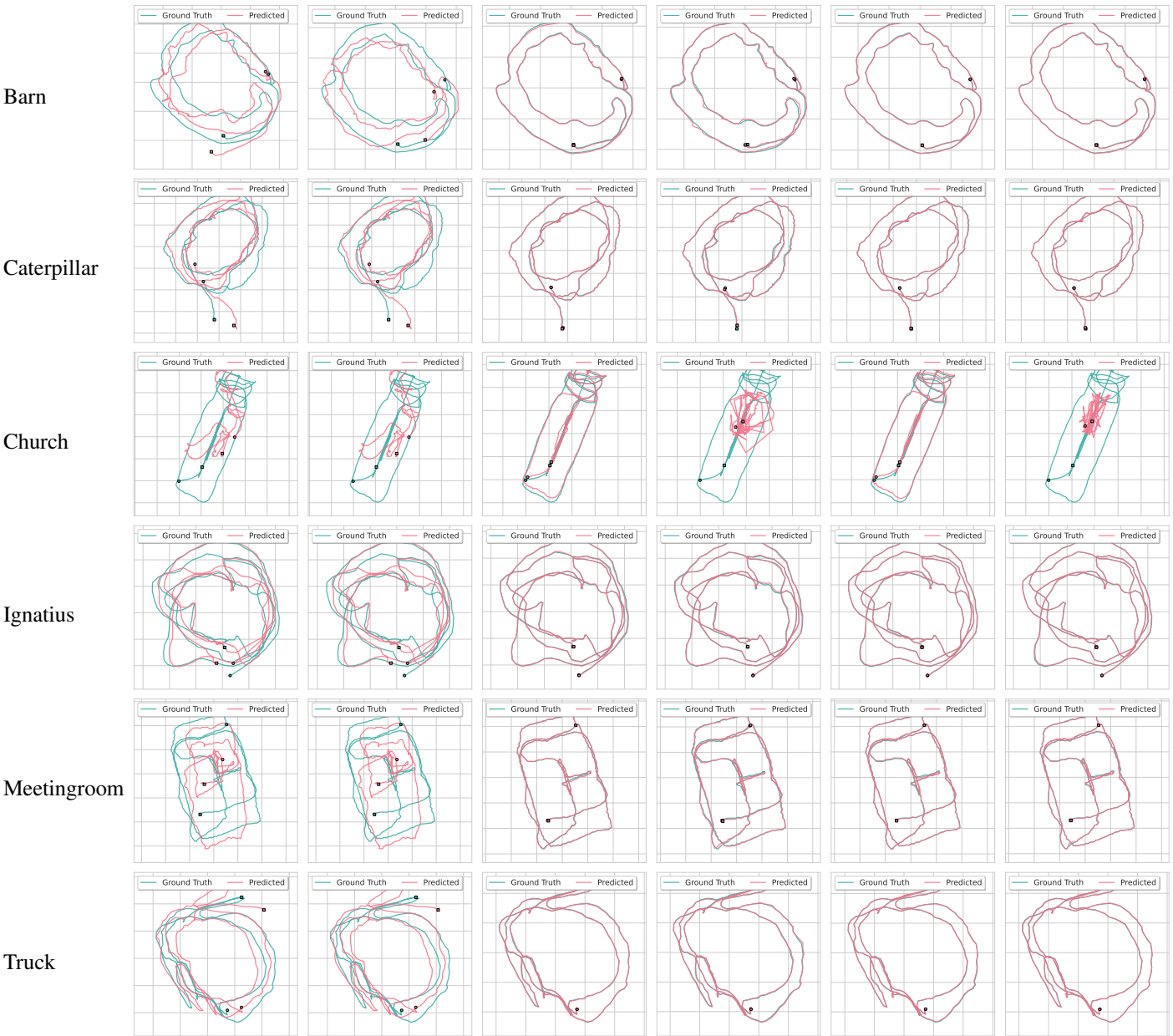


Figure 13. Histograms of two pair types (valid–valid vs. valid–distractor). The clear separation between the distributions demonstrates the robustness of DINO similarity for detecting outlier frames.



(a) TTT3R (b) CUT3R (c) π^3 (d) VGGT (e) ours+π^3 ours+VGGT

Figure 14. Qualitative pose estimation results for all scenes in the Tanks & Temples dataset.