
RaySt3R: Predicting Novel Depth Maps for Zero-Shot Object Completion

1 Technical Appendices

1.1 Quantitative analysis

Here we provide additional quantitative results of RaySt3R evaluated against the baselines.

1.1.1 Distribution of Chamfer Distance

Figure 1 shows the distributions of chamfer distance, visualized with histograms. The results suggest RaySt3R consistently produces more accurate predictions.

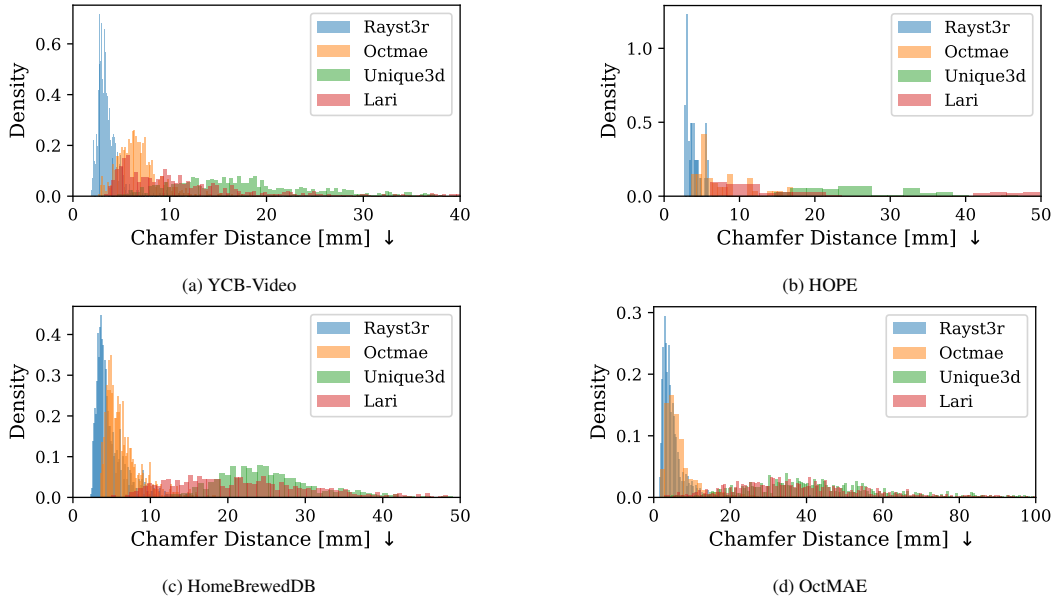


Figure 1: RaySt3R chamfer distance distribution compared against several baselines. The results suggest RaySt3R consistently produces a more favorable distribution, with higher density at smaller chamfer distance.

1.1.2 Standard Deviation

Table 1 shows the standard deviation for the baselines with public checkpoints. The results suggest RaySt3R not only produces the most competitive results on average, but also consistently does so with the smallest standard deviation.

Table 1: Standard deviations (STD) of Chamfer Distance (CD) [mm] and F1-Score@10mm (F1) for multi-object scene completion methods evaluated on synthetic (OctMAE [3]) and real datasets (YCB-Video [11], HOPE [8], and HomebrewedDB [4]). Metric values are provided for context, standard deviations are denoted as subscripts. The results suggest RaySt3R consistently achieves lower metrics and STD.

Method	Synthetic		Real					
	OctMAE [3]		YCB-Video [11]		HB [4]		HOPE [8]	
	CD↓	F1↑	CD↓	F1↑	CD↓	F1↑	CD↓	F1↑
OctMAE [3]	6.48 \pm 28.52	0.839 \pm 0.104	6.40 \pm 2.269	0.800 \pm 0.066	6.14 \pm 2.394	0.819 \pm 0.074	6.97 \pm 3.446	0.803 \pm 0.082
LaRI [5]	39.22 \pm 23.89	0.283 \pm 0.143	11.41 \pm 7.422	0.658 \pm 0.195	22.23 \pm 9.650	0.414 \pm 0.146	18.64 \pm 14.23	0.528 \pm 0.202
Unique3D [9]	44.62 \pm 26.99	0.244 \pm 0.122	17.56 \pm 7.180	0.468 \pm 0.144	25.41 \pm 6.951	0.329 \pm 0.074	26.37 \pm 8.313	0.322 \pm 0.075
TRELLIS (w/ mask) [10]	61.74 \pm 69.06	0.227 \pm 0.135	22.94 \pm 12.13	0.443 \pm 0.207	35.29 \pm 62.04	0.354 \pm 0.154	20.25 \pm 10.75	0.443 \pm 0.155
TRELLIS (w/o mask) [10]	65.74 \pm 104.9	0.225 \pm 0.135	31.74 \pm 53.39	0.338 \pm 0.141	30.71 \pm 73.71	0.348 \pm 0.092	22.05 \pm 11.47	0.416 \pm 0.121
SceneComplete [1]	81.57 \pm 219.9	0.289 \pm 0.135	96.63 \pm 100.6	0.359 \pm 0.120	85.81 \pm 188.6	0.416 \pm 0.168	N/A	N/A
RaySt3R (ours)	5.21 \pm 7.836	0.893 \pm 0.085	3.56 \pm 1.194	0.930 \pm 0.038	4.75 \pm 1.976	0.889 \pm 0.070	3.92 \pm 0.9229	0.926 \pm 0.043

1.1.3 Per-Scene Results

Table 2 shows the results on each scene individually. The results suggest RaySt3R outperforms the baselines in all scenes.

Table 2: RaySt3R evaluated on each real-world scene individually. The results suggest RaySt3R outperforms all baselines on all scenes.

Dataset	Scene	Unique3D [9]		TRELLIS (w/ mask) [10]		TRELLIS (w/o mask) [10]		LaRI [5]		OctMAE [3]		RaySt3R (ours)	
		CD ↓	F1 ↑	CD ↓	F1 ↑	CD ↓	F1 ↑	CD ↓	F1 ↑	CD ↓	F1 ↑	CD ↓	F1 ↑
HB [4]	000001	35.53	0.24	68.57	0.18	62.15	0.21	30.70	0.31	7.59	0.77	5.99	0.85
	000002	33.39	0.28	62.81	0.18	55.64	0.20	31.27	0.31	7.32	0.79	5.36	0.88
	000003	24.39	0.34	33.12	0.35	26.45	0.31	26.59	0.34	4.74	0.88	3.43	0.94
	000004	30.69	0.30	72.98	0.24	29.18	0.31	27.95	0.34	4.99	0.87	3.40	0.94
	000005	22.06	0.37	23.24	0.41	23.73	0.39	18.66	0.43	5.65	0.84	4.27	0.90
	000006	25.22	0.30	34.58	0.31	23.25	0.37	16.50	0.51	6.19	0.82	4.19	0.91
	000007	24.72	0.33	23.01	0.43	24.87	0.35	23.55	0.36	6.79	0.81	4.66	0.89
	000008	24.35	0.33	20.59	0.46	22.67	0.40	12.61	0.60	6.32	0.81	4.15	0.91
	000009	22.36	0.36	50.75	0.30	20.26	0.39	17.16	0.50	5.32	0.86	3.56	0.93
	000010	25.22	0.31	26.99	0.37	51.21	0.34	28.26	0.32	5.19	0.86	3.95	0.92
	000011	26.58	0.32	21.30	0.42	25.45	0.35	18.11	0.47	5.04	0.87	3.97	0.92
	000012	17.96	0.40	24.56	0.40	18.55	0.46	19.75	0.43	9.75	0.71	7.69	0.78
	000013	19.16	0.39	20.47	0.44	19.58	0.43	20.30	0.42	9.63	0.72	7.40	0.78
HOPE [8]	000001	24.85	0.34	12.88	0.55	19.34	0.43	10.72	0.63	5.65	0.83	3.88	0.93
	000002	31.93	0.29	15.33	0.51	14.34	0.49	6.82	0.79	4.09	0.92	2.74	0.98
	000003	35.12	0.25	41.98	0.20	42.50	0.24	47.11	0.20	5.61	0.84	3.37	0.94
	000004	31.79	0.25	24.08	0.36	26.36	0.36	19.83	0.43	11.47	0.66	5.54	0.87
	000005	20.11	0.36	16.96	0.46	21.11	0.41	16.02	0.45	7.56	0.78	3.89	0.93
	000006	19.46	0.40	11.40	0.58	14.94	0.50	7.94	0.74	6.23	0.79	3.15	0.95
	000007	22.83	0.33	14.42	0.51	18.29	0.43	13.31	0.57	8.41	0.73	5.39	0.84
	000008	22.82	0.32	19.88	0.41	14.80	0.47	9.38	0.66	5.23	0.87	3.64	0.96
	000009	38.66	0.24	35.41	0.23	37.06	0.26	44.74	0.18	4.60	0.87	3.18	0.95
	000010	16.17	0.45	10.20	0.64	11.78	0.56	10.54	0.62	15.63	0.67	4.48	0.89
YCB-Video [11]	000048	15.17	0.49	15.96	0.54	26.39	0.38	10.90	0.65	9.07	0.74	4.29	0.91
	000049	17.50	0.45	25.13	0.42	27.45	0.34	9.56	0.71	9.15	0.74	5.04	0.89
	000050	20.22	0.43	22.15	0.42	55.73	0.28	17.95	0.46	9.07	0.72	4.25	0.90
	000051	21.93	0.37	35.79	0.28	39.82	0.26	8.71	0.69	6.66	0.77	2.82	0.96
	000052	8.92	0.71	5.79	0.86	18.34	0.50	9.12	0.71	5.61	0.83	2.98	0.95
	000053	14.87	0.48	27.61	0.34	26.73	0.32	4.95	0.88	4.29	0.89	2.43	0.98
	000054	18.07	0.41	16.30	0.47	23.96	0.40	13.32	0.53	6.96	0.78	3.93	0.91
	000055	31.03	0.27	38.87	0.26	44.97	0.23	9.02	0.70	6.64	0.77	3.68	0.92
	000056	11.95	0.59	14.29	0.56	20.55	0.43	7.55	0.79	6.27	0.80	3.34	0.94
	000057	24.99	0.34	33.36	0.32	47.29	0.21	28.94	0.31	5.49	0.84	3.71	0.92
	000058	12.21	0.55	19.72	0.43	25.45	0.32	7.79	0.77	6.49	0.79	3.13	0.94
	000059	13.94	0.53	20.37	0.42	24.17	0.38	9.08	0.70	5.60	0.84	3.10	0.94

1.1.4 Single-Object Evaluation

We target cluttered scenes, as is common in robot applications. However, several of the baselines are trained only on single-object scenes. We evaluate RaySt3R and baselines in a single-object setting from the YCB-Video dataset [11]. For each frame, we evaluate on the largest object with no occlusions, which is often assumed by single-view RGB-based object reconstruction methods such as TRELLIS [10]. Table 3 shows the results of the experiments, and suggests RaySt3R outperforms the baselines in real-world single-object scenes.

Table 3: YCB-V single-object evaluation

Method	CD ↓	F1 ↑
TRELLIS [10]	22.37	0.56
LaRI [5]	13.75	0.51
OctMAE [3]	8.34	0.74
Ours (RaySt3R)	4.06	0.91

Table 4: Completeness results

Method	OctMAE	YCB-Video	HomebrewedDB	HOPE
Unique3D [9]	47.29	20.27	27.72	30.18
TRELLIS with mask [10]	48.12	22.43	26.22	20.74
TRELLIS [10]	45.34	28.20	23.23	20.28
LaRI [5]	41.21	11.74	24.60	19.66
OctMAE [3]	20.27	8.49	8.69	7.69
Ours (RaySt3R)	7.80	3.67	5.93	4.66

1.1.5 Additional Evaluation Metrics

For the evaluations in the main paper we follow the protocol from prior works [3]. Here we provide additional evaluation metrics for an even more thorough evaluation. The reported Chamfer distance up to now is the average of accuracy and completeness, Table 4 and Table 5 show completeness and accuracy evaluated separately. Furthermore, Table 6 shows the root mean square error (RMSE) averaged between accuracy and completeness. The results suggest RaySt3R outperforms the baselines in accuracy, completeness, and RMSE.

1.2 Runtime Analysis

Table 7 compares the inference speed between methods. The results suggest that OctMAE and LaRI are the fastest available methods, followed by RaySt3R. RaySt3R is slower than OctMAE and LaRI mostly as a result of querying 22 novel views instead of a single view (LaRI) and a volumetric approach (OctMAE). RaySt3R proves to be considerably faster than TRELLIS and SceneComplete.

1.3 Ablations

1.3.1 Baseline ground truth alignment

Unique3D [9], LaRI [5] and TRELLIS [10] predict shapes in a canonical space. We fit their predictions to the ground truth for evaluation. Note that we do not perform an alignment step for RaySt3R, by allowing the baselines access to the ground truth points, we give them the benefit of the doubt. All other baselines predict points directly in the coordinate frame of the input camera.

To align points from a canonical frame, we first scale the prediction such that its oriented bounding-box (OBB) extent matches that of the ground truth points. Second, we align the prediction’s OBB center with that of the ground truth points. Third, we do two passes of brute force rotational orientation search in a grid search manner to minimize the chamfer distance. Finally, we apply iterative closest point (ICP) [7] for the local alignment of the predictions. In the following ablations, we randomly sample 30 scenes from each real-world dataset for computational tractability.

Rotational grid search resolution: We ablate the number of steps during the first brute-force search for the optimal rotational alignment with the ground truth point cloud. Figure 2 shows the results. The average chamfer distance declines with an increasing number of steps. A value of 20 steps marks a point beyond which further increases yield diminishing improvements in chamfer distance.

Scale grid search: We use PCA to estimate the parameters of the OBB, which may not be the optimal scale. For this ablation, we first scale the predictions using the OBB as described above. Then, we scale the object with a constant and apply the rotational alignment grid search with 20 steps.

Table 5: Accuracy results

Method	OctMAE	YCB-Video	HomebrewedDB	HOPE
Unique3D [9]	45.87	14.86	23.09	22.57
TRELLIS with mask [10]	84.11	23.46	42.71	19.77
TRELLIS [10]	94.42	35.27	38.19	23.82
LaRI [5]	42.61	11.08	19.87	17.62
OctMAE [3]	5.93	5.06	4.28	7.21
Ours (RaySt3R)	2.63	3.45	3.58	3.19

Table 6: RMSE results

Method	OctMAE [3]	YCB-Video [11]	HomebrewedDB [4]	HOPE [8]
Unique3D [9]	62.23	24.16	34.27	36.12
TRELLIS with mask [10]	84.94	31.55	45.55	27.34
TRELLIS [10]	88.32	42.26	40.32	30.53
LaRI [5]	55.88	15.72	30.54	25.26
OctMAE [3]	21.72	11.31	10.30	14.52
Ours (RaySt3R)	9.33	6.17	7.85	6.37

We evaluate scales in the range from 0.65 to 1.35 with a step size of 0.05, the results are shown in Figure 3.

1.3.2 Mask ablation

We have demonstrated RaySt3R outperforms the baselines in the real world, even with imperfect masks. It remains unclear how sensitive our method is to the input mask. Figure 4 shows a sensitivity analysis on false positive and false negative entries in the mask, evaluated on the OctMAE dataset. False positives are pixels in the background marked as foreground, while false negatives are pixels in the foreground marked as background. We study noise in the range of 0 to 0.2, meaning 0% to 20% of the foreground and background pixels are incorrect. The results suggest RaySt3R is more robust to false positives, likely thanks to the confidence and mask filtering steps. The performance degrades substantially with a large number of false negatives.

1.3.3 Confidence Ablation

Table 8 shows the Chamfer Distance (CD) in mm for varying confidence threshold. The results suggest each dataset has a different optimal confidence threshold. Overall, RaySt3R is not sensitive to the choice of confidence threshold and consistently outperforms the baselines under all confidence thresholds.

1.3.4 Hyper Parameter Ablation

We study RaySt3R’s performance with a single set of hyperparameters instead of two sets for synthetic and real datasets. The hyperparameters varied are only the settings for sampling novel views, we set a larger radius for the OctMAE dataset. Increasing the radius reduces spatial resolution, but may help

Table 7: Runtime comparison on $1 \times \text{RTX } 4090$. *Per object, as reported by original paper.

Method	Runtime [ms]
OctMAE [3]	56
LaRI [5]	283
TRELLIS [10]	11,409
SceneComplete* [1]	$\sim 30,000$
Ours (RaySt3R) w/ Torch Attn	1200
Ours (RaySt3R) w/ Flash Attn	568

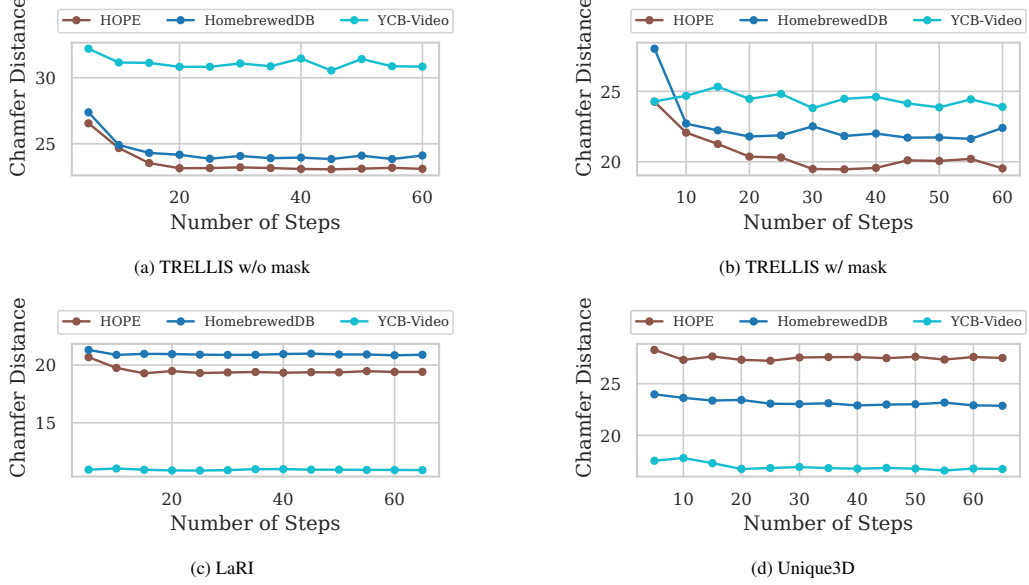


Figure 2: **Rotation ablation:** Chamfer Distance averaged over the evaluation split of HomebrewedDB [4], HOPE [8], and YCB-Video [11] with different numbers of steps during the rotational alignment grid search.

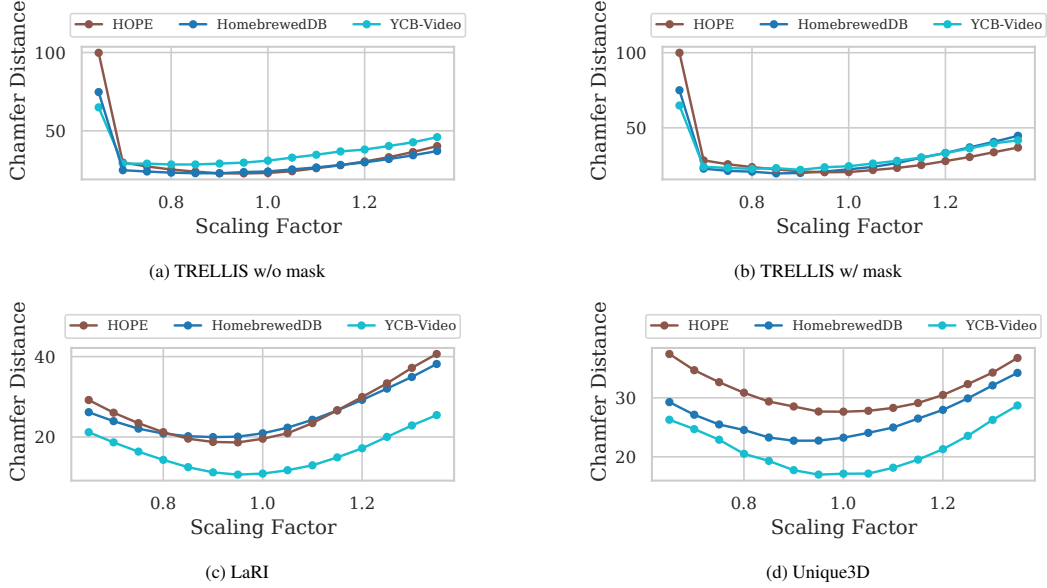


Figure 3: **Scaling ablation:** Chamfer distance averaged over the evaluation split of HomebrewedDB [4], HOPE [8], and YCB-Video [11] for different initial scaling factors.

capture the full shape of larger objects. The results in Table 9 suggest that RaySt3R outperforms the baselines even with a single set of hyperparameters..

1.4 Implementation details

1.4.1 Evaluation metrics

Consistent with prior work [3], we adopt the Chamfer Distance (CD) and F1-Score (F1) metrics for evaluation.

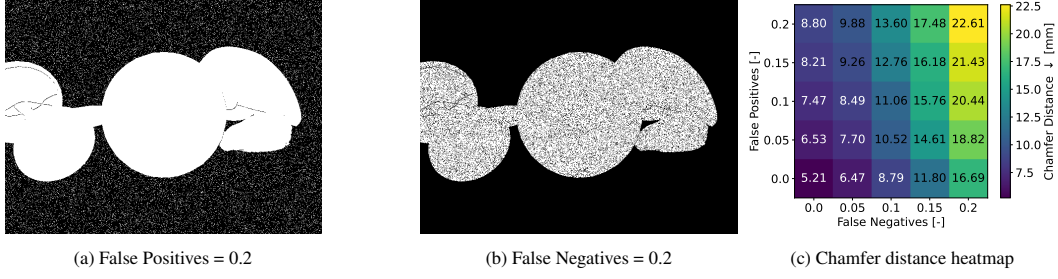


Figure 4: RaySt3R input mask noise ablation. The results suggest RaySt3R is more robust against false positive noise, as compared to false negative noise. Colors in heatmap altered for visual clarity only.

Table 8: Confidence ablation, Chamfer Distance in mm.

Method	Conf	YCB-Video [11]	HomebrewedDB [4]	HOPE [8]
Ours (RaySt3R)	0.00	3.59	4.49	3.99
Ours (RaySt3R)	1.25	3.53	4.50	3.99
Ours (RaySt3R)	2.50	3.55	4.58	3.98
Ours (RaySt3R)	3.75	3.56	4.67	3.94
Ours (RaySt3R)	5.00	3.56	4.75	3.92
Ours (RaySt3R)	6.25	3.57	4.85	3.94
Ours (RaySt3R)	7.50	3.58	4.95	3.97
Ours (RaySt3R)	8.75	3.62	5.11	4.05
Ours (RaySt3R)	10.00	3.65	5.27	4.16
OctMAE	N/A	6.40	6.14	6.97

Chamfer Distance (CD). The chamfer distance $CD(Q, Q^{gt})$ between a predicted point cloud Q and a ground truth point cloud Q^{gt} is defined as the average of two asymmetric terms:

The accuracy (forward chamfer term) measures how well the predicted points approximate the ground truth:

$$A = \frac{1}{|Q|} \sum_{q_{pd} \in Q} \min_{q_{gt} \in Q^{gt}} \|q_{pd} - q_{gt}\| \quad (1)$$

The completeness (backward chamfer term) measures how well the ground truth is covered by the predicted points:

$$C = \frac{1}{|Q^{gt}|} \sum_{q_{gt} \in Q^{gt}} \min_{q_{pd} \in Q} \|q_{gt} - q_{pd}\| \quad (2)$$

The full Chamfer distance is:

$$CD(Q, Q^{gt}) = \frac{A + C}{2} \quad (3)$$

F1-Score. The F1-Score@ η evaluates the geometric match between predicted and ground truth point clouds under a distance threshold η , using:

Precision (prediction accuracy under threshold):

$$P = \sum_{q_{pd} \in Q} \frac{(\min_{q_{gt} \in Q^{gt}} \|q_{gt} - q_{pd}\|) < \eta}{|Q|} \quad (4)$$

Recall (ground truth completeness under threshold):

$$R = \sum_{q_{gt} \in Q^{gt}} \frac{(\min_{q_{pd} \in Q} \|q_{gt} - q_{pd}\|) < \eta}{|Q^{gt}|} \quad (5)$$

The final F1 score is:

$$F1 = \frac{2PR}{P + R} \quad (6)$$

Table 9: Hyperparameter study, Chamfer Distance [mm]. RaySt3R with a single set of hyper parameters samples novel views further away from the origin, as is beneficial in the synthetic OctMAE [3] dataset.

Method	OctMAE [3]	YCB-Video [11]	HomebrewedDB [4]	HOPE [8]
OctMAE [3]	6.48	6.40	6.14	6.97
RaySt3R	5.21	3.56	4.75	3.92
RaySt3R w/ single parameter set	5.21	4.19	5.20	4.66

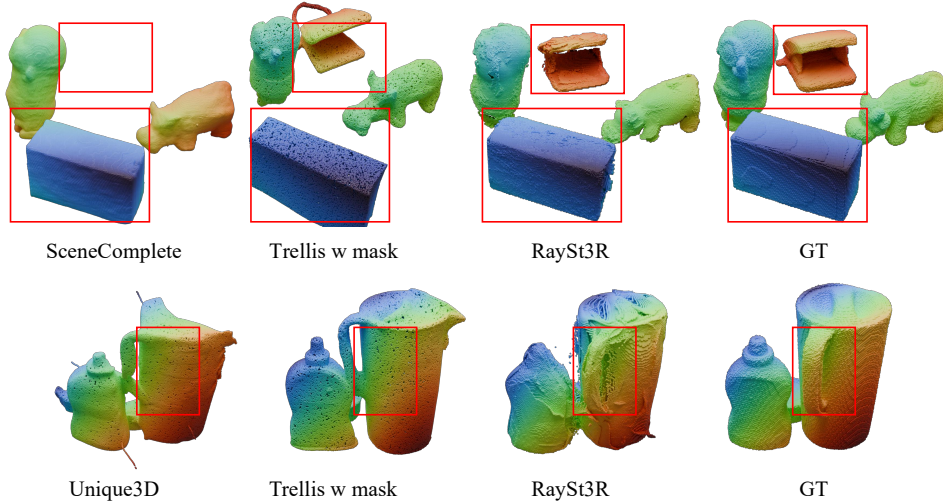


Figure 5: Highlight on relative object placement and aspect ratio. The examples suggest related works such as TRELLIS [10] and Unique3D [9] may produce misaligned geometries in their predictions. We observe minor misalignment in SceneComplete [1], and objects missing, as we describe in the paper.

1.4.2 Visual features from DINOv2

We use DINOv2 [6] to extract visual features, and select the ViT-L with registers. As prior work has shown [2], aggregating features from several intermediate layers may lead to a performance improvement over only considering the last layer. We aggregate features from layers 4, 11, 17, 23 and project them to the ViT token size with a linear layer.

1.5 Qualitative evaluation

1.5.1 Object placement and aspect ratio

We describe our findings of misalignment and incorrect aspect ratios in the paper. It can be challenging to observe these deficiencies, we highlight them in Figure 5. The examples suggest related works such as TRELLIS [10] and Unique3D [9] may produce misaligned geometries in their predictions. We observe minor misalignment in SceneComplete [1], and objects missing, as we describe in the paper.

1.5.2 Baseline comparison

The following Figures provide additional qualitative evaluation of the baselines and RaySt3R.

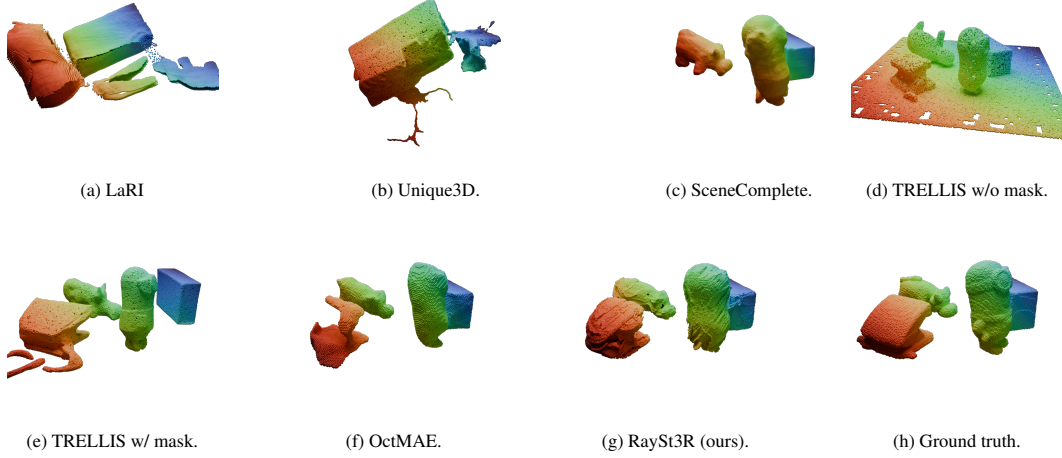


Figure 6: Comparison of different methods on scene 000003, frame 000061 of HomebrewedDB [4].

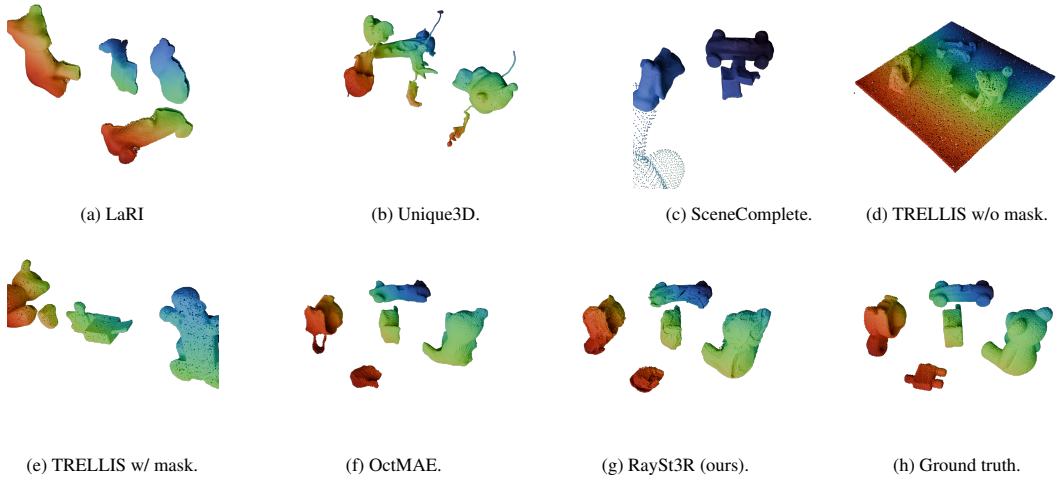


Figure 7: Comparison of different methods on scene 000004, frame 000176 of HomebrewedDB [4].

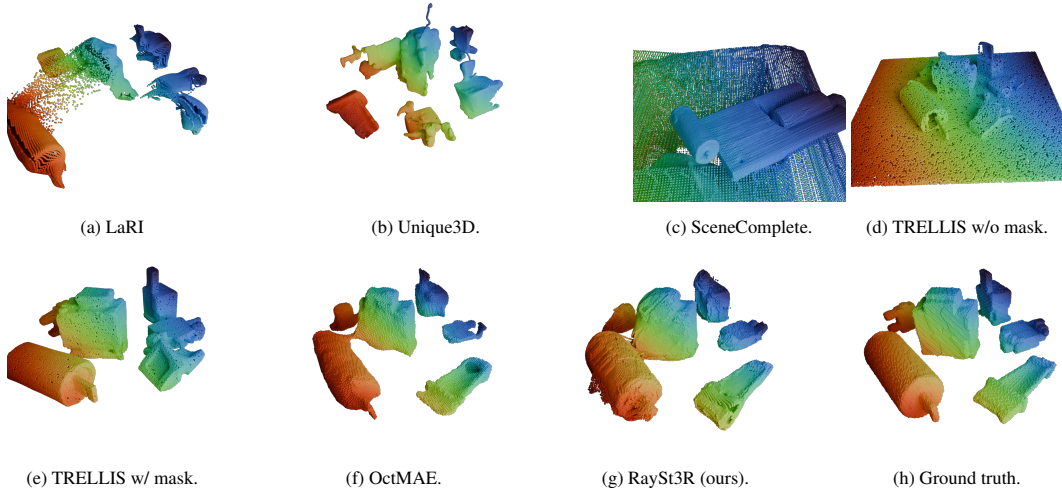


Figure 8: Comparison of different methods on scene 000009, frame 000096 of HomebrewedDB [4].

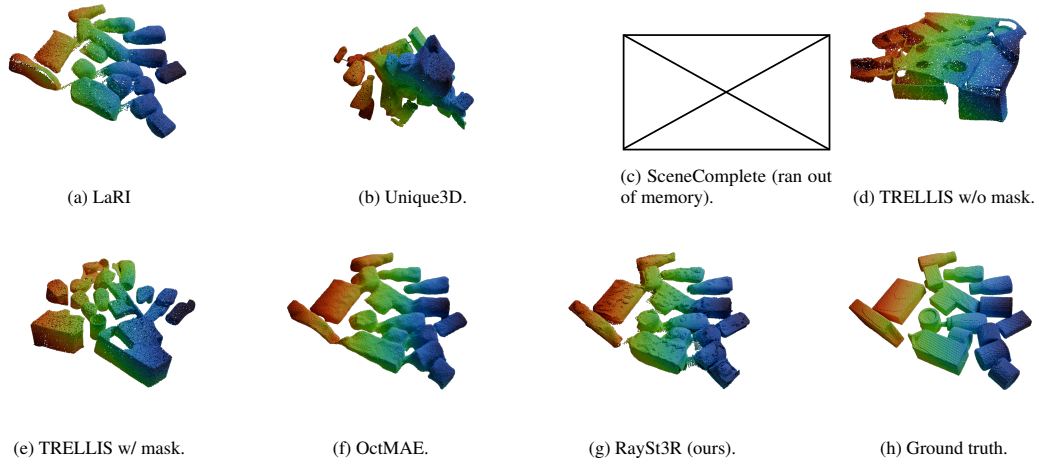


Figure 9: Comparison of different methods on scene 000002, frame 000002 of HOPE [8].

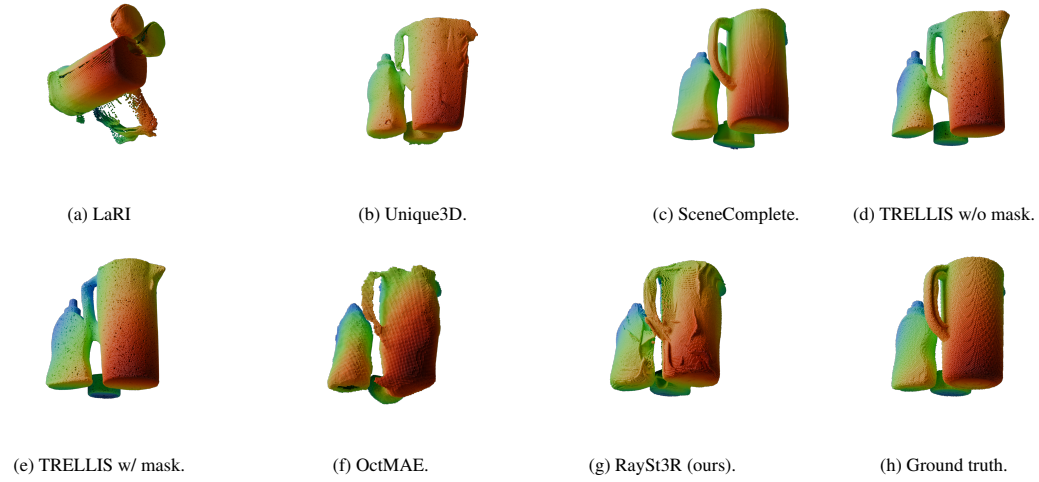


Figure 10: Comparison of different methods on scene 000052, frame 000650 of YCB-Video [11].

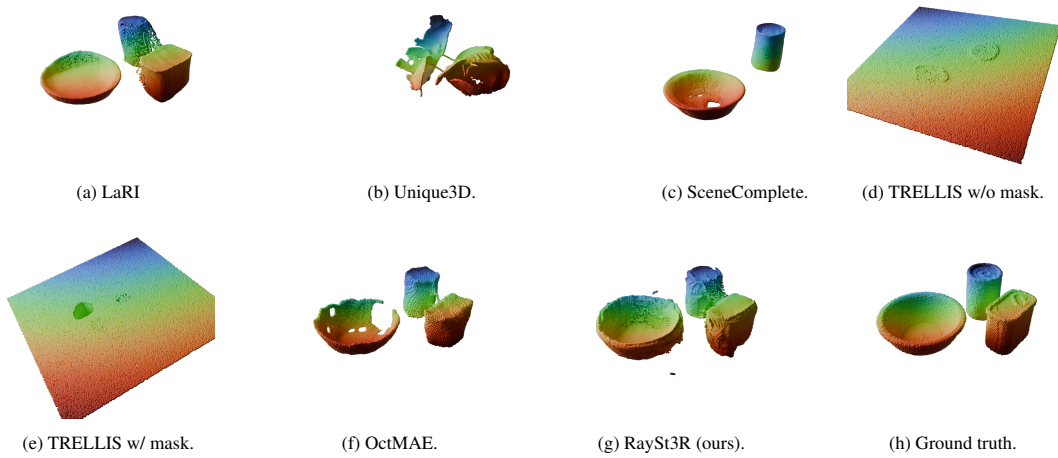


Figure 11: Comparison of different methods on scene 000053, frame 000075 of YCB-Video [11].

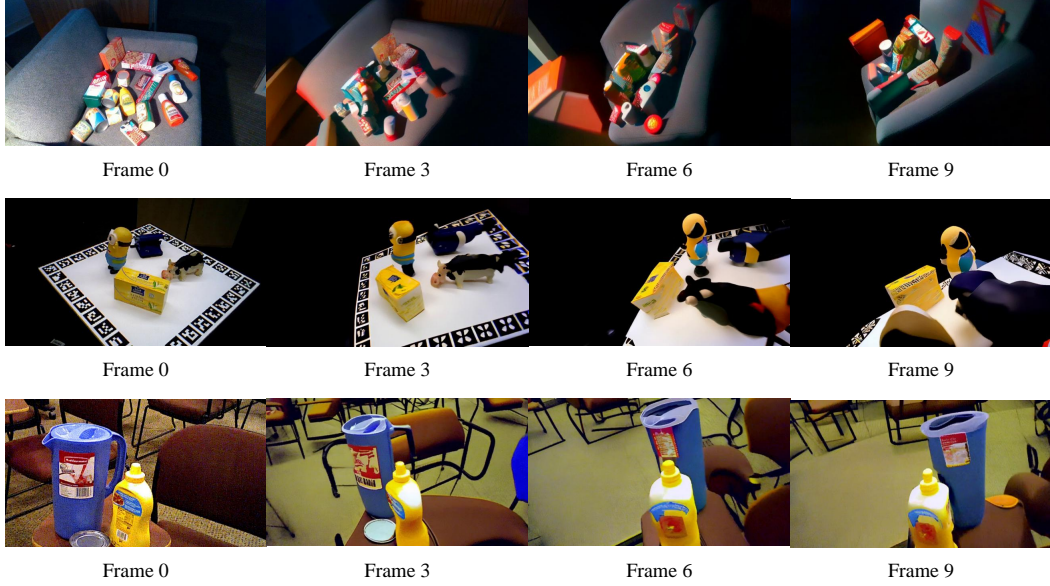


Figure 12: Qualitative results for RGB predictions from ViewCrafter [12] on a HOPE [8] (top), HomebrewedDB [4] (middle), and YCB-Video [11] (bottom) scene. The rendered trajectory is a circle around the center of the scene. The results suggest ViewCrafter [12] is unable to produce 3D-consistent images.

1.5.3 ViewCrafter qualitative results

Recent work has demonstrated video diffusion models may be used to render novel views, and subsequently synthesize geometry [12]. In this context, we observe the images synthesized by ViewCrafter [12] on scenes from the three real-world datasets used in this paper (YCB-Video [11], HOPE [8], and HomebrewedDB [4]). We define the camera trajectory as a circle on the sphere centered in the scene. Figure 12 shows frames produced by ViewCrafter [12]. The results suggest ViewCrafter [12] is unable to reconstruct geometrically consistent images.

References

- [1] A. Agarwal, G. Singh, B. Sen, T. Lozano-Pérez, and L. P. Kaelbling. Scenecomplete: Open-world 3d scene completion in complex real world environments for robot manipulation, 2024.
- [2] A. El-Nouby, M. Klein, S. Zhai, M. A. Bautista, V. Shankar, A. Toshev, J. M. Susskind, and A. Joulin. Scalable pre-training of large autoregressive image models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org, 2024.
- [3] S. Iwase, K. Liu, V. Guizilini, A. Gaidon, K. Kitani, R. Ambrus, and S. Zakharov. Zero-shot multi-object scene completion, 2024.
- [4] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. *ICCVW*, 2019.
- [5] R. Li, B. Zhang, Z. Li, F. Tombari, and P. Wonka. Lari: Layered ray intersections for single-view 3d geometric reasoning. In *arXiv preprint arXiv:2504.18424*, 2025.
- [6] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.
- [7] A. Somani, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 698–700, 1987.

- [8] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.
- [9] K. Wu, F. Liu, Z. Cai, R. Yan, H. Wang, Y. Hu, Y. Duan, and K. Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image, 2024.
- [10] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.
- [11] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018.
- [12] W. Yu, J. Xing, L. Yuan, W. Hu, X. Li, Z. Huang, X. Gao, T.-T. Wong, Y. Shan, and Y. Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.