Hand-held Object Reconstruction from RGB Video with Dynamic Interaction

Supplementary Material

In Sec. A, we provide more implementation details of our method and datasets. Section B includes additional visualizations and comprehensive per-video quantitative results. Finally, in Sec. C, we further explore the limitations of our method and discuss potential solutions.

A. Experiment Details

A.1. Pose Initialization Details

For mesh generation, we prompt ChatGPT with "describe the object interacting with the hand" and use its response as input for the text-to-3D model. During the registration of the 3D model to the image, the generated mesh V is first normalized to fit within a unit cube. Instead of relying on real camera intrinsics, we set the camera intrinsics based on the width I_w and height I_h of the input images as follows:

$$\begin{bmatrix} 1.2min(I_h, I_w) & 0 & \frac{I_w}{2} \\ 0 & 1.2min(I_h, I_w) & \frac{I_h}{2} \\ 0 & 0 & 1 \end{bmatrix}, \quad (1)$$

with the intrinsics fixed during initialization. We adopt a render-and-compare approach to initialize each frame pose by minimizing:

$$\lambda_{proj}\mathcal{L}_{proj} + \lambda_{sem}\mathcal{L}_{sem}, \quad (2)$$

where λ_{proj} and λ_{sem} are both set to 1. Specifically, we sample numerous pose candidates, with rotations uniformly drawn from $SO(3)$ and translations chosen by minimizing the difference between the diagonals of the projected model's tight bounding box and the object mask bounding box. We then select the top five poses with the lowest $\mathcal{L}_{sem}$ values and choose the one closest to the previous frame's pose as the starting point for the Eq. (2) optimization (optimization conducted with only the chosen pose for 100 iterations, others will be deprecated). After obtaining the initial poses for all frames, we refine them to enforce temporal smoothness across the input sequence for another 200 iterations.

A.2. Shape-Pose Joint Optimization Details

During joint optimization, we apply a coarse-to-fine strategy for pose refinement, following the approach in [9]. Specifically, we weight the positional encoding of a 3D point x as $\lambda(x) = (x, \dots, w_k(n_s) \cdot \sin(2^k\pi x), w_k(n_s) \cdot \cos(2^k\pi x), \dots)$, where $k \in [0, L-1]$. In our paper, we set L to 6. Here, $n_s \in [0, N_s]$ denotes the current training step, and N_s is the total number of training steps. The weight function w_k is defined as:

$$w_k(n_s) = \frac{1}{2}\left[1 - cos(clamp(\frac{2n_sL}{N_s} - k, 0, 1) \cdot \pi)\right]. \quad (3)$$

Starting from $n_s = 0$, the positional encodings are gradually activated, with the training loss as follows:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{color} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{mask}\mathcal{L}_{mask} \\ + \lambda_{normal}\mathcal{L}_{normal} + \lambda_{corres}\mathcal{L}_{corres}, \end{aligned} \quad (4)$$

where $\lambda_{mask} = 0.5$, $\lambda_{eik} = 0.1$, $\lambda_{normal} = 0.1$, $\lambda_{corres} = 10$ are set empirically. For pose outlier voting, we use a local window length of 10, set the standard deviation to $\lambda = 2$, and define the Sampson inconsistency rate threshold as 0.3. When calculating $d_{Sampson}$, we only consider correspondence points with a confidence w_p greater than 0.5. To avoid suboptimal results from earlier training stages, we reinitialize the parameters of the corresponding low-quality poses and SDF weights.

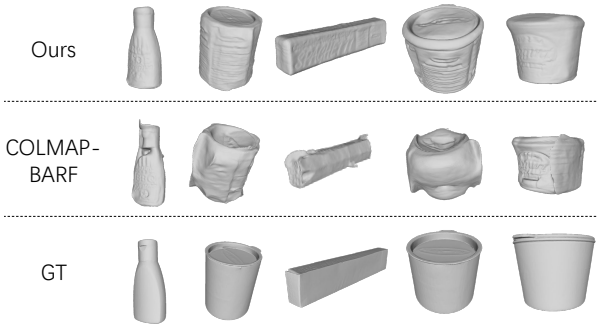

Figure 1. **Further qualitative comparison with the baseline COLMAP-BARF for two-handed object reconstruction on HOPE.** Our method (top row) delivers significantly better reconstruction quality than the baseline (middle row), with results exhibiting fewer artifacts.

A.3. Dataset Details

In Tab. 1, we list the sequence names from the HO3D [3] dataset used for reconstruction, with underlined items corresponding to those in Hampali et al. [4]. The objects from the HOPE [12] and HOD [5] datasets used for evaluation are shown in Tab. 2 and Tab. 3, respectively.

B. Additional Results

B.1. Additional Qualitative Results

Additional qualitative results of two-handed object reconstruction on HOPE. In Fig. 1, we present more reconstruction results on the HOPE dataset. As discussed in

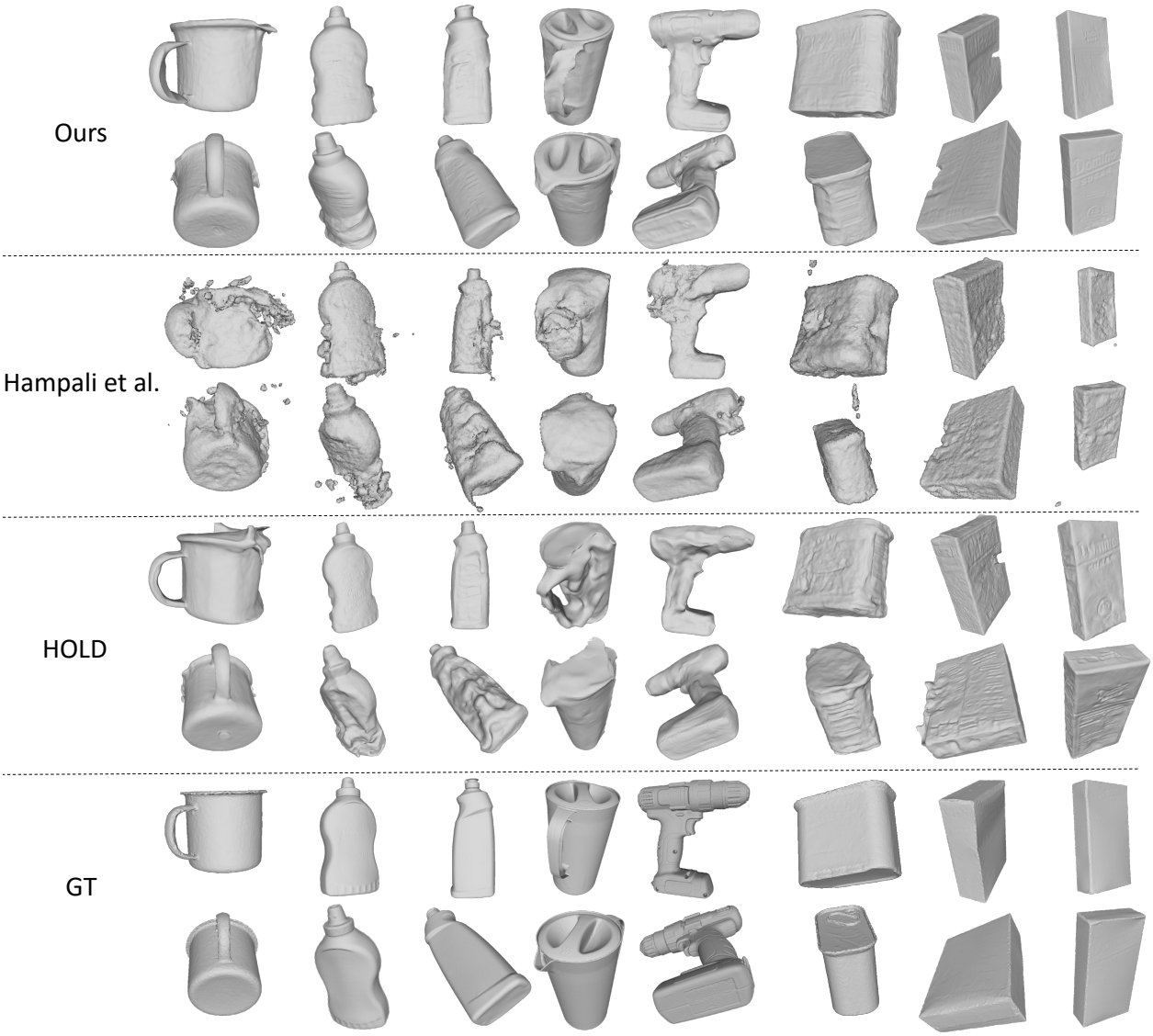

Figure 2. **Comparison with Hampali et al. [4] and HOLD [2] on the same HO3D sequences from two different views.** Our method (top) demonstrates significantly better reconstruction quality compared to Hampali et al. [4] (second row), with more detailed results and fewer artifacts. For reference, the third row includes reconstruction meshes from HOLD [2], where our method still demonstrates superior quality.

the main paper, COLMAP-BARF's results exhibit numerous artifacts, while our method achieves higher fidelity.

Comparison with Hampali et al. [4] and HOLD [2] on the same HO3D sequences. Since the code for Hampali et al. [4] is unavailable, we follow the approach of HOLD [2] and use the same reconstruction sequences (underlined sequences in Tab. 1), comparing them with their released 3D models in point cloud format. As shown in Fig. 2, our results capture finer details and exhibit fewer artifacts, consistent with the quantitative results presented in the main paper. For reference, we also include the reconstruction results from HOLD on the same sequences, where our method still demonstrates superior quality.

Generated priors. We visualize the generated priors for HO3D, HOD, and HOPE in Fig. 3, Fig. 4 and Fig. 5, respectively. Although there are discrepancies in shape and texture, our proposed semantic consistency allows these priors to be used for initializing object poses.

Comparison of different generated priors. We compare the reconstructed meshes and pose trajectories along with the generated priors from different text-to-3D models: Genie [11] (our used, Gen_{genie}) and Shap-e [7] (Gen_{shap-e}). As shown in Fig. 2, although Shap-e generates priors of lower quality compared to Genie, it still reconstructs rea-

Object	Seq.
cracker box	MC1
cracker box	MC6
sugar box	ShSu10
sugar box	ShSu12
mustard bottle	SM2
mustard bottle	SM4
meta can	GPMF12
meta can	GPMF14
banana	BB12
banana	BB14
pitcher	AP10
pitcher	AP13
bleach cleanser	ABF14
bleach cleanser	SB10
mug	SMu1
mug	SMu40
power drill	ND2
power drill	MDF14
scissors	GSF12
scissors	GSF14

Table 1. **HO3D sequences for evaluation.** The underline sequences are the same with Hampali et al. [4].

Object
BBQSauce
Butter
Cherries
Milk
Spaghetti
TomatoSauce
Tuna
Yogurt

Table 2. **HOPE sequences for evaluation.**

Object
Rubber Duck
Robot
Cat
AirPods
David
Giuliano
Marseille

Table 3. **HOD sequences for evaluation.**

sonable meshes and poses in most cases, demonstrating the robustness of our system to variations in priors, even when there are significant discrepancies in geometry or texture (*e.g.*, Drill, Pitcher, and Cracker box). However, it is worth noting that in cases where Shap-e's priors are of particularly low quality (*e.g.*, Scissors), the reconstructed results may exhibit more artifacts. Higher-quality generated priors can further enhance our results. Additionally, we include results from HOLD, where Gen$_{shap-e}$ consistently outperforms HOLD in most cases, highlighting the effectiveness of the generated priors.

	RPE$_t(cm)\downarrow$	RPE$_r(^\circ)\downarrow$	ATE$(m)\downarrow$
GT	**2.019**	**3.049**	**0.077**
GT w/o $\mathcal{L}_{sem}$	4.256	6.304	0.202
Gen$_{genie}$	3.629	4.738	0.115
Gen$_{genie}$ w/o $\mathcal{L}_{sem}$	6.011	8.652	0.255

Table 4. **The effectiveness for $\mathcal{L}_{sem}$ for pose initialization.** GT indicates the use of ground-truth 3D models for pose initialization, while Gen$_{genie}$ refers to using the generated priors from Genie for pose initialization. w/o $\mathcal{L}_{sem}$ means aligning the 3D model to the 2D image without $\mathcal{L}_{sem}$.

Init.	Pose initialization		Joint optimization with our method		
	RPE$_t(cm)\downarrow$	RPE$_r(^\circ)\downarrow$	RPE$_t(cm)\downarrow$	RPE$_r(^\circ)\downarrow$	CD $(cm^2)\downarrow$
Hand	2.844	3.541	1.597	2.724	1.04
COLMAP	3.061	4.429	2.682	3.937	1.80
Ours	**2.693**	**3.531**	**1.523**	**2.441**	**0.53**
COLMAP	5.127	7.558	4.67	6.396	2.39
Ours	**2.729**	**3.905**	**1.938**	**2.924**	**0.40**

Table 5. **Comparison of different pose initialization methods.** "Hand" refers to estimate object motion using hand motions.

B.2. Additional Quantitative Results

The effectiveness of $\mathcal{L}_{sem}$ for pose initialization. Table 4 shows the pose initialization results that compared with variants without $\mathcal{L}_{sem}$. The results indicate that relying solely on masks to initialize poses using generated priors results in significant errors, even when ground-truth meshes are used. $\mathcal{L}_{sem}$ effectively resolves ambiguities and enhances pose accuracy. This explains the inferior reconstruction results when omitting $\mathcal{L}_{sem}$ in pose initialization.

Comparison of different pose initialization methods. In Tab. 5, we evaluate the initial poses and final results obtained with our proposed optimization method using eight HO3D sequences with different pose initializations. The top rows correspond to four sequences where the hand firmly grasps the object, while the bottom rows correspond to four sequences with freely moving objects. For hand pose as a proxy, we compare using HOLD's estimation. It outperforms COLMAP in most cases, particularly for smaller objects. However, larger objects often occlude the hand, making pose estimation more challenging. A key limitation of using the hand as a proxy is its reliance on a firm grip, which can degrade reconstruction quality. COLMAP performs well on objects with rich features but can completely fail on more challenging ones. Meanwhile, our joint optimization consistently improves results across different pose initialization methods, demonstrating its robustness.

Per-sequence quantitative results. We present per-sequence quantitative results for all three datasets in Tab. 6 (HO3D), Tab. 7 (HOD), and Tab. 8 (HOPE).

C. Discussions

In this paper, we propose a novel system for reconstructing hand-held objects during dynamic interactions involving one or both hands. While the method produces high-quality meshes, it has certain limitations.

Firstly, the effectiveness of our method depends on the quality of the generated priors, which may not perform well with unique objects that current generators struggle to represent accurately. Recent advancements in 3D generators [1, 10] have enabled the creation of increasingly higher-quality priors. In Fig. 7, we present a failure case. Our method relies on semantic consistency for pose initialization. However, in some instances, DINO may produce nearly identical features for different sides of an object, resulting in incorrect pose estimations. Recent progress in semantic feature extraction [14] may help address this limitation. To further improve reconstruction quality, our method can adopt the approach of HOLD [2], which leverages hand-object interaction priors [6], refines poses using reconstructed meshes, and retrains the network. Additionally, the joint optimization of neural implicit fields is computationally intensive, and adopting alternative representations [8] may enhance efficiency.

Figure 3. **Generated prior using Genie for HO3D.**

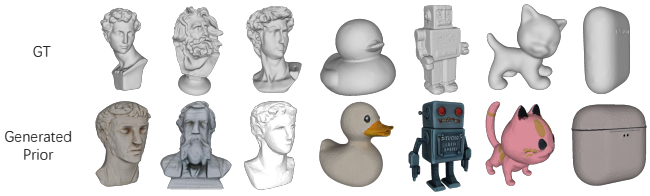

Figure 4. **Generated prior using Genie for HOD.**

Figure 5. **Generated prior using Genie for HOPE.**

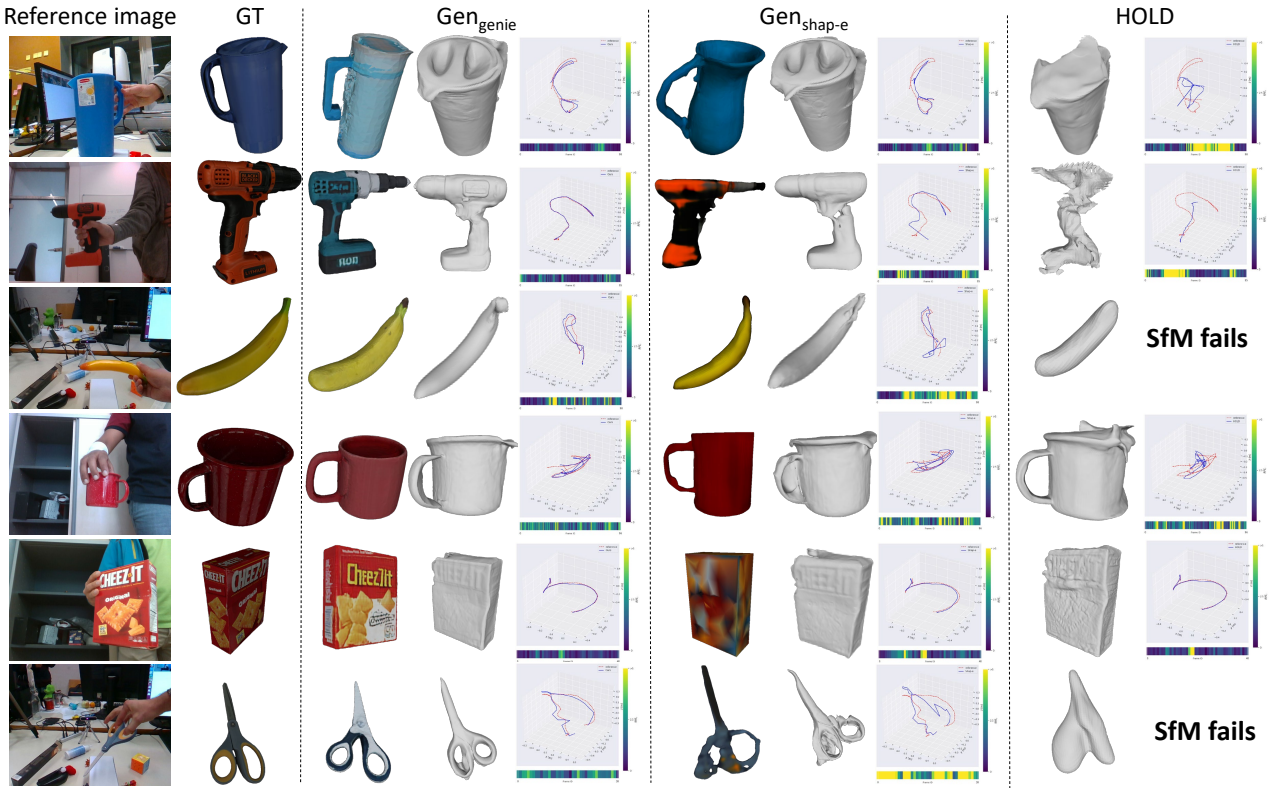

Figure 6. **Comparison of different priors.** We present visualizations of the generated priors, reconstructed meshes, and pose trajectory results from two text-to-3D models: Genie [11] (second column) and Shap-e [7] (third column). Despite producing lower-quality priors, Shap-e successfully reconstructs reasonable meshes and poses in most cases, highlighting the robustness of our system to varying priors. Moreover, its results outperform those from HOLD (last column).

Seq.	Metric	IHOI [13]	HOLD [2]	Ours
MC1	CD (cm^2) ↓	2.67	**0.46**	0.79
	RPE$_t$(cm) ↓	-	**1.946**	2.164
	RPE$_r$(°) ↓	-	**2.543**	2.600
	ATE(m)↓	-	0.067	**0.060**
MC6	CD (cm^2) ↓	3.10	0.26	**0.19**
	RPE$_t$(cm) ↓	-	0.769	**0.736**
	RPE$_r$(°) ↓	-	1.199	**1.165**
	ATE(m)↓	-	0.019	**0.015**
ShSu10	CD (cm^2) ↓	0.57	0.42	**0.24**
	RPE$_t$(cm) ↓	-	2.606	**2.490**
	RPE$_r$(°) ↓	-	4.482	**3.251**
	ATE(m)↓	-	0.206	**0.081**
ShSu12	CD (cm^2) ↓	1.84	**0.17**	0.68
	RPE$_t$(cm) ↓	-	1.594	**1.475**
	RPE$_r$(°) ↓	-	3.894	**2.663**
	ATE(m)↓	-	0.251	**0.045**
SM2	CD (cm^2) ↓	0.32	0.45	**0.14**
	RPE$_t$(cm) ↓	-	5.196	**2.204**
	RPE$_r$(°) ↓	-	7.947	**3.328**
	ATE(m)↓	-	0.333	**0.057**
SM4	CD (cm^2) ↓	0.62	**0.46**	0.58
	RPE$_t$(cm) ↓	-	5.271	**3.412**
	RPE$_r$(°) ↓	-	7.838	**4.637**
	ATE(m)↓	-	0.264	**0.092**
GPMF12	CD (cm^2) ↓	1.38	**0.17**	**0.17**
	RPE$_t$(cm) ↓	-	6.511	**1.773**
	RPE$_r$(°) ↓	-	11.529	**2.791**
	ATE(m)↓	-	0.190	**0.033**
GPMF14	CD (cm^2) ↓	0.88	0.17	**0.05**
	RPE$_t$(cm) ↓	-	9.285	**1.498**
	RPE$_r$(°) ↓	-	13.589	**1.987**
	ATE(m)↓	-	0.408	**0.030**
BB12	CD (cm^2) ↓	4.37	1.94	**1.66**
	RPE$_t$(cm) ↓	-	-	**2.024**
	RPE$_r$(°) ↓	-	-	**3.612**
	ATE(m)↓	-	-	**0.047**
BB14	CD (cm^2) ↓	1.33	2.11	**0.48**
	RPE$_t$(cm) ↓	-	-	**3.34**
	RPE$_r$(°) ↓	-	-	**4.767**
	ATE(m)↓	-	-	**0.201**

(a) **Per-sequence comparison on the HO3D dataset. (Part 1)**

Seq.	Metric	IHOI [13]	HOLD [2]	Ours
AP10	CD (cm^2) ↓	6.47	1.79	**0.76**
	RPE$_t$(cm) ↓	-	1.859	**1.319**
	RPE$_r$(°) ↓	-	2.011	**1.716**
	ATE(m)↓	-	0.155	**0.090**
AP13	CD (cm^2) ↓	5.38	2.72	**0.32**
	RPE$_t$(cm) ↓	-	4.163	**1.214**
	RPE$_r$(°) ↓	-	5.650	**1.504**
	ATE(m)↓	-	0.364	**0.117**
ABF14	CD (cm^2) ↓	0.84	0.69	**0.54**
	RPE$_t$(cm) ↓	-	4.725	**3.564**
	RPE$_r$(°) ↓	-	6.874	**5.543**
	ATE(m)↓	-	0.435	**0.158**
SB10	CD (cm^2) ↓	0.71	4.58	**0.39**
	RPE$_t$(cm) ↓	-	4.719	**3.093**
	RPE$_r$(°) ↓	-	5.764	**3.567**
	ATE(m)↓	-	0.368	**0.129**
SMu1	CD (cm^2) ↓	5.58	**1.45**	1.54
	RPE$_t$(cm) ↓	-	2.741	**2.102**
	RPE$_r$(°) ↓	-	4.445	**3.619**
	ATE(m)↓	-	0.205	**0.075**
SMu40	CD (cm^2) ↓	0.67	1.16	**0.37**
	RPE$_t$(cm) ↓	-	6.501	3.009
	RPE$_r$(°) ↓	-	9.868	3.576
	ATE(m)↓	-	0.379	0.082
ND2	CD (cm^2) ↓	0.71	3.23	**0.24**
	RPE$_t$(cm) ↓	-	2.166	**1.053**
	RPE$_r$(°) ↓	-	3.086	**1.652**
	ATE(m)↓	-	0.248	**0.017**
MDF14	CD (cm^2) ↓	3.38	0.43	**0.41**
	RPE$_t$(cm) ↓	-	6.245	**1.499**
	RPE$_r$(°) ↓	-	4.870	**2.003**
	ATE(m)↓	-	0.354	**0.038**
GSF12	CD (cm^2) ↓	28.28	6.73	**0.23**
	RPE$_t$(cm) ↓	-	-	**1.150**
	RPE$_r$(°) ↓	-	-	**1.908**
	ATE(m)↓	-	-	**0.024**
GSF14	CD (cm^2) ↓	26.42	6.99	**0.47**
	RPE$_t$(cm) ↓	-	-	**2.463**
	RPE$_r$(°) ↓	-	-	**3.342**
	ATE(m)↓	-	-	**0.058**
Mean	CD (cm^2) ↓	4.78	1.82	**0.51**
	RPE$_t$(cm) ↓	-	4.143	**2.129**
	RPE$_r$(°) ↓	-	5.974	**2.962**
	ATE(m)↓	-	0.265	**0.072**

(b) **Per-sequence comparison on the HO3D dataset. (Part 2)**

Table 6. **Per-sequence comparison on the HO3D dataset.**

Seq.	IHOI [13]	HOLD [2]	Ours
Rubber Duck	7.25	2.70	**0.34**
Robot	3.89	0.89	**0.29**
Cat	7.50	1.51	**0.18**
AirPods	6.06	1.46	**1.39**
David	5.81	0.52	**0.21**
Giuliano	9.32	0.39	**0.36**
Marseille	11.11	1.43	**0.34**
Mean	7.28	1.27	**0.44**

Table 7. **Per-sequence comparison on the HOD dataset.** The metric is the Chamfer Distance in unit size, where the lower value indicates the higher reconstruction quality.

Seq.	COLMAP-BARF	Ours
BBQSauce	0.37	**0.12**
Buter	1.50	**0.07**
Cherries	0.94	**0.15**
Milk	1.61	**0.25**
Spaghetti	0.82	**0.22**
TomatoSauce	0.21	**0.14**
Tuna	1.76	**1.30**
Yogurt	0.51	**0.12**
Mean	0.96	**0.29**

Table 8. **Per-sequence comparison on the HOPE dataset.** The metric is the Chamfer Distance in cm^2, where the lower value indicates the higher reconstruction quality.

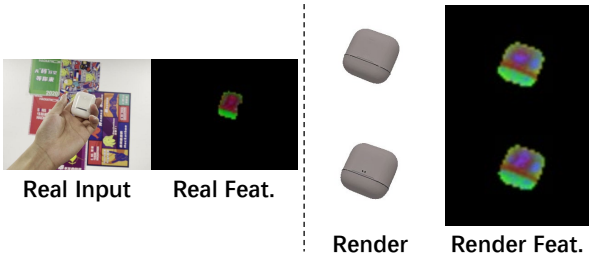

Figure 7. **Ambiguous features.** DINO generates nearly the same features for different object sides, causing incorrect poses.

References

[1] Maciej Bala, Yin Cui, Yifan Ding, Yunhao Ge, Zekun Hao, Jon Hasselgren, Jacob Huffman, Jingyi Jin, JP Lewis, Zhaoshuo Li, et al. Edify 3d: Scalable high-quality 3d asset generation. *arXiv preprint arXiv:2411.07135*, 2024. 4

[2] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 2, 4, 6, 7

[3] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 1

[4] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *arXiv preprint arXiv:2211.16193*, 2022. 1, 2, 3

[5] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1

[6] Shijian Jiang, Qi Ye, Rengan Xie, Yuchi Huo, Xiang Li, Yang Zhou, and Jiming Chen. In-hand 3d object reconstruction from a monocular rgb video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2525–2533, 2024. 4

[7] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2, 5

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 4

[9] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021. 1

[10] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv preprint arXiv:2407.02445*, 2024. 4

[11] Luma Team. Luma genie 1.0. https://www.luma-ai.com/luma-genie-1-0/. 2, 5

[12] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan 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. 1

[13] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022. 6, 7

[14] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 4