HORT: Monocular Hand-held Objects Reconstruction with Transformers Supplementary Material

Zerui Chen[1] Rolandos Alexandros Potamias[2] Shizhe Chen[1] Cordelia Schmid[1]
[1]Inria, École normale supérieure, CNRS, PSL Research University [2]Imperial College London
firstname.lastname@inria.fr r.potamias@imperial.ac.uk
https://zerchen.github.io/projects/hort.html

In supplementary material, we provide more details of our model architectures and additional results. We first present details of our model architecture in Section A. Then in Section B, we discuss additional experimental results.

A. Network Architecture

A.1. PointNet encoder

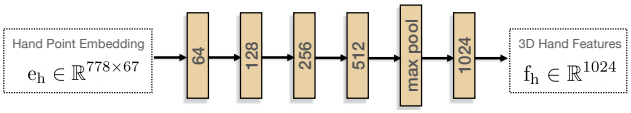

Figure 1. Network architecture used for our PointNet encoder. We use five fully-connected layers for efficiency. The number in the box denotes the dimension of features.

As shown in Figure 1, we employ five fully-connected layers to construct an efficient PointNet encoder. It takes the point clouds transformed into different hand coordinate systems ($e_h \in \mathbb{R}^{778\times67}$) and outputs 1024-dimensional global features $f_h \in \mathbb{R}^{1024}$ for the sparse point cloud decoder.

A.2. Dense point cloud decoder

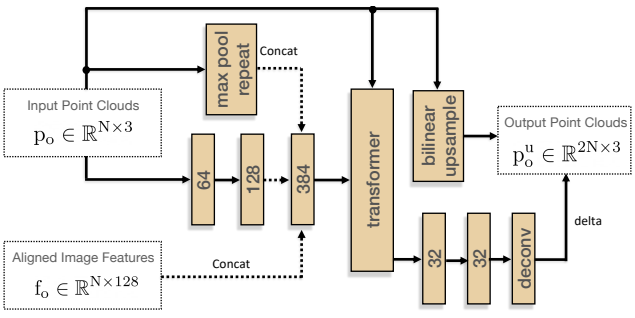

Figure 2. Network architecture used for reconstructing dense object point clouds. Given the input point clouds and its corresponding image features, the network performs point clouds upsampling. The number in the box denotes the dimension of features.

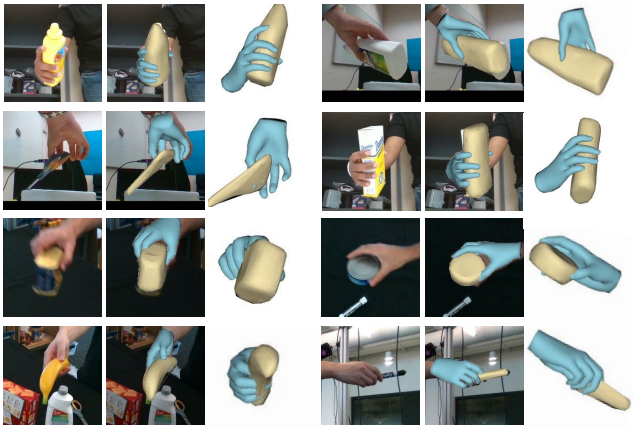

Figure 3. Qualitative results of reconstructed object meshes produced by our approach on HO3D [7] and DexYCB [2] datasets.

Figure 2 illustrates the network architecture for reconstructing dense object point clouds. Given input point clouds $p_o \in \mathbb{R}^{N\times3}$ and aligned image features $f_o \in \mathbb{R}^{N\times128}$, we first apply convolution layers and max pooling to p_o, then concatenate the processed features with f_o. As discussed in the main paper, we perform the same operations for predicted hand vertices v_h, equipping each object point with its surrounding hand context. Next, we upsample the input object point clouds via bilinear interpolation and employ a transformer model to predict 3D offsets for the initially upsampled object points. Our model consists of two consecutive blocks, as shown in Figure 2, which progressively upsample sparse point clouds $p_o^s \in \mathbb{R}^{2048\times3}$ to dense object point clouds $p_o^d \in \mathbb{R}^{16384\times3}$.

B. Experimental Results

B.1. Object mesh reconstruction results

When needed, we can easily convert our reconstructed object point clouds into meshes. As shown in Figure 3, we employ the alpha shapes algorithm [5], as implemented in the Open3D library [11], to generate high-fidelity ob-

Table 1. Comparison with previous methods on DexYCB.

	Methods	mesh	FS@5 ↑	FS@10 ↑	CD ↓	Cr ↑	Pd ↓
R1	GF [9]	✓	0.39	0.66	4.5	0.96	0.92
R2	AlignSDF [3]	✓	0.41	0.68	3.9	0.97	1.08
R3	gSDF [4]	✓	0.44	0.71	3.4	0.95	0.94
R4	D-SCO [6]	×	**0.63**	0.82	1.3	-	-
R5	HORT (Ours)	×	**0.63**	**0.85**	**1.1**	**0.98**	**0.90**
R6	HORT (Ours)	✓	0.62	**0.85**	**1.1**	**0.98**	**0.90**

Table 2. Ablation studies of hand accuracy on DexYCB dataset.

Methods	Noise	E_h ↓	FS@5 ↑	FS@10 ↑	CD ↓	C_r ↑	P_d ↓
HORT (Ours)	$\sigma = 0.0$	0.00	**0.64**	**0.88**	**1.0**	**0.98**	**0.88**
HORT (Ours)	$\sigma = 0.1$	7.67	0.60	0.85	1.3	0.96	0.93
HORT (Ours)	$\sigma = 0.5$	36.11	0.54	0.79	1.6	0.91	1.02

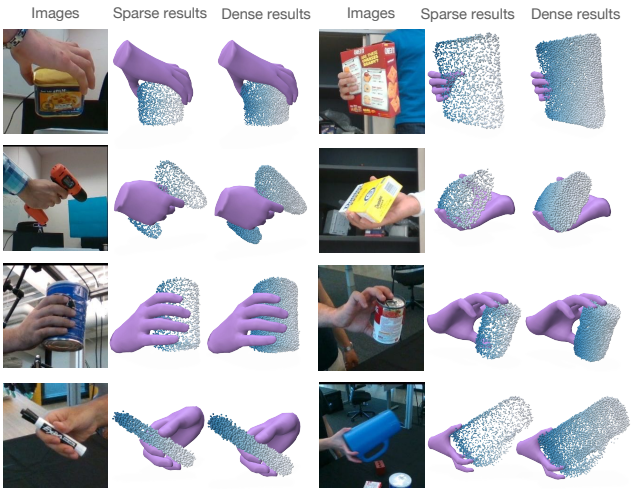

Figure 4. Qualitative comparison of our reconstructed sparse and dense object point clouds on HO3D and DexYCB datasets.

ject meshes. Table 1 presents our quantitative analysis of point cloud and mesh predictions. Here, we use predicted hand poses for different approaches on DexYCB [2] dataset. Compared to evaluations on point clouds in R5, R6 evaluates object meshes and achieves similar quantitative performance, further demonstrating the flexibility and efficiency of our point cloud representation. Moreover, compared to previous methods, HORT produces more physically plausible hand-object configurations and achieves better results in terms of Contact Ratio (Cr) and Penetration Depth (Pd).

B.2. Comparison for 3D reconstruction densities

Figure 4 qualitatively compares our reconstructed sparse and dense object point clouds. We observe that our reconstructed sparse and dense point clouds are consistent in the general shape of the object. Our dense point clouds contain more surface details for the manipulated object.

B.3. Impact of imperfect hand poses

Table 2 ablates the impact of hand accuracy to our model by gradually adding Gaussian noise to ground-truth hand

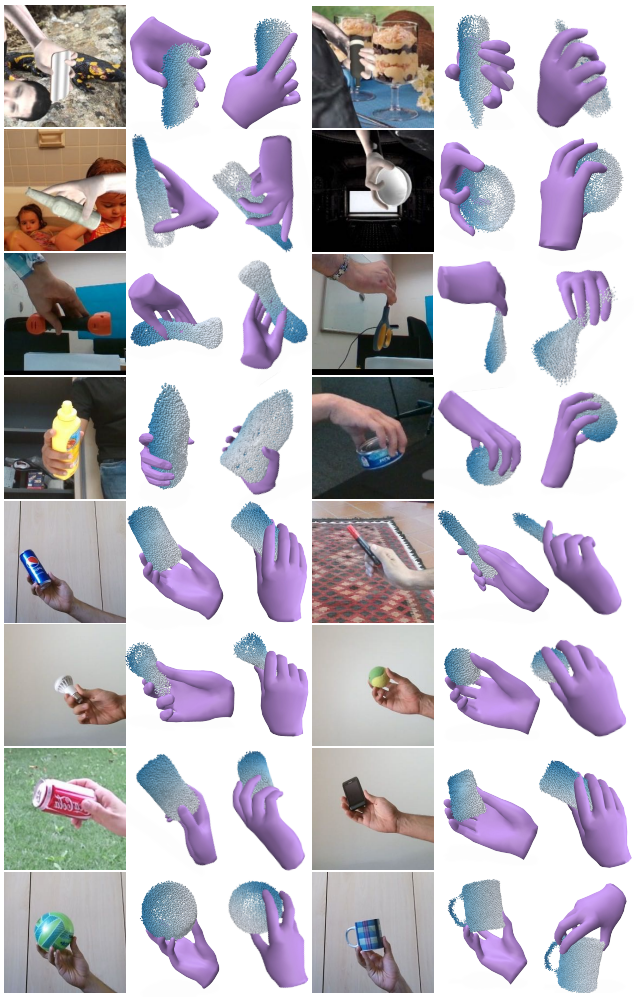

Figure 5. Qualitative results of HORT on synthetic ObMan [8], indoor HO3D [7] and DexYCB [2], and in-the-wild CORe50 [10] images. Our model shows impressive results on all these domains.

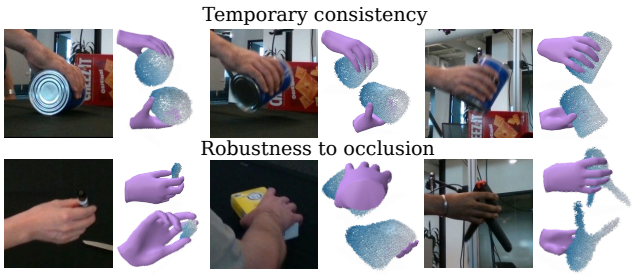

Figure 6. Qualitative analysis on DexYCB datasets.

poses. As a result, the hand joint error E_h increases from 7.67mm to 36.11mm with more noises. Our HORT model is robust to noisy hand poses and can still reconstruct plausible 3D hand-held objects.

B.4. Additional qualitative results

In this section, we present additional qualitative examples in Figure 5 to demonstrate that our approach produces

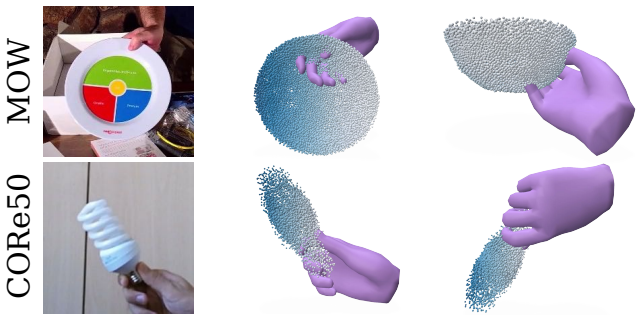

Figure 7. Failure cases on MOW [1] and CORe50 [10] images.

high-quality 3D reconstructions across various challenging scenes. Our model performs well on both the synthetic ObMan [8] dataset and the real-world HO3D [7] and DexYCB [2] datasets. Furthermore, we show that our model can generate reliable predictions on in-the-wild CORe50 [10] images, highlighting the ability of our HORT model to effectively generalize to diverse object instances and textures in unconstrained environments. Figure 6 demonstrates that our reconstructed results are temporally consistent and robust to hand occlusion.

B.5. Failure cases analysis

While HORT achieves impressive qualitative results, it struggles to accurately infer hand-object configurations when the hand is heavily occluded (Figure 7, top) and to recover fine-grained geometry for rare objects (Figure 7, bottom). Scaling training data with a broader range of objects and grasping poses could help mitigate these limitations.

References

[1] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 3

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 1, 2, 3

[3] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-Aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 2

[4] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-driven signed distance functions for 3D hand-object reconstruction. In *CVPR*, 2023. 2

[5] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 1983. 1

[6] Bowen Fu, Gu Wang, Chenyangguang Zhang, Yan Di, Ziqin Huang, Zhiying Leng, Fabian Manhardt, Xiangyang Ji, and Federico Tombari. D-SCo: Dual-stream conditional diffusion for monocular hand-held object reconstruction. In *ECCV*, 2024. 2

[7] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 1, 2, 3

[8] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2, 3

[9] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *3DV*, 2020. 2

[10] Vincenzo Lomonaco and Davide Maltoni. CORe50: a new dataset and benchmark for continuous object recognition. In *CoRL*, 2017. 2, 3

[11] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018. 1