HandBooster: Boosting 3D Hand-Mesh Reconstruction by Conditional Synthesis and Sampling of Hand-Object Interactions

Supplementary Material

Overview

This supp. material contains the following four parts:

- Part 1: We provide detailed novelty analysis on our synthetic data.
- Part 2: We provide implementation details of our method.
- Part 3: We present more experimental results, including ablation studies and qualitative comparisons.
- Part 4: We discuss the limitations of our approach and then the future work.

Part 1: Novelty Analysis on our Synthetic Data

To show the novelty of our synthetic data on hand appearance/pose, camera view, and background, we present several visualizations of the intermediate results in our method.

First, to have a global picture of the similarity between our synthetic data and the real-world data, we present further visualizations of the similarity distribution for more object categories in DexYCB and HO3D. As illustrated in Fig. A, for each randomly-selected object category, the X-axis represents the normalized similarities between the sampled synthetic and real-world grasping poses, and the Y-axis represents the number of sampled synthetic grasping poses. It can be found that only a small portion is similar to the real-world data, revealing the hand pose diversity of our synthetic data. Please refer to Sec. 3.3 in the main paper for the details. On the other hand, to demonstrate the hand poses novelty and diversity of our synthetic data, we randomly choose some grasping poses sampled by our similarity-aware distribution sampling strategies for each object category. Then, for each randomly-chosen grasping pose, we retrieve the most similar pose within the same object category in the given real-world data. We use the same similarity calculation approach as we described in our similarity-aware sampling strategies, *i.e*., computing the cosine similarity between the pose vectors (consisting of the hand pose at the canonical pose and the object rotation and translation) of the synthetic and real-world data. As Fig. D shows, most of the retrieved results are not the same as the grasping poses generated by our method, thereby demonstrating that our method can help to generate novel data different from the training ones.

Second, to obtain an intuitive understanding of the novelty and diversity of our generated data on hand appearance, camera view, and background, we show more generated cases in Fig. E. Further, to exploit the novelty of the generated data, we retrieve the nearest training sample (real-world data) for each generated data sample; note that we measure the sample distance using the feature-space vector. Similar to calculating the FID score, we use the InceptionV3 [6] model pre-trained on ImageNet [2] as the feature extractor and calculate the cosine similarity. As Fig. F shows, our generated data is not similar to its nearest neighbors in the training set, demonstrating the capability of our method to generate novel and diverse hand appearances and backgrounds. To validate the accuracy of our nearest sample retrieving approach, we also randomly select some real-world images and retrieve their top 5 nearest samples in the corresponding real-world dataset. As shown in Fig. G, all the retrieved samples are from the same video sequence of the input image and they are very similar. This demonstrates the effectiveness of this feature-space retrieval and further demonstrates the novelty of our generated data.

Part 2: Implementation Details

For the training of the diffusion model, gradient accumulation is performed every two iterations. We set the timestamp T to 1,000 and use an image resolution of 128×128. For intra-distribution sampling, we set N to 1,000 and M to 500/10 for each object category in DexYCB/HO3D, considering their different pose diversity. We randomly generate three Euler angle rotations within $[-30^\circ, 30^\circ]$ on the X, Y, and Z axes for the 3D rotation augmentation on camera views. For fairness, the diffusion models and the condition sampling are conducted separately on different data splits of DexYCB. Due to the small scale of HO3D (only 55 sequences for training and most grasping poses are similar), for HO3D, we fine-tune the diffusion model pre-trained on DexYCB for better generalization and stable training. For the 3D hand-mesh reconstruction baselines, we generate an equal number of synthetic samples as the original training set. We employ MobRecon [1] for the edge cases filtering and perform a simple hyperparameter search to determine the thresholds, yielding 10/20mm on DexYCB/HO3D for J-PE and V-PE after performing PA, and 25/50mm on DexYCB/HO3D for J-PE and V-PE before performing PA. Our hyperparameters are consistent with [1, 5, 9], except for doubling the training epochs to improve convergence. Specifically, for H2ONet [9], we follow its two-stage training strategy, where the losses related to the hand orientation are not applied in the first stage and all losses are used in the second stage. The initial learning rate is set to 0.001/0.0001 for the first/second stages, and decayed by 0.1 at the 60-th epoch for each stage (training 76 epochs in total).

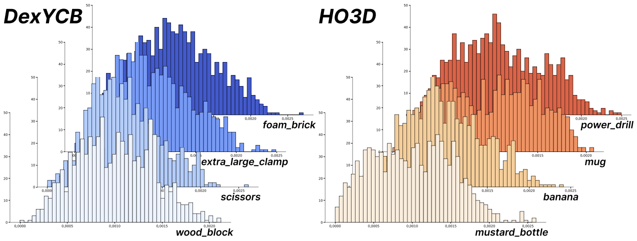

Figure A. Visualizations in sampling. For randomly-selected object categories in DexYCB and HO3D, we show the histograms of similarity between sampled real-world and synthetic poses. The X-axis represents the values of normalized similarities, and the Y-axis represents the number of sampled synthetic grasping poses.

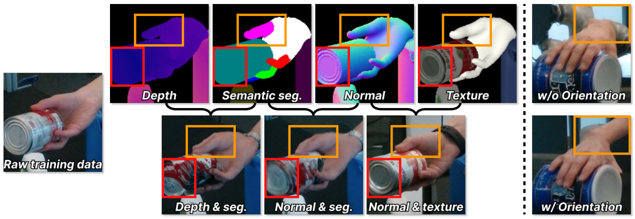

Figure B. Qualitative comparison with different conditions.

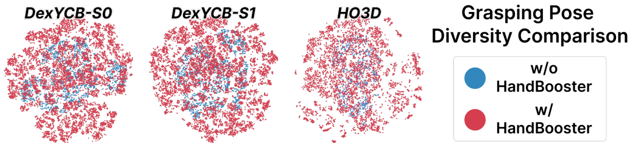

Figure C. t-SNE visualizations w/ and w/o our HandBooster.

Part 3: More Experimental Results

More Ablation Studies

To further show the effectiveness of the main components in our method, we conduct more ablation studies on the DexYCB dataset, as shown in Tab. A. We have shown the experimental results of MobRecon in our main paper. For the other two baseline methods, it can be found that our novel condition creator ("+ Novel Conditions") and similarity-aware intra- and cross-distribution sampling strategies ("+ Dist. Sampling") help to improve their performances. Finally, without our edge cases filtering module ("- Edge Cases Filtering), all metrics of the three baselines degrade slightly, revealing the effectiveness of our design.

As Fig. B shows, using the depth and segmentation maps results in unclear boundaries and chaotic object appearance. The hand orientation gives hints to generate arms, which are not included in our conditions. Removing it may cause unrealistic artifacts. Also, we provide the t-SNE plots compared with the DexYCB and HO3D datasets. As shown in Fig. C, the results indicate our diversity.

More Qualitative Comparisons

We provide more qualitative comparisons on the DexYCB and HO3D datasets, as illustrated in Figs. H and I.

	Models	*Root-relative*		*Procrustes Align.*	
		J-PE ↓	V-PE ↓	J-PE ↓	V-PE ↓
(a)	MobRecon [1]	14.20	13.05	6.36	5.59
(b)	+ Our HandBooster	**13.25**	**12.34**	**6.20**	**5.36**
(c)	- Edge Cases Filtering	14.35	13.38	6.72	6.02
(d)	HandOccNet [5]	14.04	13.09	5.80	5.50
(e)	+ Novel Conditions	12.24	11.82	5.40	5.21
(f)	+ Dist. Sampling	**11.93**	**11.53**	**5.23**	**5.05**
(g)	- Edge Cases Filtering	12.27	11.87	5.42	5.24
(h)	H2ONet [9]	14.02	13.03	5.65	5.45
(i)	+ Novel Conditions	13.17	12.75	5.23	5.12
(j)	+ Dist. Sampling	**12.95**	**12.54**	**5.16**	**5.09**
(k)	- Edge Cases Filtering	13.51	13.14	5.28	5.26

Table A. Additional ablation studies on major components.

Edge Cases Filtering

We also provide the visualizations for our generated data samples classified as "edge cases" by our edge cases filtering module. As illustrated in Fig. J, our method can successfully detect cases with artifacts on DexYCB and HO3D.

Part 4: Limitations and Future Work

Limitations. As our approach focuses mainly on generating data for single-frame 3D hand-mesh reconstruction, the consistency of hand appearance is not guaranteed across the existing data and generated data, limiting its applications on multi-frame and multi-view 3D hand-mesh reconstructions. Besides, though our method can generate images containing objects, the object type is limited by the existing dataset.

Future Work. Improving consistency during data generation has been studied in several 2D and 3D methods [3, 4, 8]. These works inspire us to utilize and design techniques to alleviate these limitations, *e.g.*, we can adopt our task-oriented designs with DDNM [7] to promote hand-object appearance consistency across different views and grasping poses, enabling us to generate multi-frame/view hand-object interaction images for downstream applications.

References

[1] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, pages 20544–20554, 2022. 1, 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[3] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *ICCV*, pages 9298–9309, 2023. 2

[4] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman.

StyleSDF: High-resolution 3D-consistent image and geometry generation. In *CVPR*, pages 13503–13513, 2022. 2

[5] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022. 1, 2

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 1

[7] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *ICLR*, 2023. 2

[8] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-HD: 3D-consistent image generation at high resolution with generative radiance manifolds. In *ICCV*, pages 2195–2205, 2023. 2

[9] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *CVPR*, pages 17048–17058, 2023. 1, 2

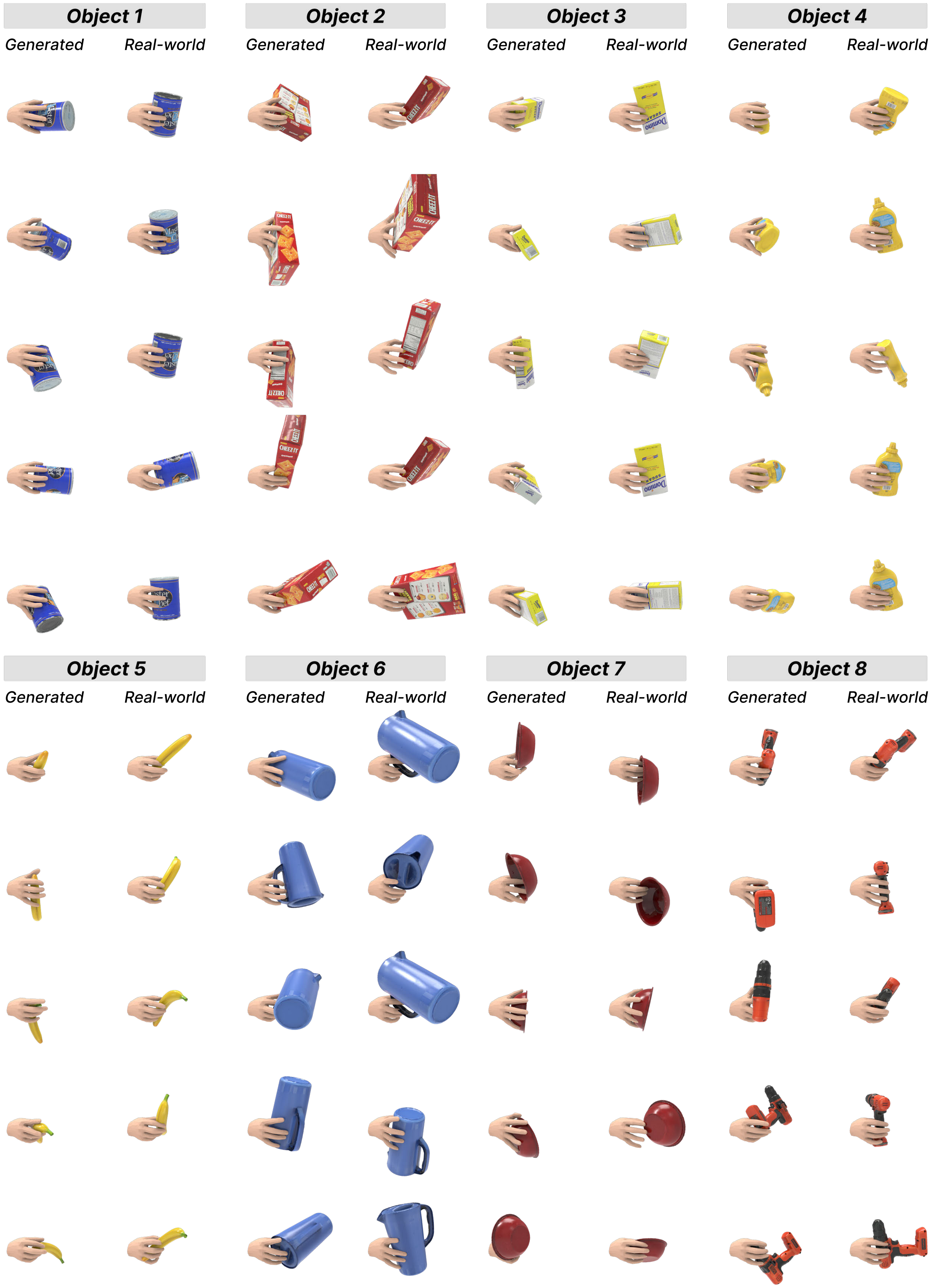

Figure D. Visualizations of the nearest grasping pose retrievals. The nearest real-world poses are not similar to our generated ones, reflecting the novelty of the poses generated by our approach.

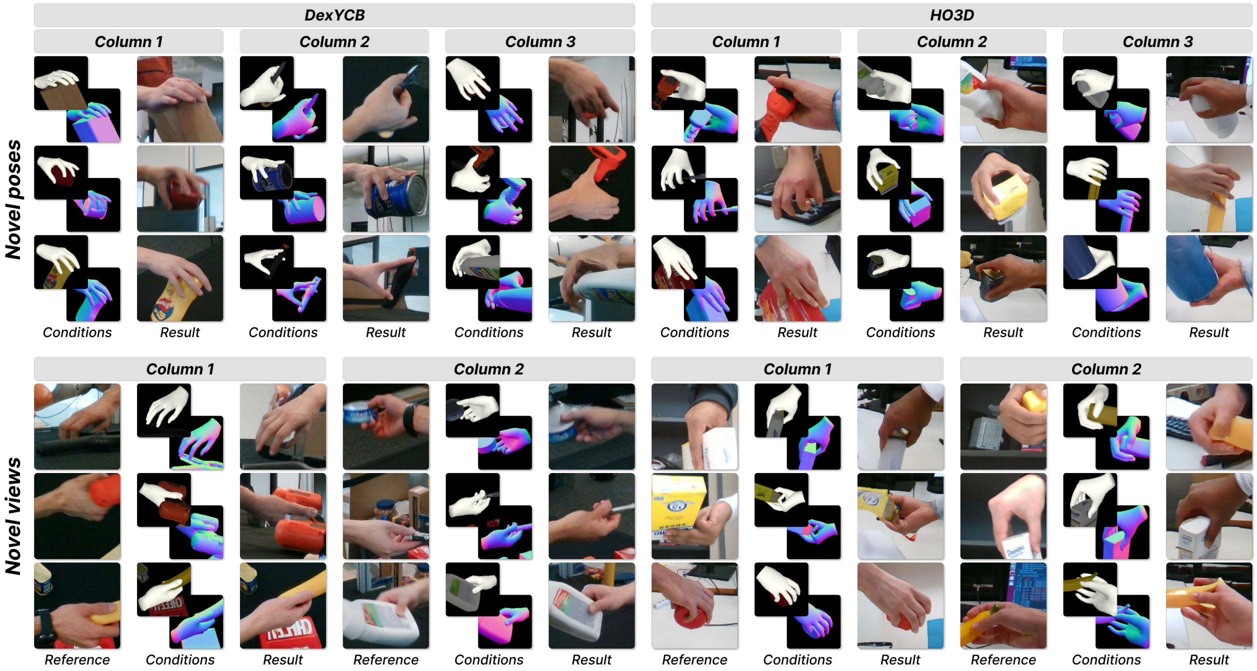

Figure E. Generated examples for novel views/poses with realistic hand appearance and backgrounds.

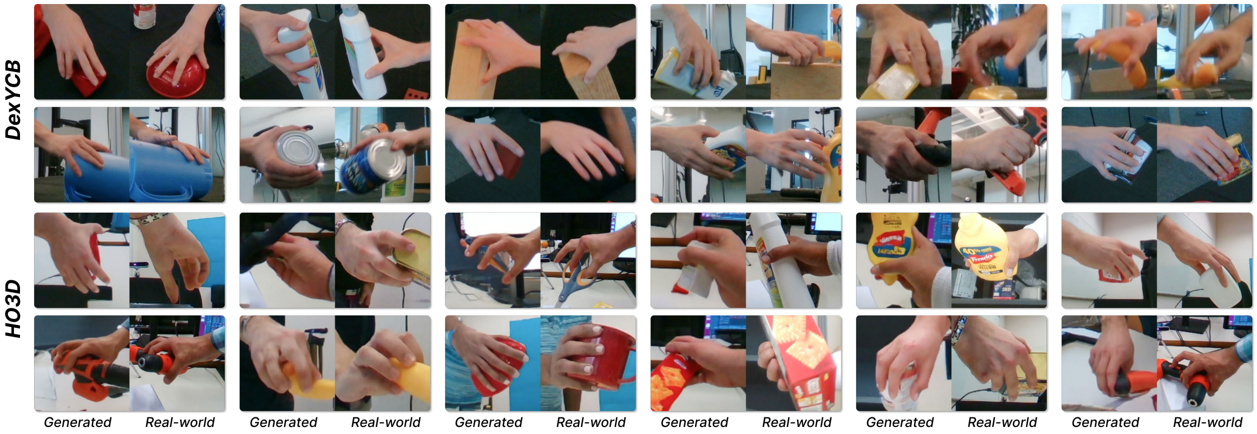

Figure F. Visualizations of the nearest samples retrieved, between our generated synthetic images and the real-world images. The nearest real-world images are not similar to our generated ones, reflecting the novelty of the images generated by our approach.

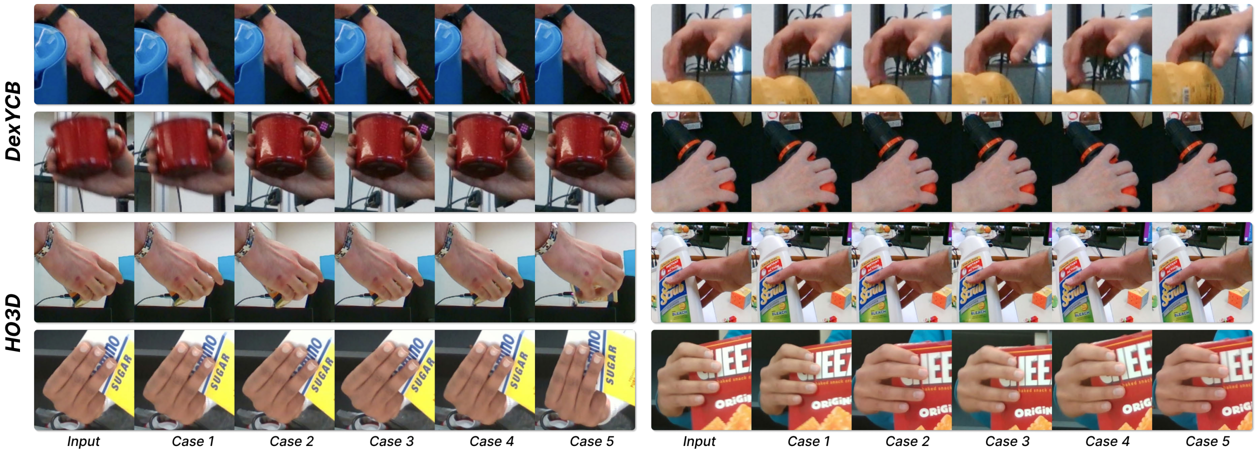

Figure G. Visualizations of the nearest samples retrieving among the real-world images. The feature-space retrieving approach finds similar cases successfully given the input, demonstrating that the retrieval mechanism can effectively find similar samples in the training set.

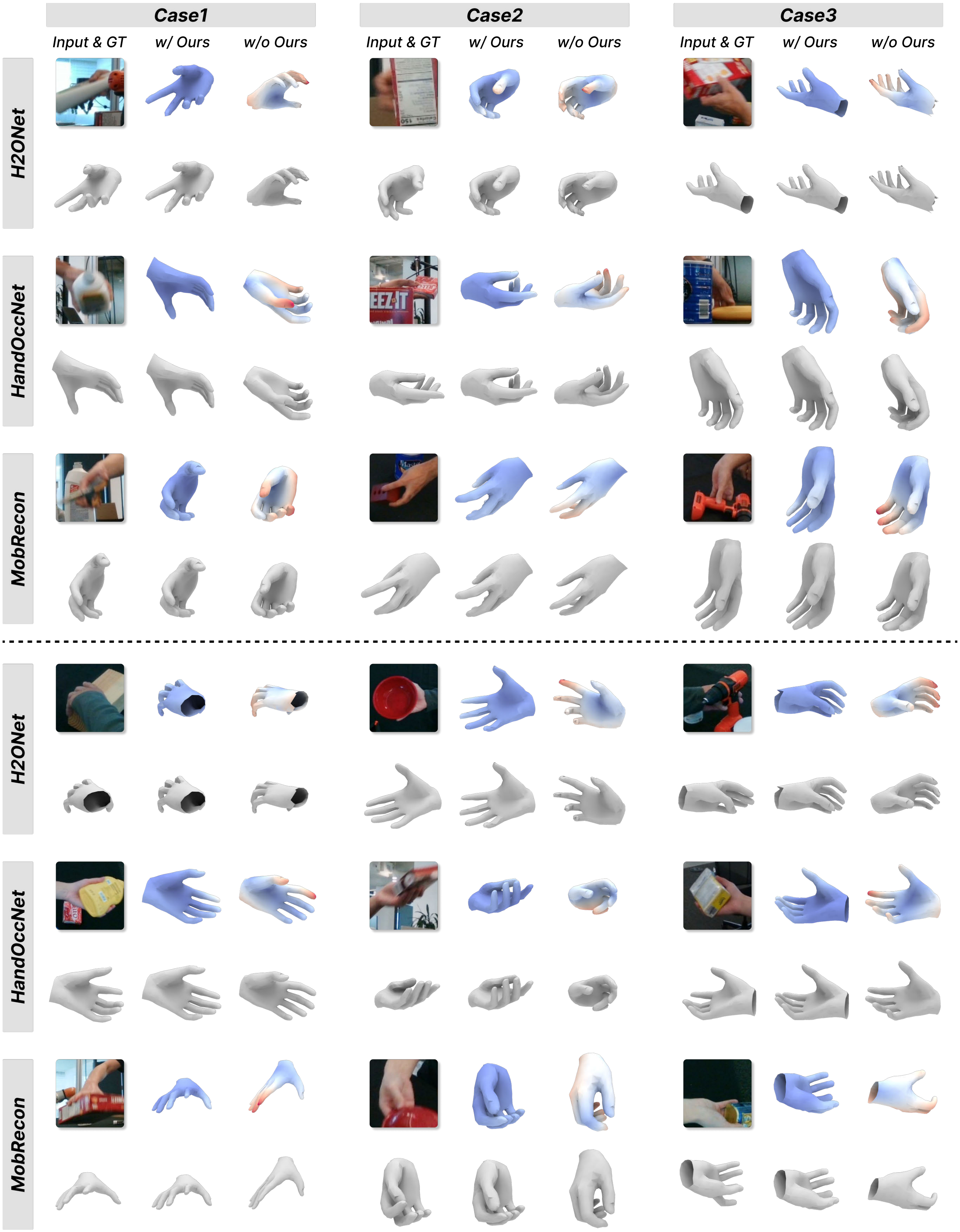

Figure H. More qualitative comparison on DexYCB (top: S0 split, bottom: S1 split).

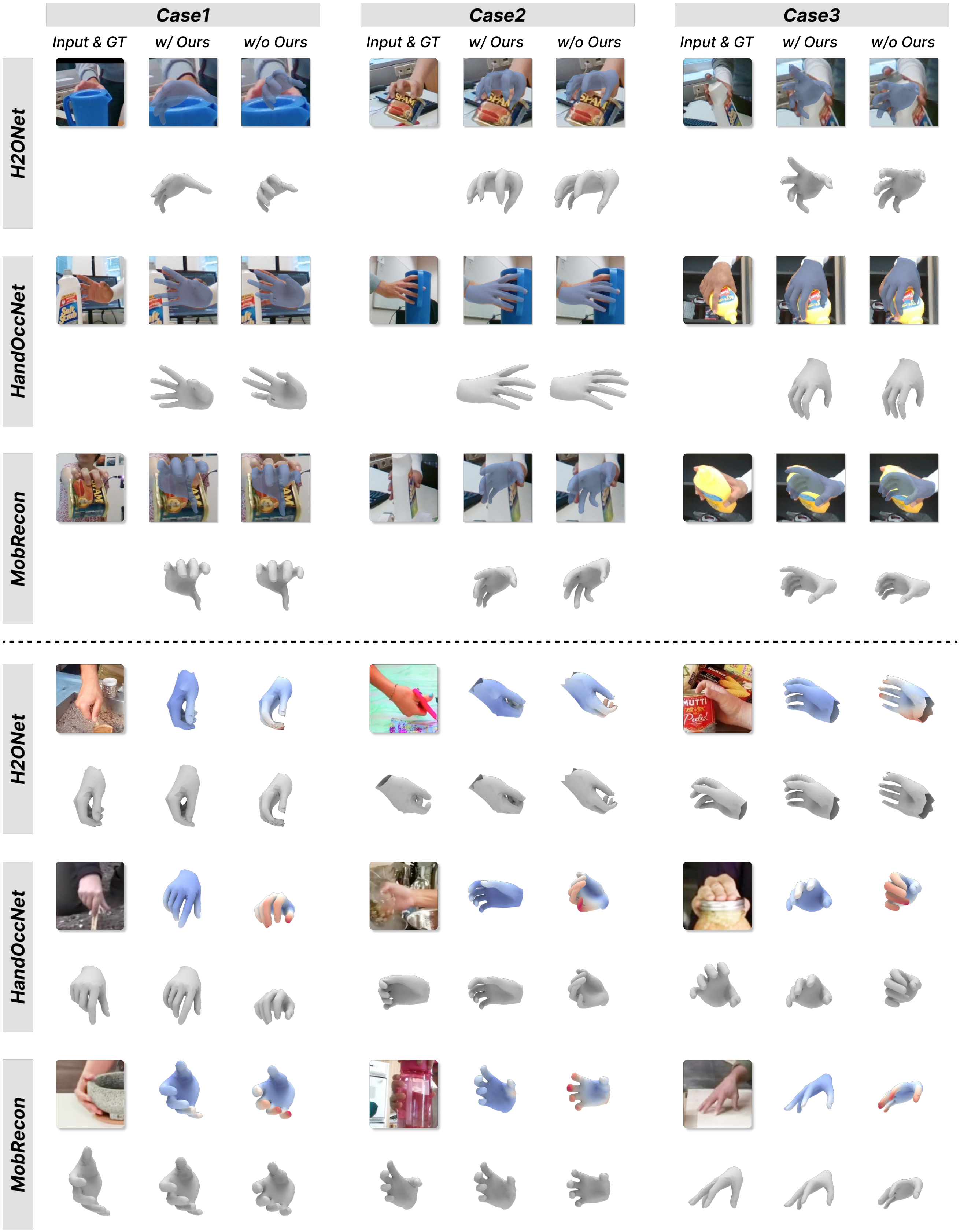

Figure I. More qualitative comparison on HO3D (top) and MOW (bottom).

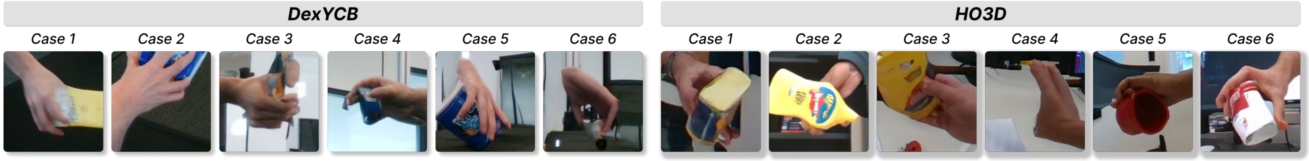

Figure J. Example randomly-picked edge cases.