Supplementary Material of D-SCo: Dual-Stream Conditional Diffusion for Monocular Hand-Held Object Reconstruction

A Architecture of Dual-Stream Denoiser

In this section, we provide the architectural details of our proposed dual-stream denoiser. The dual-stream denoiser is built upon Point-Voxel CNN (PVCNN) [8], which separately processes the point cloud in two branches, *i.e.* a point branch and a voxel branch. Specifically, for the point branch, a simple multi-layer perceptron (MLP) is applied to extract fine-grained per-point features. For the voxel branch, the point cloud is first normalized and then voxelized. The resulting voxels are then processed by a 3D U-Net for feature aggregation. As shown in Fig. A.1, we leverage four Set Abstraction (SA) layers and four Feature Propagation (FP) layers for downsampling and upsampling of the input features with the aforementioned Point-Voxel Convolution (PVConv) as the shared architecture of f_θ^1 and f_θ^2. Eventually, we employ an MLP to estimate the noise as ϵ_θ.

Noteworthy, only the first N channels of $\mathcal{F}_\theta^1$ are utilized for concatenation since the semantic features of object and hand have been integrated by f_θ^1. We use $S = 64$ in our experiments.

B Implementation Details

We implement our model in PyTorch [10] and use PyTorch3D [11] for rasterization. Our model is trained for 300K steps with a batch size of 24 on ObMan and 100K for DexYCB. We finetune 14K and 10K steps for HO3D and MOW datasets, respectively. We use a resolution of size 224 $\times$ 224 and sample 16,384 points for each object during training and inference. We utilize AdamW [7] optimizer with a base learning rate of 10^{-3} and a cosine decay schedule [9]. As

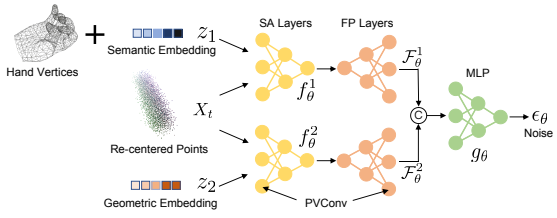

Fig. A.1: Architecture of dual-stream denoiser.

Table C.1: Jointly hand-object finetuning on ObMan.

	Object		Hand	
	F-5 ↑	F-10 ↑	F-2 ↑	F-5 ↑
Ours	0.61	0.81	0.20	0.75
Ours w/ finetuning	0.62	0.81	0.21	0.77

in DDPM [6], a linear variance schedule with beta increasing from $\beta_0 = 10^{-5}$ to $\beta_T = 0.008$ is employed in the underlying diffusion model. During the reverse process, we run 1,000 denoising steps for each object. We set $\eta_1 = 0.2$ and $\lambda_1 = \lambda_2 = 0.2$ in our experiments. All experiments are performed on a single NVIDIA A100 with 40GB GPU memory.

C Jointly Hand-Object Finetuning

Note that for fairness we primarily focus on object reconstruction conditioned on an estimated hand pose in the very same setting as related work such as iHOI [13] and DDF-HO [14]. This includes that we use the same hand poses for conditioning. However, also note that the initial hand poses, despite being reliable, can be yet noisy. Thus, to give a more complete picture, we also conducted experiments for jointly optimizing object shape and hand pose, adding a hand shape/pose loss. To this end, aside of our projective 2D object loss, we also project the hand in the very same way and employ the L1 loss between them and ground truth following

$$\mathcal{L}_{hand} = \|\mathcal{R}(X_H) - \mathcal{R}(\hat{X}_H)\|_1, \tag{1}$$

where $\hat{\bullet}$ denotes predicted results. The overall finetuning loss is then a weighted sum of all terms with

$$\mathcal{L}_{finetune} = \mathcal{L}_{denoise} + \eta_1 \mathcal{L}_{mask} + \eta_2 \mathcal{L}_{hand}, \tag{2}$$

where $\eta_2 = 0.2$ is a hyperparameter that controls the strength of jointly hand-object optimization.

In Tab. C.1, we present the F-scores for object and hand after 10K steps of finetuning. It can be seen that both the object shape and hand pose precision benefit from finetuning using such a projective hand loss.

D Centroid Prediction Ablation

In Tab. D.2, we demonstrate the effectiveness of our proposed hand-constrained centroid prediction. Aided by the hand vertices constraint, the mean centroid prediction error undergoes a noticeable decline of 13.0% with respect to the mean 3D error and 23.1% for the mean 2D error. The hand-constrained centroid prediction is responsible for providing the hand prior to the diffusion model,

Table D.2: Ablation study on centroid prediction. We show the 3D mean error (mm) in the hand wrist coordinate system and the 2D mean error in the NDC space on ObMan dataset.

	3D Error ↓	2D Error ↓
Ours w/o hand constraint	0.023	0.013
Ours	**0.020**	**0.010**

Table E.3: Ablation study on ObMan [5] and HO3D [4] datasets.

Row	Method	ObMan			HO3D		
		F-5 ↑	F-10 ↑	CD ↓	F-5 ↑	F-10 ↑	CD ↓
A0	Ours	0.61	0.81	0.11	0.41	0.63	0.34
A1	Ours Oracle	0.67	0.86	0.09	0.51	0.76	0.23
B0	A0 → w/o $\mathcal{L}_{mask}$	0.57	0.76	0.23	0.36	0.56	0.61
C0	B0 → w/o dual-stream denoiser	0.54	0.74	0.27	0.34	0.53	0.76
D0	C0 → w/o X_t^{HO} & X_t^A	0.48	0.67	0.41	0.28	0.46	0.96
D1	C0 → w/o X_t^{HO}	0.51	0.69	0.37	0.33	0.50	0.81
D2	C0 → w/o X_t^A	0.51	0.69	0.38	0.30	0.48	0.89
D3	C0 → w/ GCN hand embedding	0.52	0.71	0.30	0.34	0.53	0.82
D4	C0 → w/ global hand embedding	0.52	0.71	0.30	0.31	0.49	0.86
E0	D0 → w/o centroid fixing	0.44	0.61	0.65	0.27	0.45	1.00
E1	D0 → w/o centroid prediction network	0.32	0.45	2.48	0.23	0.36	1.31
F0	E0 → Test with GT object centroid	0.45	0.67	0.36	0.29	0.47	0.93
F1	E0 → Test with GT object pose	0.50	0.70	0.34	0.31	0.49	0.84
G0	A0 → Predicted hand pose + noise $\sigma = 0.1$	0.61	0.81	0.11	0.40	0.61	0.36
G1	A0 → Predicted hand pose + noise $\sigma = 0.5$	0.57	0.77	0.13	0.37	0.58	0.43
H0	A0 → GT hand pose	0.65	0.84	0.10	0.43	0.65	0.31
H1	A0 → GT hand pose + noise $\sigma = 0.1$	0.63	0.83	0.11	0.41	0.63	0.33
H2	A0 → GT hand pose + noise $\sigma = 0.5$	0.59	0.79	0.13	0.36	0.60	0.36
I0	F0 → GT object centroid + noise $\sigma = 0.1$	0.44	0.65	0.34	0.28	0.46	1.00
I1	F0 → GT object centroid + noise $\sigma = 0.5$	0.41	0.60	0.50	0.24	0.42	1.23

which is a crucial part of our centroid fixing scheme. Due to the precise centroid prediction, our diffusion model can essentially better focus on mere shape reconstruction.

E Robustness against hand pose and object centroid prediction quality

To demonstrate that our method is able to deal with noisy hand pose estimates, we add various levels of Gaussian noise to predicted hand joints, where σ denotes the variance of Gaussian noise (Tab. E.3). Notice that adding $\sigma = 0.1$ noise has almost no effect on the performance (G0 *vs*. A0) for both ObMan and HO3D. Further, although there is a slight decrease in performance with $\sigma = 0.5$ noise (G1 *vs*. A0), our results well indicate that our approach is robust towards weak hand pose predictions. Noteworthy, in very rare cases, the significant error in

Fig. F.2: Visualization of oracle experiments. We show the input images (first column), the ground-truth shapes (second column), and 4 reconstruction results (last 4 columns). We visualize the ground-truth shapes and the reconstruction results in the camera view (first row) and a novel view (second row).

predicted hand pose may also lead to failure (See Appendix G). Additionally, in H0 we explore the upper bound of our method using ground-truth hand poses, and in H1 / H2 we further demonstrate the robustness against noisy hand pose.

We also add different levels of Gaussian noise to the ground-truth object centroid (I0 / I1) to constitute the robustness of our diffusion model towards inaccurate object centroids.

F Qualitative Results

Qualitative Results for Oracle Experiments. In Fig. F.2, we present qualitative results for the oracle experiments. Thanks to the probabilistic formulation in diffusion models, our approach is able to generate multiple plausible shapes, demonstrating our capability of dealing with the uncertainty induced by hand- and self-occlusion.

Additional Qualitative Results of HO3D, MOW, ObMan and DexYCB. We provide additional qualitative results of D-SCo for HO3D in Fig. F.3, for MOW in Fig. F.4, for ObMan in Fig. F.5, and for DexYCB in Fig. F.6. While SDF- and DDF-based methods, including iHOI [13], gSDF [3], and DDF-HO [14], tend to result in either over-smoothed or distorted and fragmented reconstructions, our approach is capable of generating geometrically coherent point clouds with very plausible details, even for thin objects and heavily occluded parts. Further, due to the extraction of semantic and geometric hand features, our approach shows particularly strong performance in scenarios where objects are

highly occluded (Row 3 in Fig. F.6) or truncated (Row 3 in Fig. F.5, Row 1 in Fig. F.6). This again illustrates our ability to well infer invisible areas of the object.

Qualitative Results of Ablation Study. We demonstrate the qualitative results for our conducted ablation study in Fig. F.7. It can be seen that without the proposed dual-stream denoiser, the diffusion model is having difficulties making proper use of the semantic and geometric hand prior, leading to inferior results. Further, without the proposed centroid fixing scheme, the semantic feature projection is inaccurate and unstable, which again leads to worse reconstructions. Note that the two samples both undergo severe occlusion, having only 40% of the object visible. Due to our proper modeling of the hand-object interaction, D-SCo proves great robustness against hand- as well as self-occlusion.

G Failure Cases and Limitations

As shown in Fig. G.8, the failure cases are primarily provoked by two reasons. 1) When multiple objects are present in an image without contacting with the hand, the model can be confused about which object should be modeled (Row 1). Nonetheless, within the same video sequence, the model is able to make accurate predictions when the object comes into contact with the hand (Fig. F.6 (Row 5)). 2) Although our approach shows strong robustness against noisy hand poses, the model may predict a less-detailed shape for extreme cases. In Fig. G.8 (Row 2), we visualize our reconstruction results along with the hand. In this very rare case, there is a significant error in predicted hand pose as provided by [5], leading to a subpar object reconstruction (Middle). Note that, as aforementioned, we focus on hand-held object reconstruction conditioned on an estimated hand pose under the very same setting as iHOI [13] and DDF-HO [14]. When instead using the ground-truth hand pose as condition, our approach again produces reasonable results (Right).

Though the unstructured and order-agnostic nature of point clouds naturally suits the highly flexible nature of diffusion models, the point cloud representation may face the surface reconstruction problem in downstream tasks. Finally, D-SCo inherits common drawbacks of diffusion models such as a typical slow inference. In particular, a single reconstruction requires $\sim$ 310s for sampling 1,000 steps with a batch size of 24. To further improve the method, we could speed up inference time with techniques such as DDIM [12] sampling.

H Ethics Statement

We evaluate our approach on four publicly available datasets ObMan [5], HO3D [4], MOW [1] and DexYCB [2]. The real-world HO3D, MOW, and DexYCB datasets are well-designed, containing well-balanced samples of different skin colors. The synthetic ObMan dataset also possesses diverse skin colors. In particular, the employed rendering process ensures a considerable proportion of diverse colors

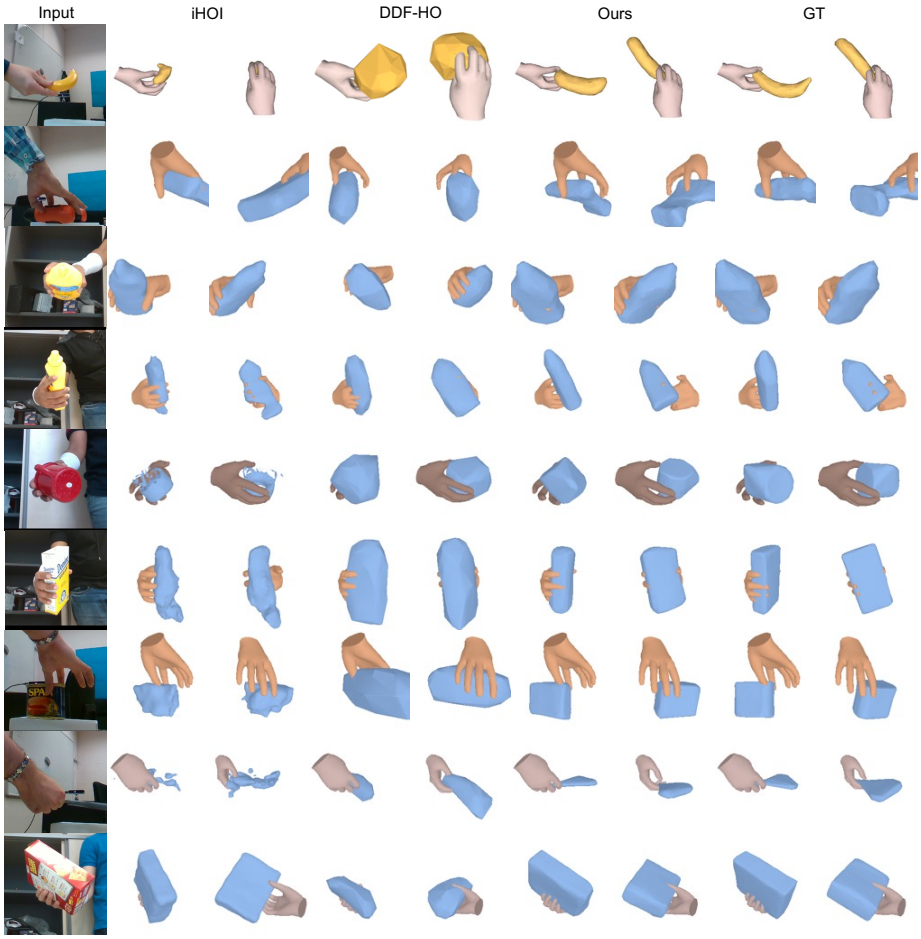

Fig. F.3: Additional qualitative results on the HO3D [4] dataset. For the reconstruction results of each method and the ground-truth shapes, we visualize in the camera view (left) and a novel view (right).

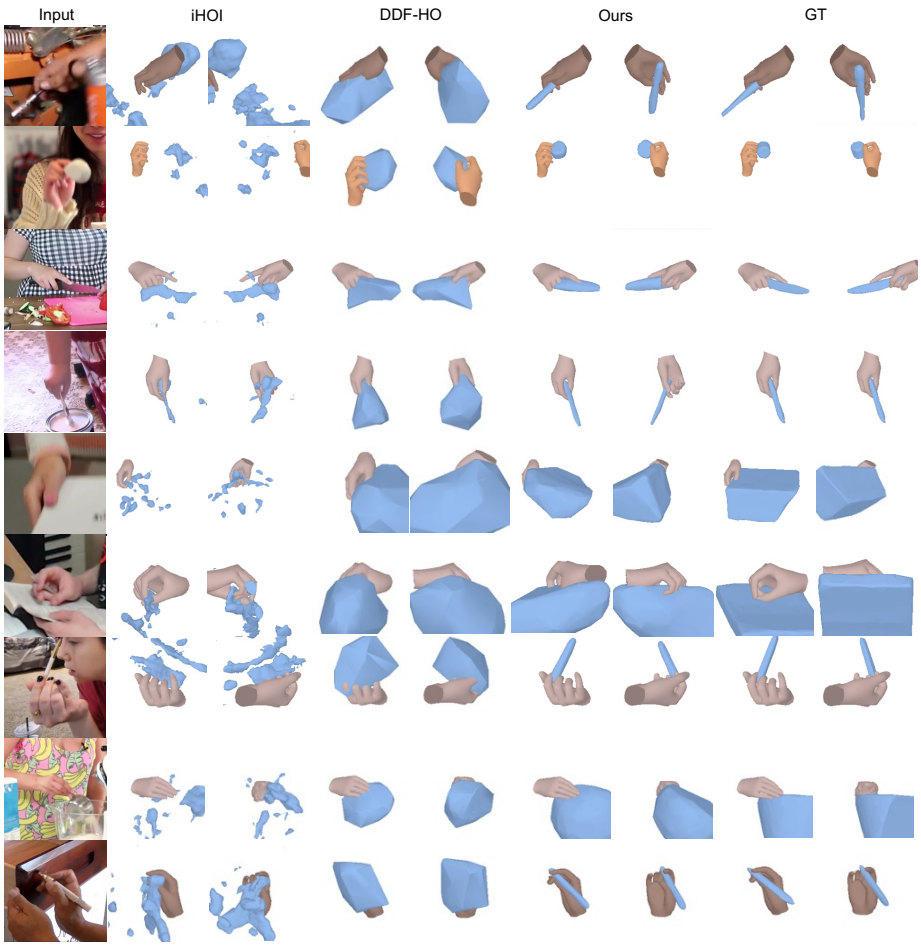

Fig. F.4: Additional qualitative results on the MOW [1] dataset. For the reconstruction results of each method and the ground-truth shapes, we visualize in the camera view (left) and a novel view (right).

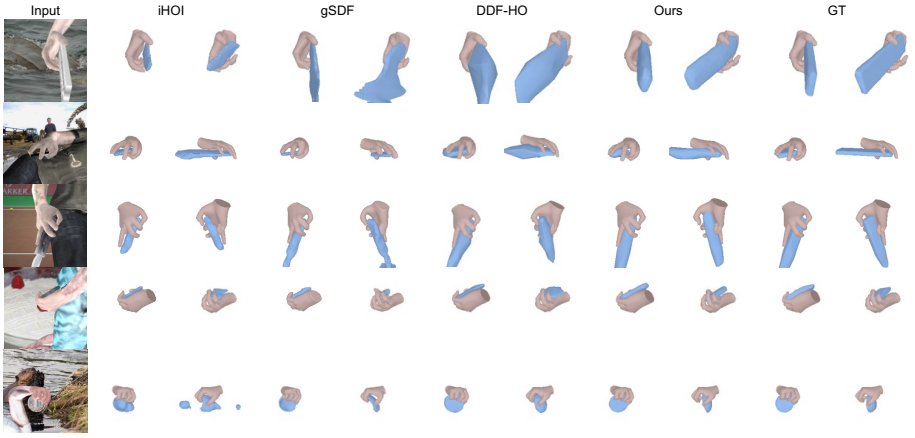

Fig. F.5: Additional qualitative results on the ObMan [5] dataset. For the reconstruction results of each method and the ground-truth shapes, we visualize in the camera view (left) and a novel view (right).

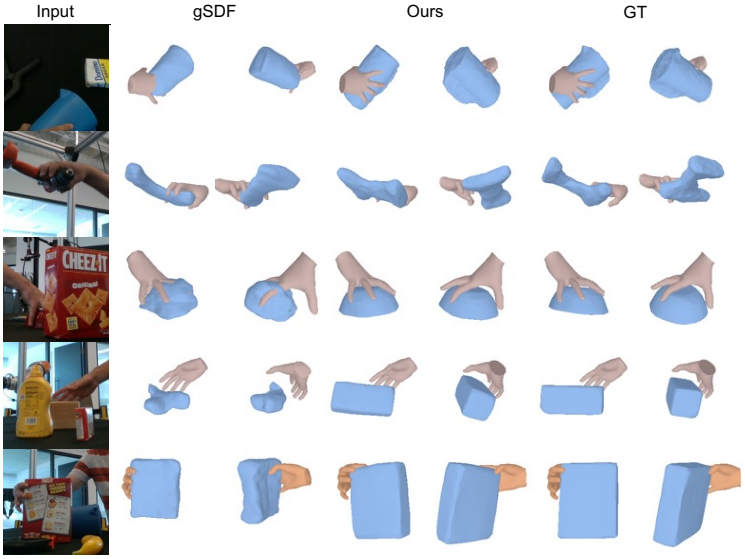

Fig. F.6: Additional qualitative results on the DexYCB [2] dataset. For the reconstruction results of each method and the ground-truth shapes, we visualize in the camera view (left) and a novel view (right).

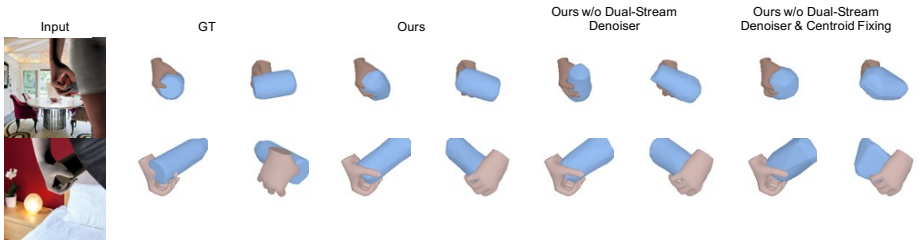

Fig. F.7: Qualitative results of ablation study. We show the ground-truth shapes and our reconstruction results w/ or w/o dual-stream denoiser and centroid fixing.

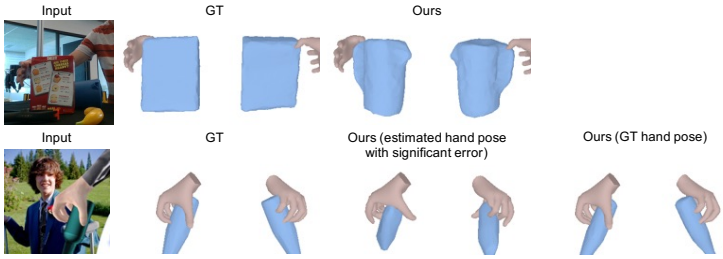

Fig. G.8: Visualization of failure cases. We show the ground-truth shapes and our reconstruction results in the camera view (left) and a novel view (right).

for the hands. Moreover, D-SCo adopts an off-the-shelf hand pose estimator, fully focusing on the hand-held object reconstruction.

Potential negative societal impact. Although the publicly available datasets we use have considered the diversities of scenes, objects, and persons, there remain potential biases and underrepresentation. Additionally, substantial computing power and energy consumption required during training and inference stages may have a negative impact on the environment.

References

1. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV. pp. 12417–12426 (2021)
2. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., Kautz, J., Fox, D.: DexYCB: A benchmark for capturing hand grasping of objects. In: CVPR (2021)
3. Chen, Z., Chen, S., Schmid, C., Laptev, I.: gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In: CVPR. pp. 12890–12900 (2023)
4. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR. pp. 3196–3206 (2020)

5. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR. pp. 11807–11816 (2019)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
8. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. NeurIPS **32** (2019)
9. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2016)
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
11. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 (2020)
12. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
13. Ye, Y., Gupta, A., Tulsiani, S.: What's in your hands? 3D reconstruction of generic objects in hands. In: CVPR. pp. 3895–3905 (June 2022)
14. Zhang, C., Di, Y., Zhang, R., Zhai, G., Manhardt, F., Tombari, F., Ji, X.: Ddf-ho: Hand-held object reconstruction via conditional directed distance field. NeurIPS **36** (2024)