Template Free Reconstruction of Human-object Interaction with Procedural Interaction Generation

Supplementary Material

In this supplementary, we first discuss in more detail about our implementation for the ProciGen and HDM in Appendix A. We also present the statistics of our generated ProciGen dataset. We then show more results and experimental analysis of our method in Appendix B. We conclude with a discussion of limitations and future works.

A. Implementation Details

We describe in more details of the implementation of our ProciGen and HDM. Our code for both data generation and reconstruction will be made publicly available.

A.1. ProciGen Data Generation

Correspondence estimation. We use the implementation from ART [115] for our autoencoder, which uses PointNet [66] as encoder and 3-layer MLPs as decoder. We sample 8000 points from the mesh surface and train the network with bidirectional Chamfer distance. To ensure reconstruction quality, we overfit one network per category. Each model is trained for 5000 epochs. We report an average reconstruction error of around 7mm for our autoencoders, which indicates highly accurate reconstructions.

Contact transfer and optimization. We use a threshold of $\sigma = 2$cm to find points that are in contact. The loss weights for our contact based loss optimization are: $\lambda_c = 400, \lambda_n = 6.25, \lambda_{\text{colli}} = 9, \lambda_{\text{init}} = 6.25 \cdot 10^4$.

Rendering. We use blender to render our synthesized human-object interactions. We choose one set of 4 camera configurations from BEHAVE [7] and another set of 6 camera configurations from InterCap [35]. For each synthesized interaction, we additionally add small random global rotation and translation to have variations of camera viewpoints. We render the interactions with an empty background since our network also takes images with background masked out as input. We add lights at fixed locations with random light intensities. Our blender scene and rendering code will also be made publicly available.

A.2. HDM: Hierarachical Diffusion Model

We use the modified Point Voxel CNN from [117] as the network for our joint diffusion p_θ, segmentation g_ϕ, and separate diffusion models p^h_θ, p^o_θ. The input images are cropped and resized to 224×224. The joint diffusion model diffuses in total 16384 points while the separate models diffuse 8196 points each. We use the MAE [30] as the image feature encoder. We additionally stack the human and object masks as well as distance transform as additional image features, same as PC^2 [60]. We train our diffusion models for a total of 500000 steps with batch size 20. We use a linear scheduler without warm-up for the forward diffusion process, in which beta increases from $1 \cdot 10^{-5}$ to $8 \cdot 10^{-3}$. For the network optimization, we use AdamW optimizer with linear learning rate decay starting from $3 \cdot 10^{-4}$ and decreasing to 0 during the course of training. The diffusion models are trained with the standard diffusion training scheme [32]. To train the segmentation model, we add small Gaussian noise to the GT point clouds and project them to obtain image features. The loss is then computed between the prediction and recomputed GT labels on the points with noise. To speed up training, we train stage 1 (g_ϕ, p_θ) and stage 2 (p^h_θ, p^o_θ) models separately. For each stage, it takes around 4 days to train on a machine with 4 A40 GPUs.

Camera estimation. Recall from Sec. 3.2.3 that a camera translation is required to project the normalized point clouds back to the image. This needs to be estimated from input when GT camera pose is not available, especially for generalization to diverse datasets. The camera translation consists of three unknowns, which requires at least two point pairs of 3D location and 2D-pixel coordinates. We empirically choose the Gaussian point center and one edge of the point cloud. The idea is to have the initial Gaussian point clouds cover the 2D human object interaction region and the 3D center is projected to 2D crop center.

Formally, let $\mathbf{p}_c = (c_x, c_y)$ be the center coordinate of the 2D interaction region, w be the width of the 2D interaction square crop, $\mathbf{p}_e = (\sigma, 0, z)$ be a 3D point near the edge of the Gaussian sphere with unknown depth z. Given camera projection matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t}_c$, we define the following equations:

$$\mathbf{K}\mathbf{t}_c = \mathbf{p}_c; \;\; \mathbf{K}(\mathbf{p}_e + \mathbf{t}_c) = \mathbf{p}^{\text{2D}}_e \tag{6}$$

The first equation projects origin to $\mathbf{p}_c$ and the second equation projects $\mathbf{p}_e$ to the middle right of the 2D crop $\mathbf{p}^{\text{2D}}_e = (c_x + w/2, c_y)$. This is a linear system of four equations with four unknowns (camera translation and depth z), leading to a unique solution for the translation $\mathbf{t}_c$. We empirically set σ to different values for different categories based on the estimation error on the BEHAVE training set. Furthermore, we compute $\mathbf{p}_c$ as the centroid of all 2D points inside the human and object masks. From Fig. 6, Fig. 17, Fig. 16, Fig. 18, Fig. 19 and , Fig. 20, it can be seen that our method can reconstruct human and object well on different datasets using our estimated translation.

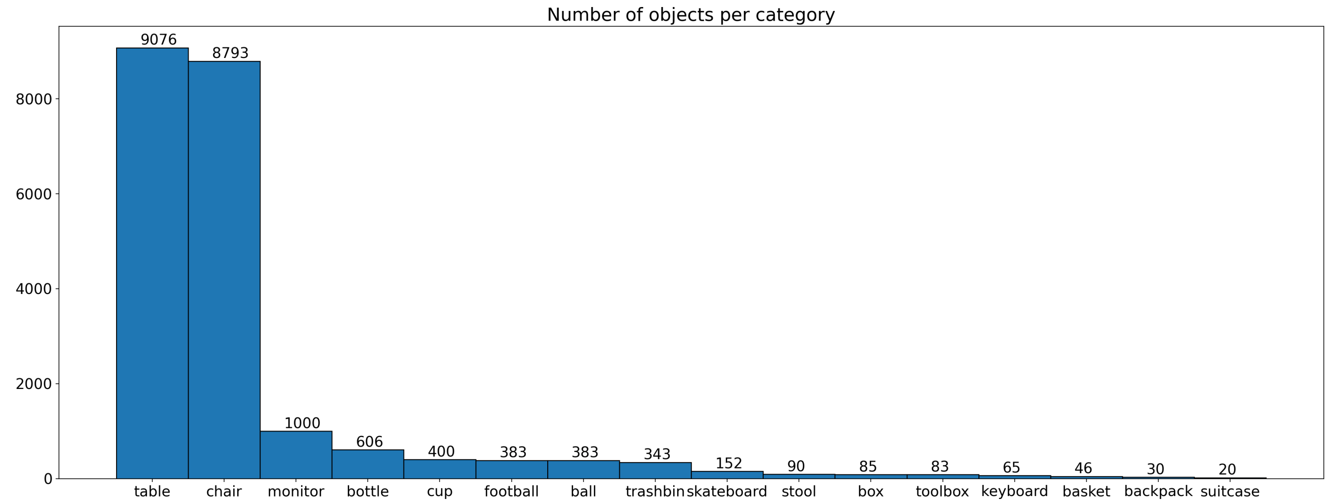

Figure 7. The number of objects per category we used to generate our ProciGen dataset. It can be seen that the shape variations are dominated by tables and chairs, which are also the categories with the most complex shapes.

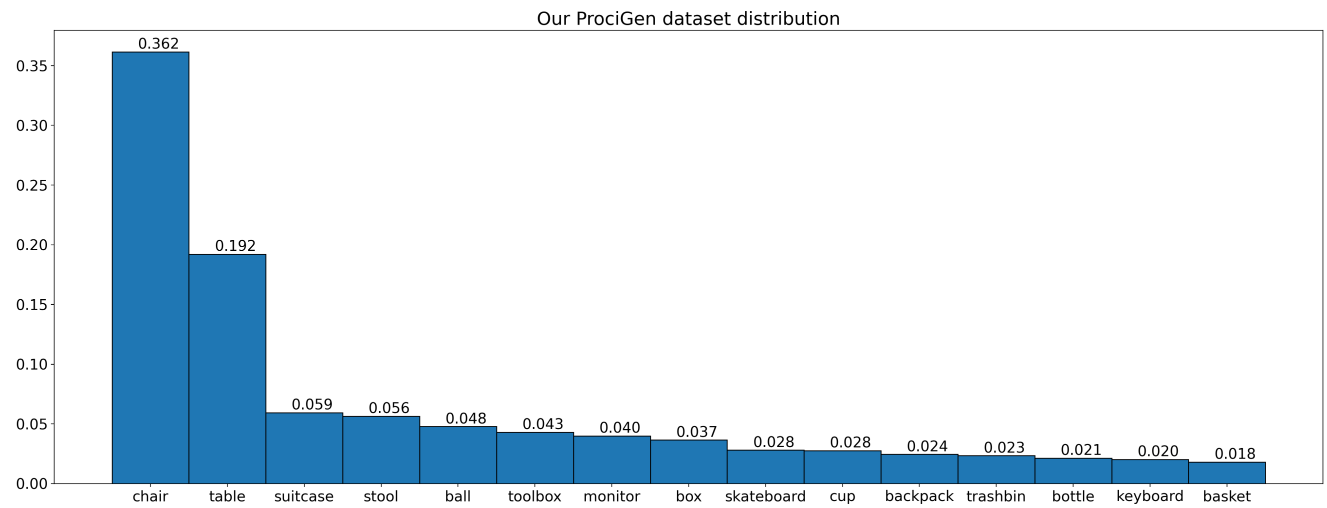

Figure 8. Distribution of interactions per object category. Our dataset features interaction data with very diverse object shapes, which is not possible via real data capture.

A.3. Dataset statistics

We generate our ProciGen dataset based on interactions from BEHAVE [7] and InterCap [35], human scans from MGN [4], object shapes from ShapeNet [12], Objaverse [19] and ABO [15]. When generating our data, we mainly consider the variation of object shapes and interaction poses while the object sizes remain the same. We also try to avoid large imbalances among objects. Therefore, chairs and tables are two dominant categories as they have the most geometry and interaction pose variations (Fig. 7, Fig. 8). Other categories have similar amounts of synthetic data as they have similar amounts of interaction poses. The difference comes from failures in joint optimization due to irregular mesh.

In total we rendered 1.1M interaction images with 21555 different object shapes. The distribution for object shapes and interactions per category are shown in Fig. 7 and Fig. 8. Our dataset has very diverse object shapes, especially for chairs and tables whose geometry also varies a lot in reality. Our procedural generation method is a scalable solution and it allows for generating large-scale interaction datasets with great amount of variations which is not obtainable via capturing real data.

Method	CHORE ↑			PC2 ↑	Ours ↑		
	Hum.	Obj.	Comb.	Comb.	Hum.	Obj.	Comb.
Chair	0.373	0.491	0.443	0.407	**0.384**	**0.521**	**0.463**
Ball	0.330	0.388	0.374	0.424	**0.395**	**0.517**	**0.471**
Backpack	**0.399**	**0.509**	**0.469**	0.436	0.397	0.457	0.444
Table	0.304	0.455	0.389	0.470	**0.379**	**0.642**	**0.517**
Basket	0.301	0.266	0.292	**0.381**	**0.412**	**0.297**	0.364
Box	0.352	0.347	0.362	0.409	**0.414**	**0.401**	**0.424**
Keyboard	0.335	0.412	0.383	0.450	**0.353**	**0.606**	**0.493**
Monitor	0.358	**0.412**	**0.395**	0.377	**0.368**	0.348	0.370
Suitcase	0.400	0.477	0.443	0.404	**0.431**	**0.484**	**0.462**
Stool	0.351	0.479	0.424	0.443	**0.394**	**0.543**	**0.479**
Toolbox	0.281	0.330	0.306	0.398	**0.373**	**0.400**	**0.403**
Trashbin	0.376	0.402	0.398	0.387	**0.407**	**0.414**	**0.422**
BEHAVE all	0.345	0.426	0.397	0.423	**0.392**	**0.498**	**0.457**
Chair	**0.389**	0.468	0.433	0.470	0.384	**0.604**	**0.500**
Cup	0.412	0.538	0.510	0.566	**0.487**	**0.601**	**0.578**
Skateboard	**0.520**	0.684	0.612	0.578	0.491	**0.739**	**0.624**
Bottle	0.426	0.501	0.495	0.592	**0.549**	**0.582**	**0.593**
InterCap all	0.406	0.513	0.469	0.506	**0.440**	**0.607**	**0.534**

Table 6. Per-category F-score@0.01m comparison. Note that PC2 cannot separate human-object hence we only report the combined error, and that CHORE requires template meshes. Our method outperforms baselines for almost all categories.

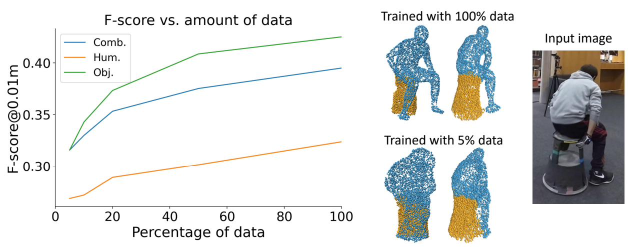

Figure 9. Reconstruction performance vs. amount of data. It can be seen that more data leads to better results.

B. Additional Experiments and Results

B.1. Per-category reconstruction accuracy

We report the accuracy of each category in Tab. 6. Our method consistently outperforms baselines in almost all categories. While the improvements in numbers look small, the visual difference is quite significant, as shown in the paper Fig. 4, Fig. 6.

B.2. Performance vs. amount of data

We show in Table 3 that our data contributes a lot to improve the reconstruction accuracy. To further understand the data contribution, we train our model for the same epochs with different amounts of synthetic data and test on BEHAVE images without fine-tuning. The performance vs. data plot is shown in Fig. 9. More data consistently leads to better performance both quantitatively and qualitatively.

B.3. Analysis of T_0 for our HDM

In our second stage, we first add noise to the clean predictions from stage one until step $t = T_0$, and then run the reverse diffusion process from $t = T_0$ to $t = 0$. We evaluate the performance of our method under different values of T_0 in Figure 10. There is a trade-off for the number of forward steps T_0: with a larger T_0, less interaction information and

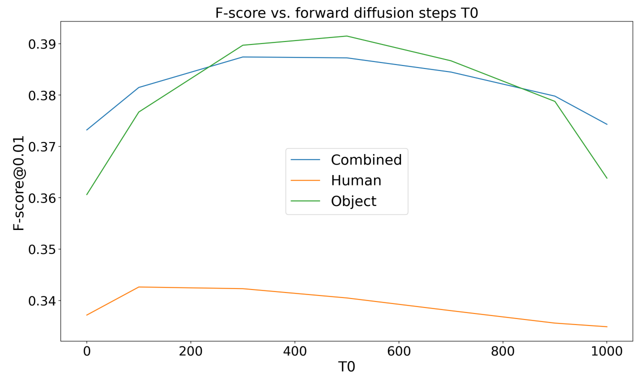

Figure 10. The performance of our method using different intermediate step T_0 for the input to our second stage diffusion. Methods are evaluated using F-score@0.01m. At $T_0 = 500$, we obtain a good balance between human and object performance.

noisy details are preserved and the network predicts sharper detail but less faithful to initial prediction and interaction constraints. It can be seen that $T_0 = 500$ is a good balance between shape fidelity and interaction coherence.

B.4. Shape fidelity

Our method predicts dense and clean point clouds which are ready for accurate surface extraction. We show in Fig. 11 that high-quality meshes can be extracted from our predicted point clouds. More specifically, we use screened Poisson surface reconstruction for the human points using normals estimated by MeshLab. For the object, we first use Delaunay triangulation to obtain triangle mesh. We then run fusion-based waterproofing [77] to obtain a watertight mesh. We also apply Delaunay triangulation and waterproofing to PC^2 [60] predictions and results are shown in Fig. 11. It can be seen that PC^2 predictions have missing structure and noisy point clouds, leading to low-quality meshes. In contract, we can extract high-quality meshes directly from our point cloud reconstructions, without any post processing.

B.5. Interaction semantics

Our method predicts the segmentation of human and object, allowing separate manipulation which is important for downstream applications. To demonstrate this, we use Text2txt [13] to generate textures for the meshes extracted from PC^2 and our predicted point clouds. Other methods such as Paint-it [106] are also applicable here. We show the reconstruction and generated textures in Fig. 13. It can be seen that PC^2 predictions are noisy and it does not reason human and object separately. This leads to low-quality mesh and generating coherent texture for a combined mesh of human and object is difficult. On the contrary, our method separate human and object while also predicting high quality individual shapes. This allows generating high quality texture and changing textures for human and object

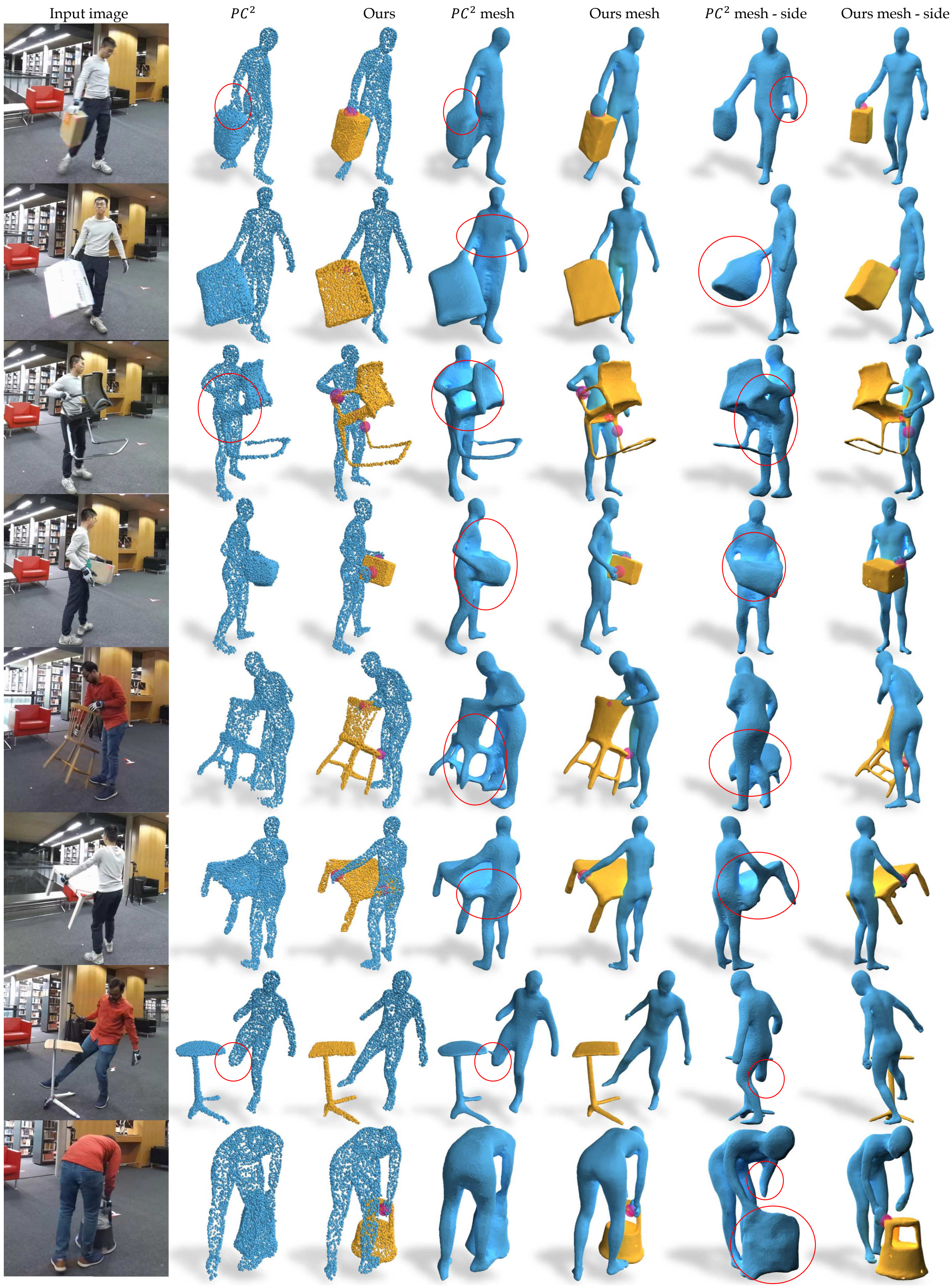

Figure 11. Comparing the shape fidelity of our method with PC^2 on the BEHAVE [7] dataset. PC^2 does not separate human and object and its prediction is noisy, leading to inaccurate meshes. Our method predicts clean point clouds with human object segmentations, allowing us to extract high-quality mesh surfaces.

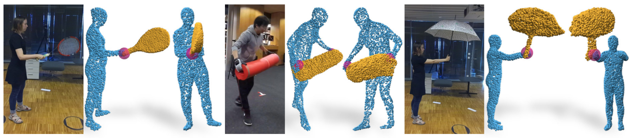

Figure 12. Out of distribution generalization. Our method can reconstruct some categories that are *unseen* in training data.

differently.

B.6. More generalization results

We show more generalization comparison on the InterCap [35] dataset in Fig. 15. Note that all objects from InterCap are unseen during training time. It can be seen that PC^2 trained on BEHAVE [7] only cannot generalize to objects from InterCap. Training PC^2 with our ProciGen dataset allows better generalization ability but its shape prediction is still noisy. Furthermore, PC^2 cannot segment human and object, which is important to reason the interaction semantics and manipulate them separately. Our method generalizes well to InterCap and reconstructs high quality shapes with interaction semantics.

Our method trained *only* on our synthetic ProciGen dataset generalizes well to other datasets. We show results on NTU-RGBD [52], SYSU [33] and challenging in the wild COCO [49] images in figure Fig. 16, Fig. 17 and Fig. 18, Fig. 19, Fig. 20 respectively. Note that our method is trained only on our synthetic ProciGen dataset and not fine-tuned on any images from these datasets. It can be seen that our method generalizes to different datasets with diverse object shapes, without requiring any template meshes.

For quantitative evaluation, we focus on 15 object categories that are seen from our synthetic data (Tab. 6). We test our method on three additional categories from BEHAVE and InterCap that are unseen and have GT data. The F-scores (human/object/combined) are: 0.465/0.453/0.479 (tennis racket), 0.333/0.332/0.361 (yoga mat), 0.375/0.305/0.360 (umbrella), 0.353/0.443/0.420 (all seen categories). We also show example reconstructions in Fig. 12. Our method can reconstruct *unseen* categories.

C. Limitations and Future Work

We present a scalable solution to synthesize large amount of interaction dataset which allows training methods with strong generalization ability. We also propose a model for obtaining high quality human, object shapes and also interaction semantics, without any template shapes. We demonstrate the generalization ability of our method on diverse datasets. Our template-free reconstruction method is a promising first step towards real in-the-wild reconstruction.

Nevertheless, there are still some limitations to the current approach. First, our ProciGen data generation method always starts with a seed interaction pose sampled from an existing interaction dataset. This limits the diversity in terms of interaction poses. Future works can explore generative models such as Object-Popup [64] to further diversify the interaction pose. It is also highly desirable to combine the large human pose variations from AMASS [58], which can further improve the robustness of reconstruction methods to challenging poses.

Secondly, our method struggles to predict accurate human shapes when large chunk of the human body is occluded, see Fig. 14. This is because our method is purely template-free and only use the network to learn the human and object shape priors. Future works can try to further explore human shape or pose constraints to regularize network training and predictions. In addition, our hierarchical diffusion model are designed for human object interaction, which is applicable for general bilateral interaction cases like human-human, hand-hand, and hand-object interactions. However, it cannot handle multi-person or multi-object interactions. We leave these for future works.

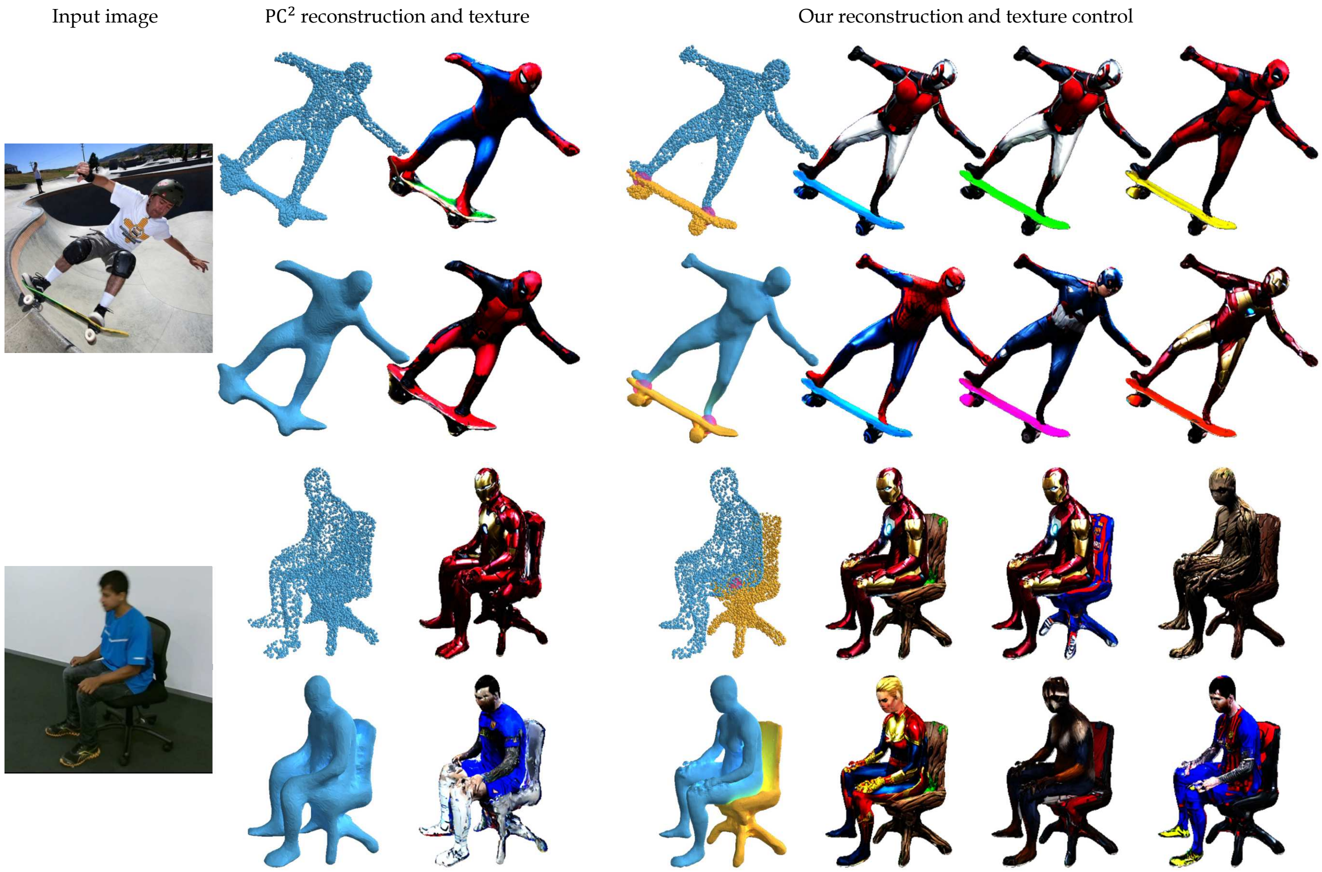

Figure 13. Comparing textures generated for meshes extracted from PC^2 [60] and our predicted point clouds. Textures are obtained using Text2txt [13]. PC^2 predicts human and object as one joint point cloud with noisy points, which leads to inaccurate mesh surfaces and it is difficult to generate textures for this combined mesh. It also does not allow changing human and object textures separately. Our method predicts high quality point clouds with segmentation. This enables us to extract high-fidelity mesh, which is important for generating high-quality texture and manipulating human and object differently.

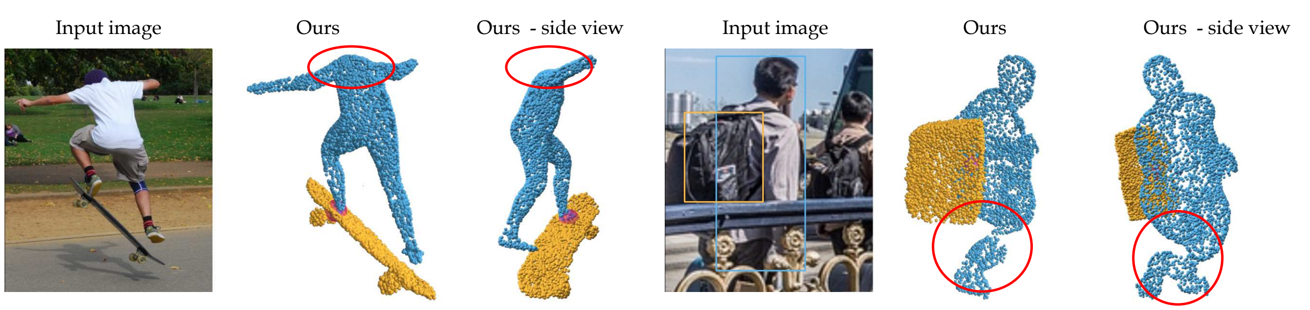

Figure 14. Example failure cases of our method. Our method can fail when large parts of human body are invisible, leading to incoherent human shape reconstructions. Future works can explore human body shape priors to regularize the network predictions.

Figure 15. Comparing generalization performance on InterCap [35]. All objects are unseen during training time. PC^2 trained only on BEHAVE [7] has limited generalization ability. Training PC^2 with our ProciGen improves generalization but it still cannot reason human and object separately and the predicted points are noisy. Our method trained only on our ProciGen generalizes well to InterCap objects even they are completely unseen.

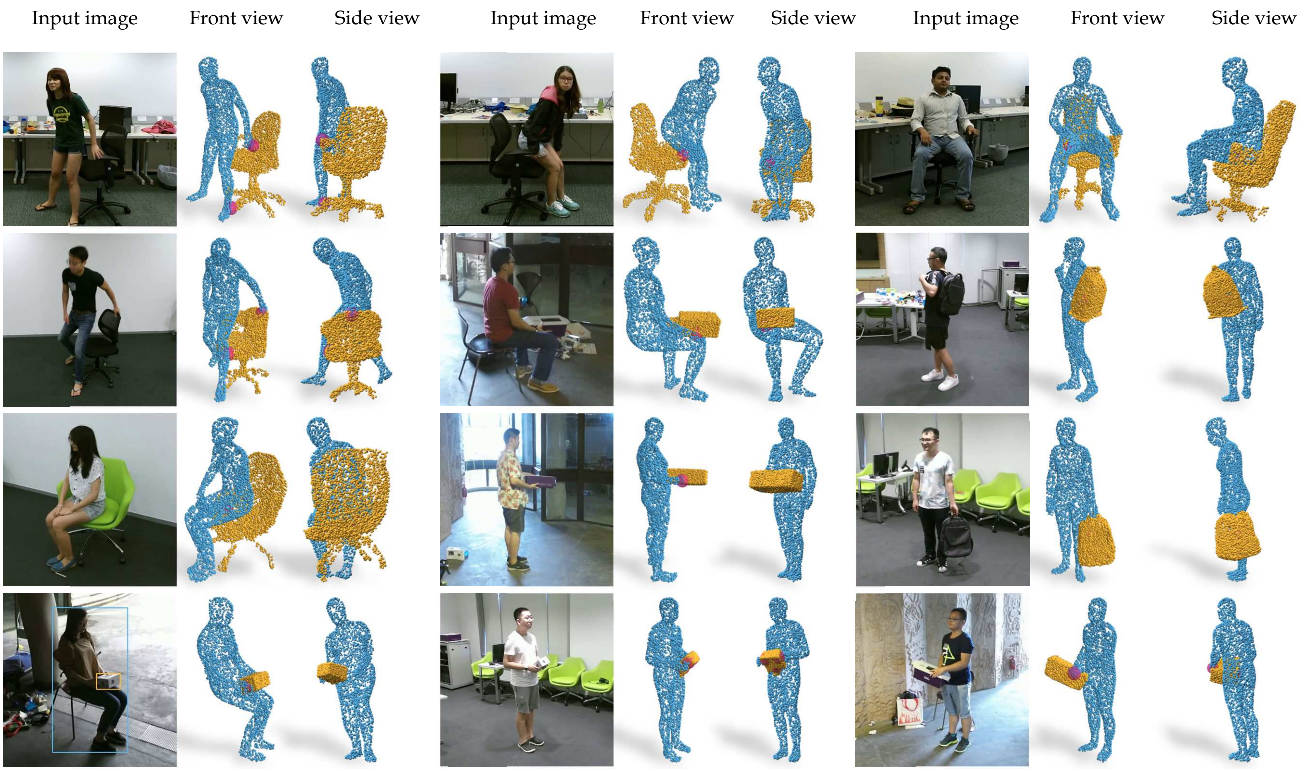

Figure 16. Generalization results on NTU-RGBD [52] dataset. Our method can reconstruct different objects faithfully under various camera viewpoints and lighting conditions, without relying on any template shapes.

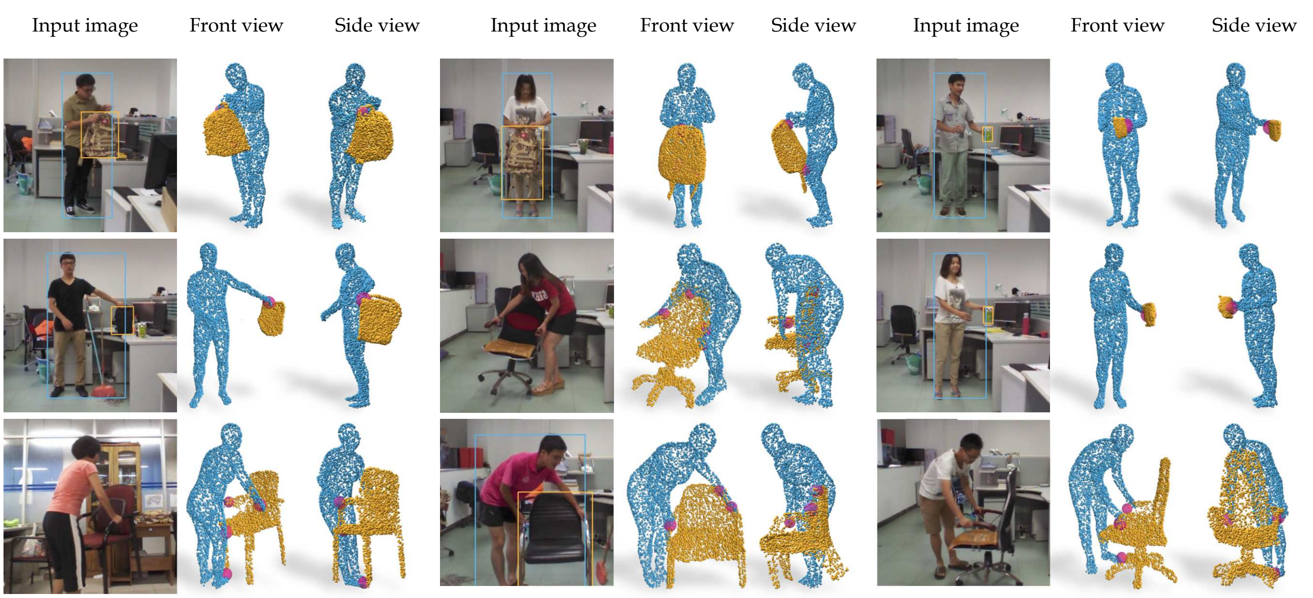

Figure 17. Generalization results on SYSU action [33] dataset. Our method can reconstruct different real-life human and objects during challenging interactions and occlusions.

Figure 18. Generalization results to COCO [49] dataset. Our method can reconstruct high-quality human and object from in the wild images which has very diverse shape variations, without using any template shapes.

Figure 19. Generalization results to COCO [49] dataset. Our method reconstructs diverse object shapes in the wild.

Figure 20. Generalization results to COCO [49] dataset. Our method can reconstruct challenging human and object pose as well as shapes without using any template shapes.

driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 3, 6

References

[1] https://renderpeople.com/. 1, 2

[2] http://virtualhumans.mpi-inf.mpg.de/people.html. 8

[3] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds, 2018. 3

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2, 4, 6

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[7] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 6, 7, 8, 5

[8] Sandika Biswas, Kejie Li, Biplab Banerjee, Subhasis Chaudhuri, and Hamid Rezatofighi. Physically plausible 3d human-scene reconstruction from monocular rgb image using an adversarial learning approach. *IEEE Robotics and Automation Letters*, 8(10):6227–6234, 2023. 2

[9] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 1, 2

[10] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[11] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, 2020. 2

[12] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1, 2, 4, 6

[13] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-

[14] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[15] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 6, 2

[16] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 4

[17] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Gregory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[18] Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 1, 6, 2

[20] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award. 2

[21] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *IEEE Computer Vision and Pattern Recognition*, 2021. 3

[22] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *CVPR*, 2020. 2

[23] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 2

[24] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from

body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 2

[25] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human–object interaction and scene changes from human motion. In *International Conference on 3D Vision (3DV)*, 2024. 2

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

[27] Sookwan Han and Hanbyul Joo. Chorus : Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15835–15846, 2023. 2

[28] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision*, 2019. 2

[29] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, New Orleans, LA, USA, 2022. IEEE. 5, 1

[31] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2, 4, 1

[33] JF Hu, WS Zheng, J Lai, and J Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2186–2200, 2017. 5, 8

[34] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 2

[35] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 1, 2, 6, 7, 8, 5

[36] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2

[37] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 1

[38] Li Jiang, Zetong Yang, Shaoshuai Shi, Vladislav Golyanik, Dengxin Dai, and Bernt Schiele. Self-supervised pretraining with masked shape prediction for 3d scene understanding. In *CVPR*, 2023. 2

[39] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 2

[40] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream, 2023. 2

[41] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[42] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *8th International Conference on 3D Vision*, pages 333–344. IEEE, 2020. 2

[43] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv:2306.01567*, 2023. 4

[44] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[45] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4

[46] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 1

[47] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)*, 2022. 2

[48] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024. 2

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 7, 8, 5, 9, 10, 11

[50] Feng Liu and Xiaoming Liu. Learning implicit functions for topology-varying dense 3d shape correspondence, 2020. 3

[51] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[52] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5, 8

[53] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Eric Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. *arXiv preprint arXiv:2304.12281*, 2023. 2

[54] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund T, Zexiang Xu, and Hao Su. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[55] 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, 2023. 1, 2

[56] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 5

[57] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics*. ACM, 2015. 4

[58] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 5

[59] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 2

[60] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *CVPR*, 2023. 2, 4, 5, 6, 7, 8, 1, 3

[61] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5

[62] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer, 2022. 2

[63] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[64] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[65] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2

[66] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, abs/1612.00593, 2017. 3, 1

[67] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023. 2

[68] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 4

[69] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 1

[70] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 1, 2

[71] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[72] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 2

[73] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2

[74] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision (ECCV)*, pages 516–533, 2022. 2

[75] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 4

[76] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019. 2, 4

[77] David Stutz and Andreas Geiger. Learning 3d shape completion under weak supervision. *CoRR*, abs/1805.07290, 2018. 3

[78] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[79] Ramana Sundararaman, Riccardo Marin, Emanuele Rodola, and Maks Ovsjanikov. Reduced representation of deformation fields for effective non-rigid shape matching. *Advances in Neural Information Processing Systems*, 35, 2022. 3

[80] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[81] Yu Tao, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Dai Quionhai, Hao Li, G. Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2

[82] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? 2019. 6

[83] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[84] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[85] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Posendf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 1

[86] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8001–8013, 2023. 2

[87] Michal J. Tyszkiewicz, P. Fua, and Eduard Trulls. Gecco: Geometrically-conditioned point diffusion models. *ICCV*, abs/2303.05916, 2023. 2

[88] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, 2016. 4

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5

[90] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *International Conference on 3D Vision (3DV)*, 2022. 2

[91] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *CVPR*, 2023. 2

[92] Christopher Wewer, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. Simnp: Learning self-similarity priors between neural points. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[93] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing, 2018. 2

[94] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 1, 2

[95] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 4, 6, 7, 8

[96] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 6

[97] Xianghui Xie, Xi Wang, Nikos Athanasiou, Bharat Lal Bhatnagar, Chun-Hao P. Huang, Kaichun Mo, Hao Chen, Xia Jia, Zerui Zhang, Liangxian Cui, Xiao Lin, Bingqiao Qian, Jie Xiao, Wenfei Yang, Hyeongjin Nam, Daniel Sungho Jung, Kihoon Kim, Kyoung Mu Lee, Otmar Hilliges, and Gerard Pons-Moll. Rhobin challenge: Reconstruction of human object interaction, 2024. 2

[98] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 2

[99] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[100] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 2

[101] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12988, 2023. 2

[102] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng

Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 2

[103] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 2

[104] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling pre-trained vision foundation models. *arXiv preprint arXiv:2303.12786*, 2023. 2

[105] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3959–3970, 2022. 2, 4

[106] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[107] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[108] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[109] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 2

[110] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 4

[111] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, 2022. 4

[112] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, William T Freeman, and Jiajun Wu. Learning to Reconstruct Shapes From Unseen Classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1

[113] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[114] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[115] Keyang Zhou, Bharat Lal Bhatnagar, Bernt Schiele, and Gerard Pons-Moll. Adjoint rigid transform network: Task-conditioned alignment of 3d shapes. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022. 3, 1

[116] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[117] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5826–5835, 2021. 5, 1

[118] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2, 4