CARI4D: Category Agnostic 4D Reconstruction of Human-Object Interaction

Supplementary Material

In this supplementary, we discuss in more details about our CoCoNet architecture, joint optimization and runtime performance. We then discuss limitations and failure cases. Please refer to our supplementary video for 4D reconstruction results and baseline comparisons.

6. Implementation Details

We detail our implementations in this section. Note that our code and pretrained models will be fully released with detailed documentation to enable reproduction of our results.

6.1. CoCoNet Details

Network architecture. We plot the architecture diagram of our CoCoNet in Fig. 7. We adopt DINOv2 [29] as our image encoder and keep the base DINO for RGB image frozen while the small DINO model for xyz map and mask is fine tuned. We replace the first convolution layer in DINOv2 small to consume five channels (xyz + human object masks) instead of three RGB channels and finetune it end to end with other network layers. The attention block is spatiotemporal attention module similar to SV3D [45]. Specifically, given feature grid of shape (b, t, d, h, w), we first reshape and perform spatial attention on feature (bt, hw, d) to obtain $\mathbf{F}_s$, followed by a temporal attention on feature (bhw, t, d) to obtain $\mathbf{F}_t$. We then blend spatial feature $\mathbf{F}_s$ and temporal feature $\mathbf{F}_t$ with a learnable factor $\alpha \in [0, 1]$: $\mathbf{F} = (1-\alpha)\mathbf{F}_s + \alpha\mathbf{F}_t$. We use two spatiotemporal attention blocks in our network and temporal window size is $t = 96$. Following the spatiotemporal attention are five MLP heads that regress delta updates of object rotation $\Delta\mathbf{R} \in \mathbb{R}^6$, object translation $\Delta\mathbf{t}^o \in \mathbb{R}^3$, SMPL pose $\Delta\boldsymbol{\theta} \in \mathbb{R}^{144}$, SMPL shape $\Delta\beta \in \mathbb{R}^{10}$, and SMPL translation $\Delta\mathbf{t}^h \in \mathbb{R}^3$.

Training details. We train our model on BEHAVE [3] and HODome [72] dataset following the data preprocessing discussed in Sec. 3.3. We use AdamW optimizer with a learning rate of 1e-4 and an effective batch size of 16 for 38k steps. The training process takes about 38 hours on 8×A100@80GB GPUs.

6.2. Joint Optimization

We discuss in more details the objective function for our contact aware joint optimization (Sec. 3.4). Given a sequence of refined human-object poses $\{\mathcal{H}_i, \mathcal{O}_i\}_{i=1}^N$ and binary hand contact labels $\{\mathbf{c}_i\}_{i=1}^N$ predicted by our CoCoNet, we aim at improving its contact realism and avoiding penetration via optimizing the pose parameters $\{\mathcal{H}_i, \mathcal{O}_i\}_{i=1}^N$. The overall objective function is defined as:

$$L = \lambda_c L_c + \lambda_{\text{j2d}} L_{\text{j2d}} + \lambda_m L_m + \lambda_{\text{pen}} L_{\text{pen}} + \lambda_{\text{acc}} L_{\text{acc}} \quad (3)$$

where $L_c, L_{\text{j2d}}, L_m, L_{\text{pen}}, L_{\text{acc}}$ are the contact loss, 2D body joint reprojection loss, object mask loss, penetration loss and acceleration loss respectively.

The contact loss L_c penalizes large distances between hand and object when there is contact: $L_c = \sum_i d(\mathbf{J}_i^h, \mathbf{O}_i') \cdot \mathbf{c}_i$, where $d(\cdot, \cdot)$ computes the closest distance from two hand joints $\mathbf{J}_i^h \in \mathbb{R}^6$ to the posed object points $\mathbf{O}_i'$.

The 2D joint projection loss L_{j2d} minimizes the distance between projected 2D joints $\pi(\mathbf{J}(\mathbf{H}_i))$ and detected 2D joints $\hat{\mathbf{J}}_i$ from input image using openpose [1]: $L_{\text{j2d}} = \sum_i ||\hat{\mathbf{J}}_i - \pi(\mathbf{J}(\mathbf{H}_i))||_2^2$, where $\mathbf{J}(\cdot)$ regresses the 3D body joints from SMPL mesh $\mathbf{H}_i$ and $\pi(\cdot)$ projects the 3D joints to 2D image.

The 2D object mask loss L_m is occlusion aware and defined as: $L_m = \sum(M - I \circ S)^2$, where M is the input object mask, I is a non-occlusion indicator (1 if pixel belongs only to this object, 0 else) and S is the rendered object silhouette. The non-occlusion indicator is derived from human and object masks. The human and object masks may have overlap regions due to imperfect segmentation. To address this, pixels that belong only to object mask are assigned value 1 for the non-occlusion indicator. This avoids computing loss on regions where the object is occluded, see illustration in PHOSA [75].

We define the penetration loss using Volumetric SMPL [27] which learns a function Φ_{SMPL} that predicts a signed distance given a query point $q \in \mathbb{R}^3$, formally: $\Phi_{\text{SMPL}} : \mathbb{R}^3 \mapsto \mathbb{R}$. The penetration loss is hence defined as: $L_{\text{pen}} = \sum_i \sum_{q \in \mathbf{O}_i'} \text{ReLU}(-\Phi_{\text{SMPL}}(q))^2$. In practice we sample 6000 points on the object mesh surface as query points to compute the penetration loss.

The acceleration loss L_{acc} avoids large jitters by pushing the acceleration to be zero. In general: $L_{\text{acc}} = ||\mathbf{x}_i - 2\mathbf{x}_{i-1} + \mathbf{x}_{i-2}||_2^2$, here $\mathbf{x}$ includes 3D human body joints $\mathbf{J}(\mathbf{H})$ and object poses.

The loss weights are $\lambda_c = 200, \lambda_{\text{j2d}} = 0.03, \lambda_m = 0.002, \lambda_{\text{pen}} = 2.0$ and we use $\lambda_{\text{acc}} = 600$ for human body joints, $\lambda_{\text{acc}} = 1000$ for object poses. We use Adam optimizer with linear learning rate decay (start with 1e-3 learning rate) to optimize the parameters for 3000 steps. For efficiency we add penetration loss only in the last 1200 steps.

6.3. Runtime Performance

We report the average runtime of different methods to finish processing a video sequence of 300 frames using

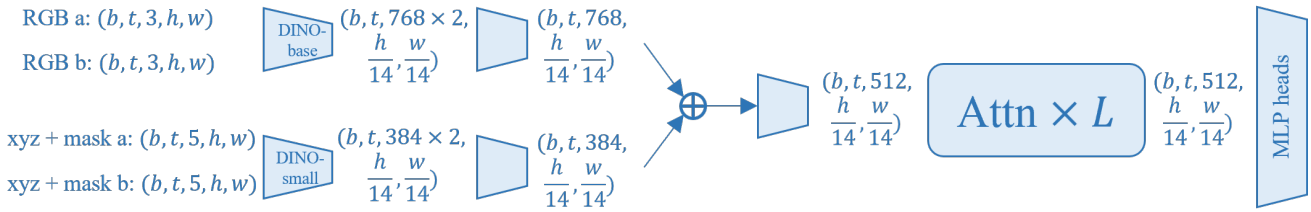

Figure 7. **CoCoNet architecture**. Here b, t, h, w denote batch size, temporal window size, image height and width respectively. We follow a render-and-compare paradigm, hence RGB a and RGB b denote the image from input observation and rendering respectively, same for xyz map and mask (human and object stacked together).

one A100@80GB GPU in Tab. 5. Image based approach PICO [10] requires the longest time as it optimizes one image each time and each optimization takes almost five minutes to finish. Video based methods InterTrack [59] and VisTracker [57] are faster as they leverage temporal cues and process multiple frames in parallel. However, they are still lower than our method due to complex multi-stage optimization (VisTracker [57]) or diffusion sampling process (InterTrack [59]). Our method achieves the fastest runtime while still producing the most accurate results.

Method	PICO	InterTrack	VisTracker	Ours
Runtime (min) ↓	1560	198	118	**45**

Table 5. **Average runtime (minutes)** to process one video of 300 frames. Our method is much faster than baselines while being more accurate.

7. Further Analysis

We provide additional analysis of our method under different conditions in this section.

Initialization and error propagation of foundation models. Our method relies mainly on FoundationPose(FP) and UniDepth predictions. UniDepth already provides accurate priors (9.20 cm avg. error on BEHAVE). We mitigate foundation model errors via our pose hypothesis selection and CoCoNet refinement. We now further validate robustness by injecting noise into FoundationPose and UniDepth predictions and run full CARI4D pipeline on BEHAVE (same test set as Tab 3). Plots below show the final chamfer distance (cm) of combined human-object mesh vs. injected noise magnitude, together with the error of VisTracker. Results confirm CARI4D remains stable and superior to VisTracker even under significant initialization noise, being insensitive to foundation model errors.

Reliability and impact of object occlusion. We assume 1st frame is mostly visible, which reduces the hallucination for object mesh reconstruction. We plot the final object error versus the object visibility in the reconstruction frame in the figure below. There is no clear correlation between the visibility for reconstruction and final error because the performance also depends on the occlusion in other frames and difficulty of the motion. As long as the frame used to reconstruct is reasonably visible ($> 86\%$ in our case), it does not significantly affect the final performance. We also provide the object mask for mesh reconstruction and pose tracking to identify target object.

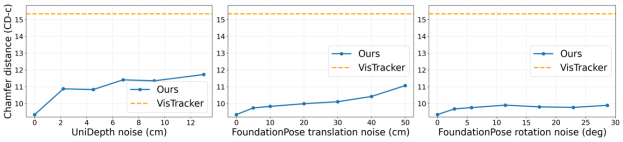

Figure 8. Sensitivity analysis of UniDepth and FoundationPose errors on the final performance.

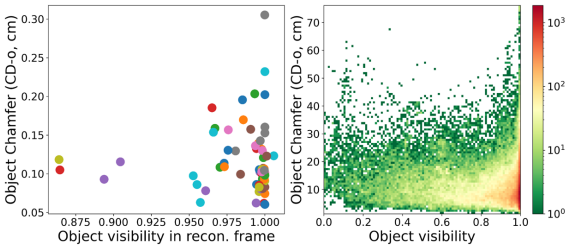

Figure 9. Error per seq. (left) and visibility distribution (right).

Interactions other than hand-objects. We did evaluate the performance other than hand only interactions as BEHAVE test set also includes sitting, leaning, shoulder carrying. We classify sequences based on interaction types and evaluate the performance in the table below. We can see that errors are similar among different interaction types, demonstrating the robustness of our method. We also show one sitting example in Fig. 10b.

Interaction type	CD-h↓	CD-o↓	CD-c↓	Acc-h↓	Acc-o↓	Contact↑
Hand mainly	8.76	11.71	7.12	1.06	0.42	0.93
Sitting	8.16	10.10	7.23	1.01	0.32	0.95
Shoulder contact	10.24	12.39	9.27	1.03	0.48	0.95
Leaning	9.37	11.44	8.28	1.05	0.31	0.93

Table 6. Performance of our method per interaction type.

8. Limitation and Future Work

As the first step towards category agnostic 4D interaction reconstruction, our method shows strong generalization per-

Figure 10. Additional qualitative examples.

Figure 11. **Failure case examples.** Our method focuses on full body interaction and the detailed hand poses are not handled, which can be important for fine-grained object manipulation task (top row). Our method thus failed to reconstruct realistic finger poses for holding the plate. Under highly dynamic motion and extreme occlusion (bottom row), FoundationPose predicts flipped object pose for initialization. Such large rotation error is not able to be corrected by our refinement process in subsequent steps, leading to inaccurate reconstruction in the end.

formance to in-the-wild videos, yet there are still some limitations. We show two typical failure cases in Fig. 11.

First, our method primarily targets full-body human-object interaction; consequently, it does not explicitly regress detailed finger articulatons. This limitation becomes particularly pronounced during interactions involving small-scale objects or those requiring fine-grained manipulation, such as grasping plates (Fig. 11, top row). In such scenarios, the absence of precise finger kinematics results in physically implausible interaction configurations, even when the full body pose is accurate. To bridge this gap, future iterations could integrate specialized hand pose estimators [31, 34, 74]. By solving for the hand parameters separately and fusing them with the full-body kinematic chain via optimization, one could achieve a holistic reconstruction that respects both macro-level body dynamics and micro-level contact physics.

Second, our method relies on FoundationPose [52] for object pose initialization, subsequently refining these estimates using human interaction cues and visual evidence. A notable dependency bottleneck arises when the initializer fails significantly. FoundationPose can occasionally predict flipped 180-degree poses under conditions of rapid motion or severe occlusion (Fig. 11, bottom row). In these instances, the error magnitude is often too large for our refinement network to correct. A promising direction is the incorporation of temporal priors and motion infilling, similar to strategies employed in VisTracker [57] or GLAMR [70]. By leveraging information from visible frames to hallucinate motion in occluded segments, we can enforce temporal smoothness and recover from initialization failures.

References

[1] https://github.com/cmu-perceptual-computing-lab/openpose. 1

[2] Stability AI. Stable video diffusion: A novel approach to image-to-video generation. *arXiv preprint arXiv:2308.09592*, 2023. Available at https://github.com/Stability-AI/generative-models. 5

[3] 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. 3, 5, 6, 7, 1

[4] 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. 2

[5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Representation Learning (ICLR)*, 2025. 2

[6] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. *ACM Trans. Graph.*, 27(3):1–9, 2008. 2

[7] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS) — Datasets abd Benchmarks Track*, 2023. 2

[8] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025. 2

[9] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), 2015. 2

[10] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *CVPR (CVPR)*, pages 1783–1794, 2025. 2, 3, 5, 6, 7, 8

[11] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *CVPR (CVPR)*, 2025. 3, 7

[12] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *CVPR*, pages 494–504, 2024. 3

[13] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2

[14] Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European conference on computer vision (ECCV)*, 2024. 2

[15] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *CVPR (CVPR)*, 2025. 2

[16] 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. 3

[17] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 3

[18] 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. 3, 5, 6, 7

[19] Chaofan Huo, Ye Shi, and Jingya Wang. Monocular human-object reconstruction in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 5547–5555, New York, NY, USA, 2024. Association for Computing Machinery. 3

[20] Hsuan I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR (CVPR)*, pages 538–549, 2024. 2

[21] 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. 2

[22] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[23] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: Generative models for human motion synthesis. *arXiv preprint arXiv:2505.01425*, 2025. 2, 7

[24] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, pages 21013–21022, 2022. 3

[25] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR (CVPR)*, pages 21013–21022, 2022. 3

[26] 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. 3

[27] Marko Mihajlovic, Siwei Zhang, Gen Li, Kaifeng Zhao, Lea Müller, and Siyu Tang. VolumetricSMPL: A neural volumetric body model for efficient interactions, contacts, and collisions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 5, 1

[28] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *CVPR*, 2024. 3, 6

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5, 1

[30] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 13463–13473, 2021. 2

[31] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 3

[32] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR (CVPR)*, 2024. 2, 3, 4

[33] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025. 2, 3, 5, 7

[34] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024. 3

[35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 2, 5

[36] 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. 3

[37] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. 2024. 2, 3, 4, 5, 7, 8

[38] Agniv Sharma, Xianghui Xie, Tom Fischer, Eddy Ilg, and Gerard Pons-Moll. Hoi3dgen: Generating high-quality human-object-interactions in 3d. In *CVPR findings*, 2026. 3

[39] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *CVPR*, pages 8623–8632, 2020. 3

[40] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular RGBD stream. *CoRR*, abs/2104.14837, 2021. 3

[41] 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. 3

[42] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 2, 3

[43] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[44] 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. 3, 7

[45] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 1

[46] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, pages 2642–2651, 2019. 5

[47] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 2

[48] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR (CVPR)*, pages 5261–5271, 2025. Oral. 2

[49] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint, 2025. 2

[50] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[51] 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

[52] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024. 2, 3, 5, 7, 8

[53] Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and

Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions, 2025. 3

[54] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. *arXiv preprint arXiv:2012.01591*, 2020. 3

[55] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR (CVPR)*, 2025. Spotlight. 2

[56] 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. 3, 5, 6

[57] 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, 3, 5, 6, 7, 8

[58] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6

[59] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. 2024. 2, 3, 6, 7, 8

[60] 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. *arXiv preprint arXiv:2401.04143*, 2024. 3

[61] Xianghui Xie, Chuhang Zou, Meher Gitika Karumuri, Jan Eric Lenssen, and Gerard Pons-Moll. Mvgbench: Comprehensive benchmark for multi-view generation models, 2025. 2

[62] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Gen-3diffusion: Realistic image-to-3d generation via 2d & 3d diffusion synergy, 2024. 2

[63] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In *Arxiv*, 2024. 2

[64] Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Infinihuman: Infinite 3d human creation with precise control. 2025. 2

[65] Pradyumna Yalandur-Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Physic: Physically plausible 3d human-scene interaction and contact from a single image. In *ACM SIGGRAPH Asia*, 2025. 3

[66] 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. 3

[67] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR (CVPR)*, 2024. arXiv preprint arXiv:2401.10891. 2

[68] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 3

[69] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, 2024. 3

[70] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR (CVPR)*, 2022. 3

[71] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022. 7

[72] 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. 3, 5, 1

[73] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m3: Capture multiple humans and objects interaction within contextual environment. In *CVPR*, 2024. 3

[74] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. *arXiv preprint arXiv:2501.02973*, 2025. 3

[75] 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. 3, 5, 1

[76] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 2

[77] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *CVPR*, pages 729–741, 2024. 3