InterTrack: Tracking Human Object Interaction without Object Templates

Supplementary Material

In this supplementary, we provide more implementation details for our tracking method and synthetic video generation. We also further analyze the design considerations and discuss typical failure cases of our method. Please refer to our supplementary video for video tracking results.

6. Implementation Details

We discuss the network architecture, training, and optimization details in this section. Our code will be publicly released with clear documentation to foster future research.

6.1. CorrAE and human optimization

For our human CorrAE, we adapt the encoder from PVCNN [43] which is also used in [47, 79]. It compresses input point clouds of shape $N \times 3$ into downsampled point feature of shape 512×16. We add two additional point convolution layer [43] to further compress it into latent vector $\mathbf{z}^h$ of shape 1024×1. The latent code is then sent to one MLP layer, followed by 6 blocks of MLP layers with residual connection. The MLP compress the latent code to 512 dimension and each block consists of three MLP layers with LeakeyReLU activation. The output dimensions of the MLPs in each block are $256, 256, 512$. The 512 dimension feature vector is then sent to a large MLP which predicts 6890 SMPL vertices as a single vector.

We train our CorrAE with a loss weight $\lambda_{\text{v2v}} = 100$ for the vertex to vertex loss and use Adam optimizer with learning rate of 3e-4. The model is trained on the GT SMPL meshes from ProciGen training set. It takes around 12 hours to finish training on 4 RTX8000 GPUs with batch size 32. The loss weights for the human optimization are: $\lambda^h_{\text{cd}} = 100, \lambda_p = 1e-5, \lambda_{\text{acc}} = 100$. We use Adam of learning rate 0.001 and stochastic gradient descent to optimize the human pose parameters, with a batch size of 256. We optimize for 2500 steps which takes $\sim$30 minutes on an A40@40GB GPU.

6.2. TOPNet and oject optimization

For the object pose TOPNet, we combine DINOv2[53] image encoder with transformer [66]. DINOv2 encodes image of shape $3 \times 224 \times 224$ into a feature grid of $768 \times 16 \times 16$. We then add three 2D convolution layers with kernel size 4, group normalization and leaky ReLU activation to further compress the feature grid into a vector of shape $1 \times 1 \times 768$. This operation is similar to the one used in MagicPony [76]. The dimension of human feature is $294 = 25 \times 6 + 24 \times 6$, which consists of 25 body joints and their velocities and SMPL body pose represented as rotation 6D[105]. We encode the human feature using two MLPs with a latent dimension 128 and output dimension 128. The human feature is then concatenated with object visibility and image feature vector and sent to transformer with 3 encoder layers [66]. Each encoder layer has 4 heads and feed forward dimension of 256. The temporal features are then sent to 3 MLP layers with output dimensions of 128, 64, and 6.

We train the model with learning rate 3e-4 (Adam optimizer) and batch size 16, temporal window size 16. It takes around 31 hours to converge on 4 RTX8000 GPUs. We train two models for all 10 categories in ProciGen-V dataset: one for large objects (chair, table, monitor) and another one for small symmetric objects (all the rest categories). The loss weights for the object optimization are: $\lambda^o_{\text{cd}} = 10, \lambda_{\text{occ}} = 0.001, \lambda^r_a = 1000, \lambda^t_a = 200, \lambda^s_a = 1000$. We optimize canonical shape and per-frame poses with a batch size of 64. For models trained on synthetic data only we optimize for 16k steps as the initial shape is less accurate, which takes around 2 hours. For models fine-tuned on real data, we optimize only 6k steps which takes 50-60 minutes on one A40 GPU.

6.3. Joint optimization

The loss weights for the human ($\mathcal{L}_{\text{hum}}$) and object ($\mathcal{L}_{\text{obj}}$) loss terms are the same as the ones used for separate optimization. The contact loss weight $\lambda_c = 10$. Note that we optimize only the SMPL body pose and object rotation parameters as this is used only for fine tuning the poses.

Similar to separate optimization, we use Adam with learning rate 0.001 for human and 6e-4 for object. We refine for 2500 steps with batch size 64, which takes in total $\sim$35 minutes on one A40 GPU.

6.4. ProciGen-Video data generation

We start from ProciGen proposed in [79] to procedurally generate interaction videos for new object shapes. The goal is to change the human and object shape and render new videos. We first sample a chunk of human and object poses from interaction sequences in real data. The human is represented using SMPL [44] pose $\Theta = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N\}$ and shape $\mathcal{B} = \{\beta_1, ...\beta_N\}$ parameters, here $1, ..., N$ are the time index. We compute dense correspondence between original object shape and new shape using an autoencoder [103], which allows transferring contacts from original shape to new shape. We also use the correspondence to initialize the pose $\mathbf{T}_i \in \mathbb{R}^{4\times 4}$ for the new object [79]. The initialization can lead to interpenetration problem, hence we further optimize the body poses Θ, shapes $\mathcal{B}$ and object transformations $\mathcal{T} = \{\mathbf{T}_1, ...\mathbf{T}_N\}$ to satisfy contacts and temporal

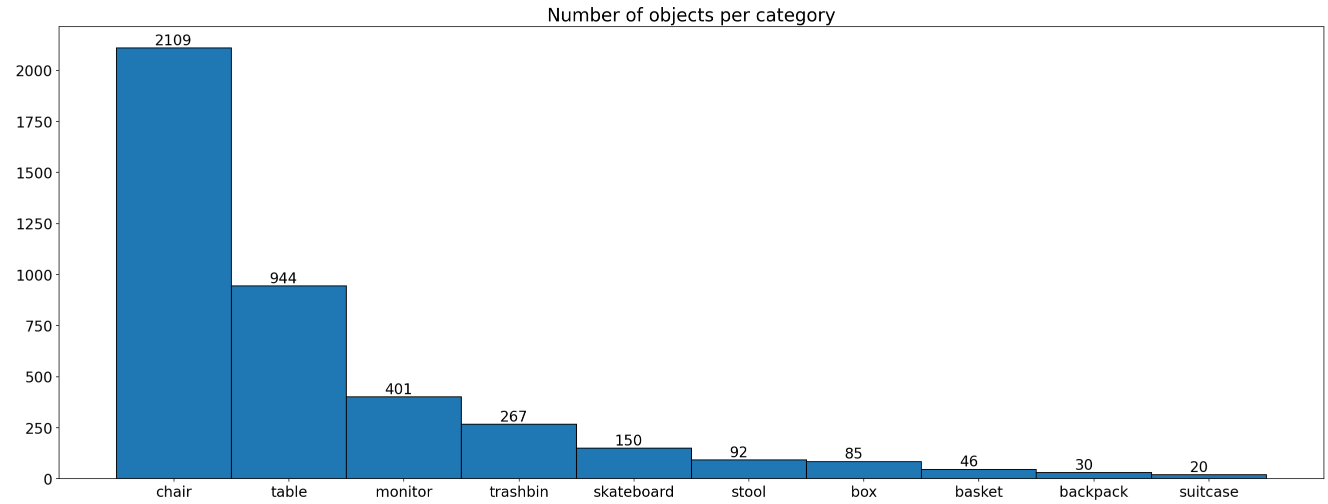

Figure 6. **Number of distinct object shapes** used in our ProciGen-V dataset. Our method is scalable and can generate interaction for new object shapes within these categories.

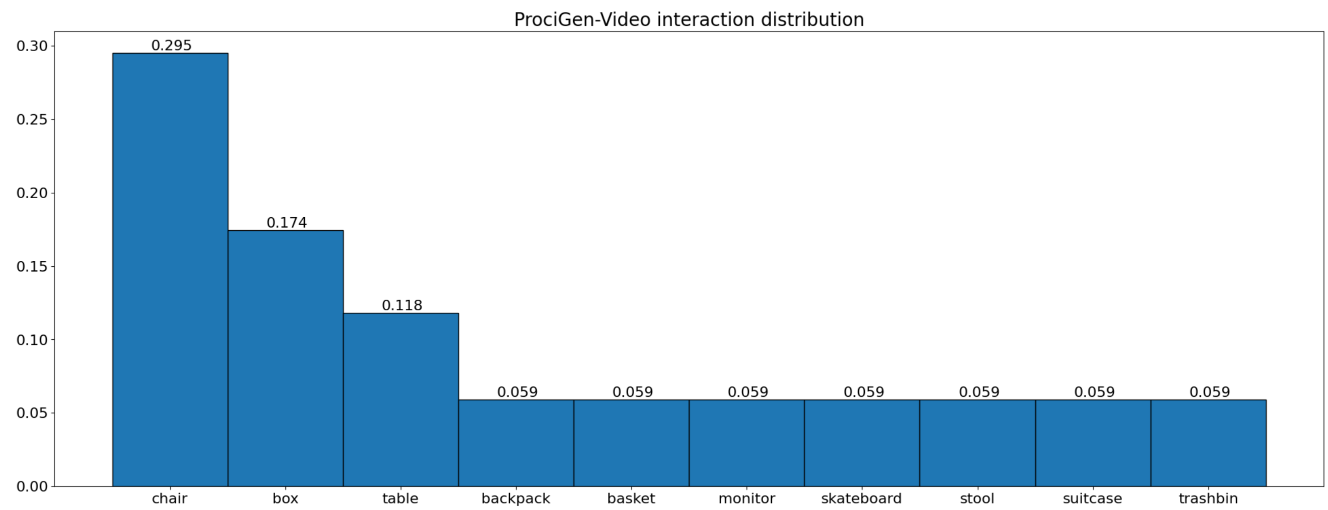

Figure 7. **Distribution of interaction sequences per category** in our ProciGen-V dataset. Our dataset is balanced for most categories except for chair which contains more complex shape and interactions.

smoothness:

$$\mathcal{L}(\Theta, \mathcal{T}, \mathcal{B}) = \lambda_c L_c + \lambda_n L_n + \lambda_{\text{colli}} L_{\text{colli}} + \lambda_{\text{init}} L_{\text{init}} + \lambda_{\text{acc}} L_{\text{acc}} \tag{5}$$

where the contact loss L_c, normal loss L_n, interpenetration L_{init} and initialization penalty L_{init} are defined in [79]. And L_{acc} is the temporal smoothness loss defined in Eq. (2) applied to a sequence of SMPL vertices. Note that we also randomly sample a body shape parameter from the MGN dataset [3] to replace the original shape for more diversity. The loss weights used are: $L_c = 400, L_n = 6.25, L_{\text{colli}} = 9, L_{\text{init}} = 100, L_{\text{acc}} = 10$.

Once optimized, we use SMPL-D registration [3] which adds per-vertex offsets to the SMPL vertices to model clothing deformation and texture. For the object, we use the original textures from the CAD model. We also add small random global rotation and translation to the full sequence to increase diversity. We render the human and object in blender with random lighting and no backgrounds. Some example renderings can be found in ADD REF.

We generate interaction videos for 10 object categories. The interaction poses are sampled from BEHAVE [6] and InterCap [27], object shapes are sampled from Objaverse [13] and ShapeNet [8]. The distribution of distinct object shapes can be found inFig. 6, and the number of interaction sequences per-category can be found in Fig. 7. The original BEHAVE and ShapeNet are captured at 30fps, we generate synthetic data at 15fps and each sequence has 64 frames (4.27 seconds). In total, we generate 8477 sequences

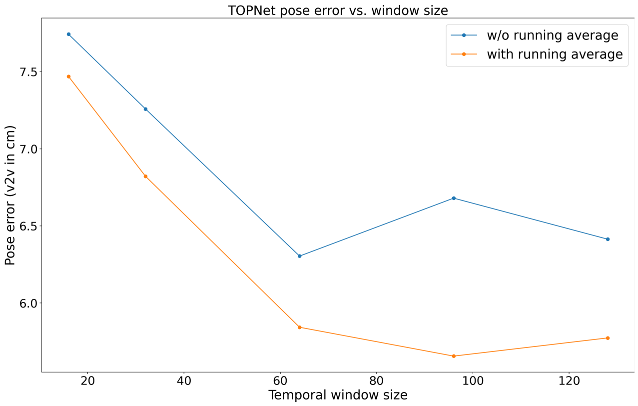

Figure 8. **Object pose error versus the temporal window size** used at inference time. The model was trained with window size=16. Averaging predictions of each frame in different sliding windows consistently leads to better pose estimations.

which amounts to 10 hours long videos. Our method can scale up to include more objects and longer videos, which is much more scalable than capturing real data.

7. Additional Analysis and Result

In this section, we provide additional analysis to the design considerations of our human and object reconstruction modeuls. We also show generalization to unseen category. Please refer to our video for more results and comparison.

7.1. Object pose TOPNet

Our TOPNet computes cross attention between W consecutive images and directly predicts W rotations for them. We train our model with $W = 16$ due to limited IO speed: with a batch size of 16, it needs to load 256 images with corresponding GT data which already takes $0.6 \sim 1$ second. Using longer window size significantly increases the training time. In contrast, we find that the learned attention weights can be applied larger window size even though the model is trained for $W = 16$ only. We plot the object pose error with different test time window size in Fig. 8. Here we report the pose error as the vertex to vertex error (cm) after applying predicted and GT rotation to the GT object vertices. We apply a sliding window of size W to process the full sequence, which means each image can appear several times at different sliding windows. We average predictions of all possible sliding windows, which also leads to smoother and more accurate pose, see Fig. 8 (with running average).

7.2. Human Reconstruction

We compare the correspondence across frames from HDM and our method in Fig. 9. HDM is image-based method and outputs point clouds without any ordering. On the other hand, our method tracks the point across the full sequence.

We argue in Sec. 3.2 that the latent space of our CorrAE is less interpretable which leads to slightly worse result compared to optimizing via SMPL layer (Tab. 5). Here we visualize another problem of optimization via the CorrAE: the surface points become less smooth, see Fig. 10. It can be seen that some points on the feet spread out from the original position, leading to a noisy surface. In contrast, optimizing via SMPL layer guarantees a smooth surface.

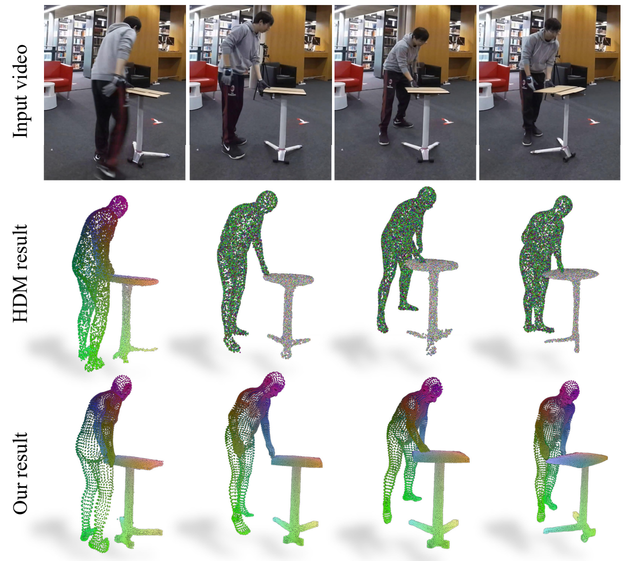

Figure 9. **Visualization of the correspondence.** HDM [79] outputs unordered points while our method consistently tracks the human and object across frames.

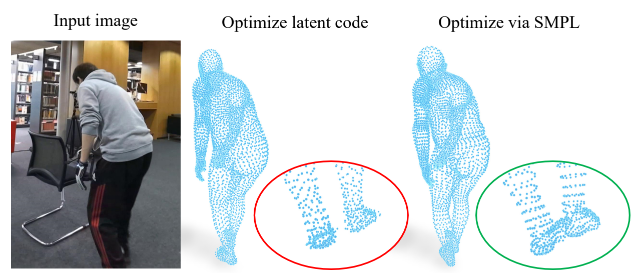

Figure 10. **The problem of optimizing CorrAE latent code**. The latent space of our CorrAE entangles human pose and shape. Optimizing it directly also leads to less smooth surface.

7.3. Generalization to unseen categories

Our model was trained on ten common daily life object categories. It works well for new object instances of the same category, as can be seen in Fig. 1 and our supplementary video. We also test our method on unseen category in Fig. 11. In general, our method can work on new categories that are similar to those seen in our training set.

8. Limitation and Failure Case Analysis

Limitations. Despite impressive performance on benchmark datasets and strong generalization to real videos, there are still some limitations of our method. First, our method

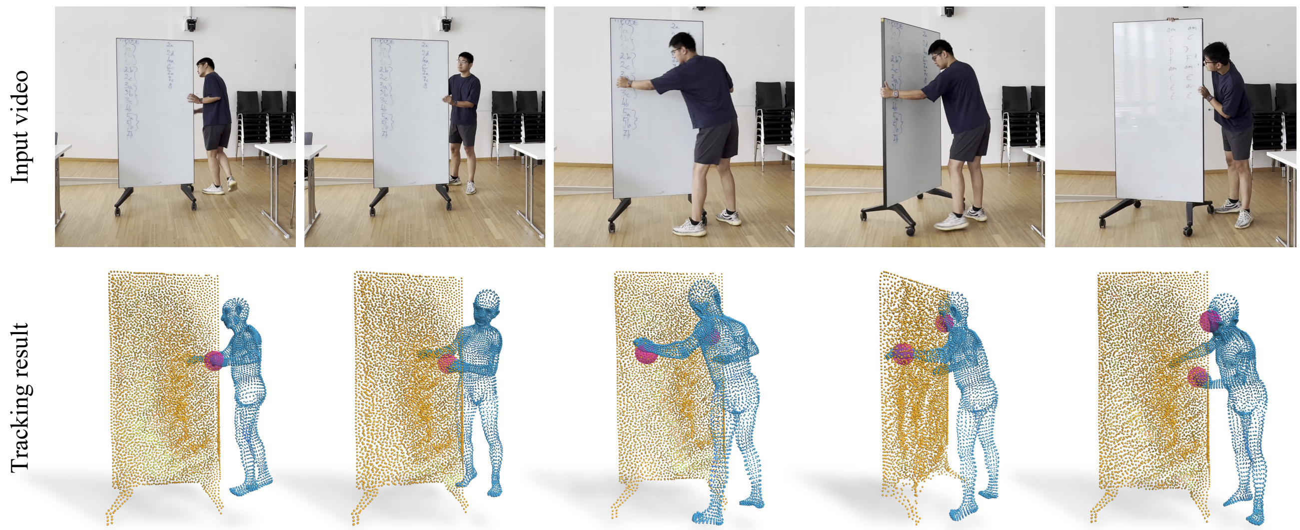

Figure 11. **Generalization to unseen category.** We test our method to unseen category blackboard. It can be seen that our method can reconstruct the shape and tracks the human object interaction.

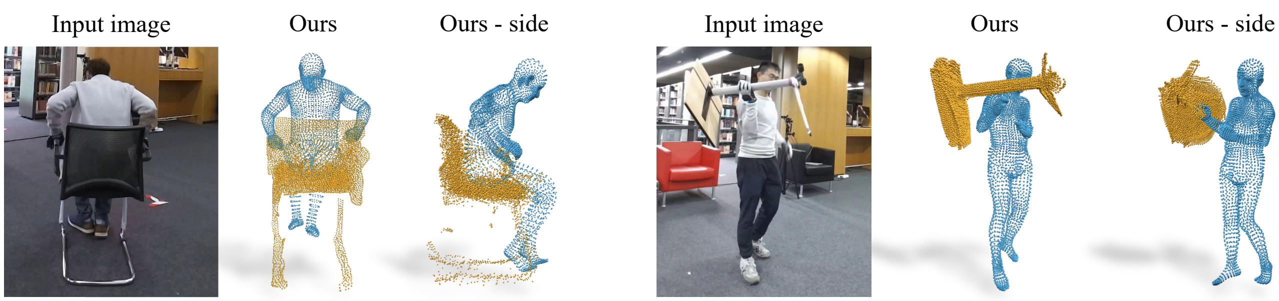

Figure 12. **Example failure cases.** Our method fails to reconstruct the object shape (left) as only one view of the object is seen in the entire video. It can also struggle to predict extreme rare pose (right), leading to less faithful shape and tracking.

does not reconstruct the textures of the human and object. Our method is easily compatible with Gaussian Splating [32] and adding colors to each point could potentially further constraint the optimization [12]. Second, our dataset contains only the categories from BEHAVE and InterCap. Future works can capture more objects or explore synthesizing interactions without real data [33]. Furthermore, our method does not deal with object symmetries explicitly. Future works can adopt good practices from object pose estimation community [14, 59, 67] to further enhance the robustness of our method. Multi-human, multi-object interaction are also interesting directions to explore. We leave these to future works.

Failure cases. We show two typical failure cases of our method in Fig. 12. Overall, our method tracks humans reliably in most cases while object tracking is more challenging due to occlusions and lack of template shapes. Our method can produce noisy object shape when there are not enough views to reason the object. In Fig. 12 left, the chair remains static in the full sequence, hence our method only receives information about the chair in back side view. The object shape aligns well with the input but the 3D structure is suboptimal. Future works can further improve our method by imposing stronger object shape prior. For example, optimizing via a well-behaved latent space which provides better output shape.

Our method can also predict noisy object pose under rare or very dynamic interaction like Fig. 12 right. In this sequence, the arm and object move very quickly, leading to noisy pose prediction which dominate the optimization and results in inaccurate shape and tracking. Training on more objects or with additional data augmentation such as Foundationpose [72] could potentially generalize better. However, Foundationpose relies on CAD model and depth input. One interesting direction is to develop methods that can iteratively improve object shape reconstruction and pose estimation. With our TOPNet, one can obtain initial object reconstruction, which should be helpful to improve object pose estimation. This iterative mutual improvement should lead to better shape and pose tracking.

Figure 13. **Example sequences from our ProciGen-Video dataset.** We generate realistic interactions with diverse object shapes. Please refer to our supplementary video for more examples.

References

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

[2] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21190–21200, 2023. 2

[3] 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. 6, 2

[4] 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. 3, 4

[5] 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. 3, 4

[6] 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. 2, 3, 5, 6, 7

[7] Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Ov9d: Open-vocabulary category-level 9d object pose and size estimation. *arXiv preprint arXiv:2403.12396*, 2024. 7

[8] 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. 5, 6, 2

[9] Yuhang Chen and Chenxing Wang. Kinematics-based 3D Human-Object Interaction Reconstruction from Single View, 2024. 3

[10] 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

[11] 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. 3

[12] Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen. Neural parametric gaussians for monocular non-rigid object reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[13] 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. 5, 6, 2

[14] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12376–12385, Montreal, QC, Canada, 2021. IEEE. 4

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 3, 6

[16] Haiwen Feng, Peter Kulits, Shichen Liu, Michael J. Black, and Victoria Abrevaya. Generalizing neural human fitting to unseen poses with articulated se(3) equivariance, 2023. 3, 8

[17] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *ECCV*, 2018. 3

[18] 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. 3

[19] 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. 3

[20] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019. 3

[21] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *CVPR*, 2023. 3

[22] 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. 2

[23] 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

[24] 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. 4

[25] 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. 3

[26] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 3

[27] 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. 2, 3, 5, 6, 7

[28] Chaofan Huo, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. StackFLOW: Monocular Human-Object Reconstruction by Stacked Normalizing Flow with Offset. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 902–910, Macau, SAR China, 2023. International Joint Conferences on Artificial Intelligence Organization. 3

[29] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. *arXiv preprint arXiv:2202.12825*, 2022. 3

[30] 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. 3

[31] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. Multiply: Reconstruction of multiple people from monocular video in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 4

[33] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models, 2024. 4

[34] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262. IEEE, 2020. 2

[35] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11127–11137. IEEE, 2021. 5

[36] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. TTA-COPE: Test-Time Adaptation for Category-Level Object Pose Estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21285–21295, Vancouver, BC, Canada, 2023. IEEE. 7

[37] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42 (6), 2023. 2

[38] Lei Li and Angela Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[39] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep Iterative Matching for 6D Pose Estimation. In *European Conference on Computer Vision (ECCV)*, page 16, 2018. 7

[40] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. *arXiv preprint arXiv:2401.02400*, 2024. 2

[41] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an RGB image. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 2, 7

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2

[43] 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. 3, 1

[44] 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. 2, 3, 4, 8, 1

[45] Riccardo Marin, Simone Melzi, Emanuele Rodolà, and Umberto Castellani. Farm: Functional automatic registration method for 3d human bodies, 2018. 3

[46] Riccardo Marin, Enric Corona, and Gerard Pons-Moll. Nicp: Neural icp for 3d human registration at scale. In *European Conference on Computer Vision*, 2024. 2, 3, 4, 8

[47] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *CVPR*, 2023. 1

[48] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[49] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2

[50] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 343–352, 2015. 2

[51] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and Robust

Novel Object Pose Estimation via One Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7

[52] David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, and Andrea Vedaldi. Keytr: Keypoint transporter for 3d reconstruction of deformable objects in videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5585–5594, 2022. 2

[53] 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

[54] Litany Or, Remez Tal, Rodolà Emanuele, Bronstein Alex M., and Bronstein Michael M. Deep functional maps: Structured prediction for dense shape correspondence. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[55] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics - TOG*, 31, 2012. 3

[56] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[57] 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

[58] Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields. *CVPR*, 2024. 3

[59] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[60] Philipp Schröppel, Christopher Wewer, Jan Eric Lenssen, Eddy Ilg, and Thomas Brox. Neural point cloud diffusion for disentangled 3d shape and appearance generation. In *CVPR*, 2024. 3

[61] Haixin Shi, Yinlin Hu, Daniel Koguciuk, Juan-Ting Lin, Mathieu Salzmann, and David Ferstl. Free-Moving Object Reconstruction and Pose Estimation with Virtual Camera, 2024. 3

[62] 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

[63] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. *CVPR*, 2022. 4

[64] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[65] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 4

[66] 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, 1

[67] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021. 4

[68] 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. 3

[69] Jiaxin Wei, Xibin Song, Weizhe Liu, Laurent Kneip, Hongdong Li, and Pan Ji. RGB-based Category-level Object Pose Estimation via Decoupled Metric Scale Recovery, 2023. 7

[70] B Wen and Kostas E Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021. 7

[71] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Muller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. *CVPR*, 2023. 2

[72] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024. 7, 4

[73] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

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

[75] 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

[76] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3d animals in the wild. In *CVPR*, 2023. 2, 5, 1

[77] 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, 3, 5, 6, 7

[78] 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

[79] 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, 4, 5, 6, 7, 1

[80] 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. 2

[81] 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

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

[83] Yuxuan Xue, Bharat Lal Bhatnagar, Riccardo Marin, Nikolaos Sarafianos, Yuanlu Xu, Gerard Pons-Moll, and Tony Tung. Nsf: Neural surface fields for human modeling from monocular depth. In *ICCV*, 2023. 2

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

[85] 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

[86] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2

[87] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *NeurIPS*, 2021. 2

[88] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2

[89] Gengshan Yang, Chaoyang Wang, N. Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *CVPR*, 2023. 2

[90] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 3

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

[92] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 3

[93] 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

[94] Raza Yunus, Jan Eric Lenssen, Michael Niemeyer, Yiyi Liao, Christian Rupprecht, Christian Theobalt, Gerard Pons-Moll, Jia-Bin Huang, Vladislav Golyanik, and Eddy Ilg. Recent Trends in 3D Reconstruction of General Non-Rigid Scenes. *Computer Graphics Forum*, 2024. 2

[95] 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

[96] 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. 2

[97] 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

[98] Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Self-Supervised Geometric Correspondence for Category-Level 6D Object Pose Estimation in the Wild, 2023. 7

[99] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8521–8531, 2024. 2

[100] 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

[101] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *CVPR*, 2021. 3

[102] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *ECCV*. Springer, 2022. 2

[103] 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, 5, 1

[104] 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

[105] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 1