Contact Estimation for Bimanual Hand-Object Manipulation

Angela Xing
Brown University
angela_xing@brown.edu

Srinath Sridhar
Brown University
srinath@brown.edu

Abstract

Reconstructing dynamic hand-object contacts is essential for realistic bimanual manipulation, yet it remains challenging due to heavy occlusions, complex surface details, and limitations in existing capture techniques. In this paper, we introduce a novel markerless capture method for accurate and efficient dynamic bimanual hand–object manipulations. Our approach leverages a dynamic, articulated representation based on 2D Gaussian surfels to model complex interactions. By binding these surfels to MANO [43] meshes, we harness the inductive bias of template models to stabilize and accelerate optimization. To evaluate dynamic contacts, we curated a new hand–object manipulation dataset with ground-truth contacts, and we demonstrate that our method achieves state-of-the-art dynamic reconstruction quality and significantly improves contact estimation accuracy.

1. Introduction

Skillful object manipulation is one of the most common, yet impressive, human physical abilities. Human manipulation of objects is highly dynamic and involves the coordinated movements of fingers in both hands to perform complex tasks. An important step in analyzing or replicating manipulations is understanding the *dynamic contacts* between hands and objects [8]. Contact provides a measure of spatial proximity [4, 47, 60] and influences the kinematics and dynamics [3, 29, 62] of the interaction.

Despite its significance, accurately capturing and reconstructing dynamic contacts remains a challenge. Existing markerless contact capture methods rely on low-dimensional parameterized models for both the hand and the object. Representations such as hand skeletons [13, 24], meshes [1], neural fields[34], or parametric hand shape models [43] are often used, but struggle to capture fine contact details. Recently, Gaussian-Splatting methods [7, 17, 19, 22, 23, 28, 33, 50, 52, 54–56] have demonstrated impressive results in accurate pixel-aligned reconstruction, and Gaussian-based hand avatars [18, 37] have proved superior to other template-based methods. However, they are not designed for capturing *dynamic* contacts with relative hand-object motion, common in complex manipulations.

We address these limitations with a method for capturing contact in complex bimanual hand-object manipulations. Our method employs multi-view markerless capture for natural manipulations, representing hands and objects with 2D Gaussian surfels [17] that accurately model surfaces and appearance for contact estimation. Specifically, we bind 2D Gaussian surfels to a parametric hand mesh [43] for each hand. Unlike grasping, where the object remains static [37], our approach supports dynamic manipulation, involving both in-hand and between-hand object motions. To support this, we initialize a model-free 2D Gaussian surfel model for the object and track its pose over time. Then, based on surfel pair distances, we can efficiently compute instantaneous and accumulated contacts.

Evaluating dynamic contacts is challenging due to the difficulty in obtaining reliable ground truth. Therefore, we introduce a dataset that provides ground truth contacts even under heavy occlusion and rapid motion. We obtain ground truth contacts using an improved paint residue method [20, 37] redesigned for gathering complex manipulations at scale. Experimental results demonstrate that our method achieves state-of-the-art reconstruction quality and contact estimation accuracy.

In summary, our contributions are:

- We introduce a method for accurate contact capture in complex bimanual hand-object manipulation sequences. Our method reconstructs *both the hand and the object* with dynamic 2D Gaussian surfels [17] for accurate surface modeling without misalignments.
- We curate and capture a dataset containing ground truth contact labels for challenging scenes with heavy occlusion and rapid motion.

2. Related Work

Capturing and Modeling Contact. Accurately capturing and modeling dynamic hand-object contact is essential for analyzing and replicating complex hand-object manipulations. However, the skeletal structure and soft tissue of

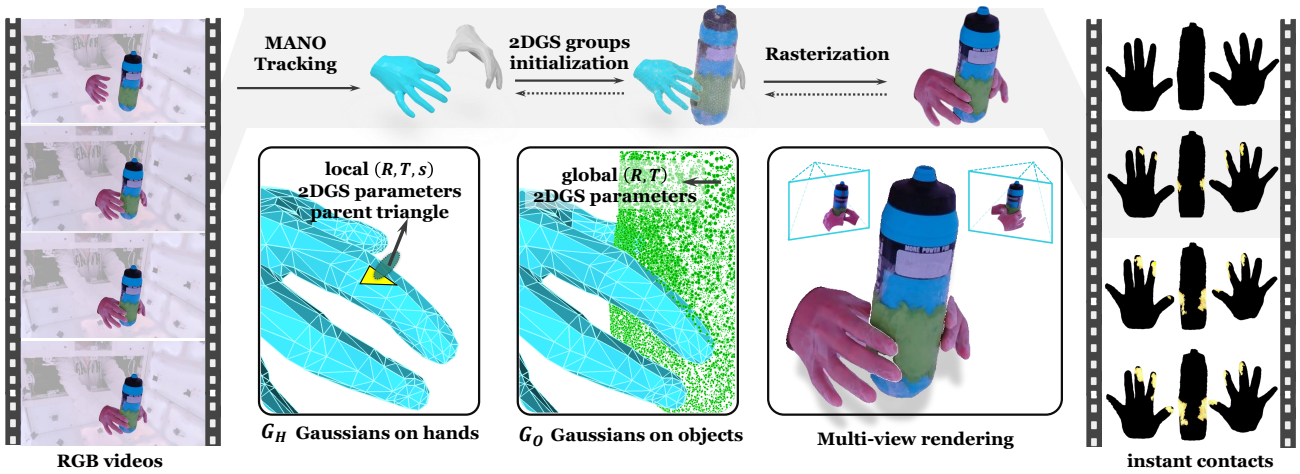

Figure 1. Our method captures dynamic contacts with a markerless system using a contact-aware dynamic articulated Gaussian representation. Given multi-view RGB videos, we bind Gaussian surfels to MANO mesh locally where they remain rigged throughout optimization. For objects, Gaussians are initialized by placing a coarse point cloud in the global coordinate space.

human hands pose significant challenges. Early methods relied on instrumented gloves [15, 27, 46], specialized sensors [12, 36, 58], or thermal imaging [4] to capture contacts, but these methods may restrain natural hand movements, affect tactile feedback, or even fail to capture dynamic changes. Existing works attempt to model dynamic contacts using hand skeletons [11, 13, 24], customized meshes [1], and MANO models [5, 9, 14, 30, 43, 47]. Although these methods exhibit impressive quality and generalization, they often suffer from misalignment and lack the fine details crucial for precise contact analysis. Inspired by the success of Gaussian Splatting, recent works [18, 37] adopt 3D Gaussians for appearance modeling that significantly improves reconstruction quality. However, these methods requires multi-view videos with minimal occlusions, a carefully designed optimization procedure for training an animatable Gaussian hand, and fail to capture dynamic contacts in manipulation scenes.

Dynamic Scene Representation. Recently, 3D Gaussian Splatting (3D-GS) [22] has become a popular representation for novel view synthesis. Building on 3D-GS, a flurry of work on dynamic scenes has emerged [7, 19, 23, 28, 33, 41, 50, 52, 54–56]. Dynamic3DGS [33] learns transformations of Gaussian primitives over time, facilitating dynamic tracking [49]. Deformable3DGS [56] and 4DGS [50] define a deformation field mapping canonical Gaussian primitives to specific time steps, but this approach struggles with new content appearing or disappearing. While these techniques perform well on general scenarios, few provide a dedicated approach for articulated objects like hands. Additionally, driving these methods with a predefined hand model or using them in complex, real-world manipulations remains a major challenge. Efforts like GaussianAvatars [38] and SurFhead [25] apply Gaussian Splatting to rig parametric models (e.g., FLAME [26]) for head avatar reconstruction [16, 40, 48, 51, 53], highlighting the potential of template-driven Gaussian Splatting. However, exploiting dymanic reconstruction for complex hand object manipulations with template models such as MANO [43] remains unexplored.

3. Preliminary

3D Gaussian Splatting (3D-GS) [22] represents a scene using anisotopic 3D Gaussian primitives. Each gaussian is defined by a mean position $\mathbf{x}_i \in \mathbb{R}^3$ and a 3D covariance matrix $\mathbf{\Sigma}_i$, where the covariance matrix $\mathbf{\Sigma}_i$ is decomposed into a rotation matrix $\mathbf{R}_i$ and a scaling matrix $\mathbf{S}_i$. Appearance is modeled by opacity $\sigma_i \in \mathbb{R}$ and color $\mathbf{c}_i \in \mathbb{R}^k$ using spherical harmonics, and the parameters are optimized via rendering loss with a tile-based rasterizer and α-blending. 2D Gaussian Splatting (2D-GS) [17] extends this to more accurately reconstruct geometry using 2D Gaussian surfels. Each surfel is defined by a mean position $\mathbf{x}$, scaling $\mathbf{s} = (s_u, s_v)$, and rotation $\mathbf{r} = (r_u, r_v)$, where r_u and r_v are tangential vectors. To enhance surface modeling, ray-splat intersection is used. We adopt 2DGS as our preferred representation to more accurately model geometry surfaces for contact analysis.

4. Method

Given multi-view videos, $\mathcal{V}$, and their camera parameters, our method accurately reconstructs the geometry, appearance, and contacts between hands and objects in a manipulation scene. Figure 1 presents our pipeline.

Initialization. To accurately capture hand–object interactions, we extract clear and consistent hand and object masks from the input multi-view images. We employ Segment Anything V2 [42] to obtain the foreground segmentation masks, $\mathcal{M}$, which include both hands and objects.

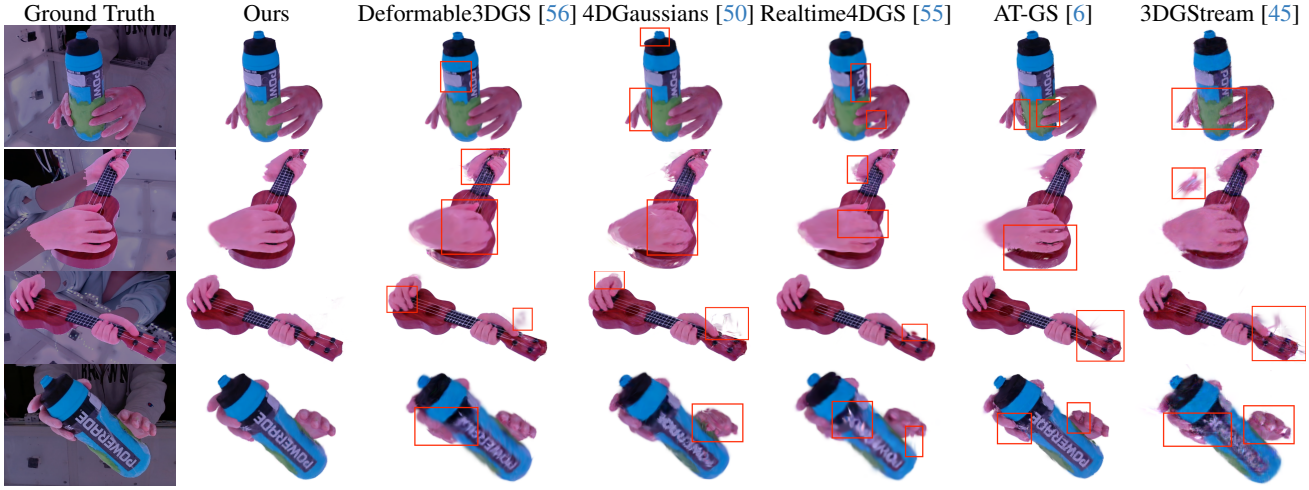

Figure 2. Qualitative comparisons on our dataset. Our method produces superior reconstruction quality with sharper novel view synthesis renderings, comparing with Deformable3DGS [56], 4DGaussians [50], Realtime4DGS [55], AT-GS [6], and 3DGStream [45]. Our method provides high quality reconstruction at the occluded regions, around the edges and at fine-grain details, where the baselines contains artifacts and blurriness. Zoom in for better views.

Facilitating the precise geometry and appearance modeling, we initialize a coarse hand surface representation using the MANO model [44]. A fully automated pipeline [10] estimates a sequence of MANO meshes $\mathcal{T}$ from the input multi-view videos to initialize the hand(s). Additionally, we initialize each object's geometry using coarse point clouds, $\mathcal{O}$, obtained either from offline scans or from the reconstruction of the first frame.

Template-based Gaussian Hand. To accurately capture hand surface geometry and appearance, we attach 2D Gaussian surfels to the triangular faces of the MANO mesh. Each surfel is defined in the local coordinate system of its parent triangle rather than moving freely in 3D space. With the MANO parameters $\mathcal{T}^j$ at time step j, the dynamics of each surfel are decoupled into a global transformation—driven by the parent triangle's motion in world coordinates—and a relative transformation within the triangle's local system. Following the approach in [39], we define each triangle's local coordinate system by setting its barycenter as the origin $\mathbf{T}$. We then construct a rotation matrix $\mathbf{R}$ by concatenating the direction vector of one edge, the triangle's normal vector, and their cross product. This matrix transforms coordinates from the triangle's local system to the global coordinate system. In the local system, each 2D Gaussian surfel is characterized by a mean position $\mathbf{x}$, rotation $\mathbf{r}$, and scaling $\mathbf{s}$. We represent the left and right hand as two separate groups of 2D Gaussian surfels, $\mathcal{G}_H = \{\mathcal{G}_H^{\text{left}}, \mathcal{G}_H^{\text{right}}\}$.

Object and Scene Composition. Given a sparse point cloud as initialization, we represent an object by a group of 2D Gaussian surfels in the world coordinate. We introduce a learnable parameter $\mathbf{P}$ to track the object's pose along the sequence. We represent all objects as $\mathcal{G}_O$ and the whole dynamic scene as $\mathcal{G} = \{\mathcal{G}_H, \mathcal{G}_O\}$. At timestep j, hands and objects are transformed to the deformed space by the corresponding MANO parameters $\mathcal{T}^j$ and pose parameters $\mathbf{P}^j$. During training, we adopt adaptive density control with binding inheritance [39] for hands and regular adaptive density control [17] for objects. During rendering, all the 2D Gaussian surfels are projected onto an image plane and rendered by a differentiable tile-based rasterizer.

Optimization. We supervise the rendered images by photometric loss, $\mathcal{L}_C$, defined in 3DGS [22]. Following 2DGS [17], we use the depth distortion term $\mathcal{L}_d$ to encourage concentration of surfels and the normal consistency loss $\mathcal{L}_n$ to approximate the surface. Following GaussianAvatars [39], we use two rigging regularization terms $\mathcal{L}_p$ and $\mathcal{L}_s$ to restrict the position and scale of hand surfels for better alignment with their parent triangles. We observe that Gaussian surfels become elongated in contact regions which result in artifacts in the estimated contact maps. Thus, to improve accuracy, we also introduce an isotropic regularization term, $\mathcal{L}_i$, that constrains the shape of Gaussian surfels. The overall loss function is:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_d + \lambda_2 \mathcal{L}_n + \lambda_3 \mathcal{L}_p + \lambda_4 \mathcal{L}_s + \lambda_5 \mathcal{L}_i, \quad (1)$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are $100, 0.005, 0.01, 1$ and 0.1 respectively.

Dynamic Contact Estimation. Leveraging our accurate hand and object surface models, we estimate hand–object contact at each frame by comparing their respective Gaussian surfels, following prior methods [9, 37, 47]. Specifically, for each 2D Gaussian surfel on the hand, we find the closest 2D Gaussian surfel on the object. The pair is considered to be in contact if the Euclidean distance is less than a pre-defined threshold τ.

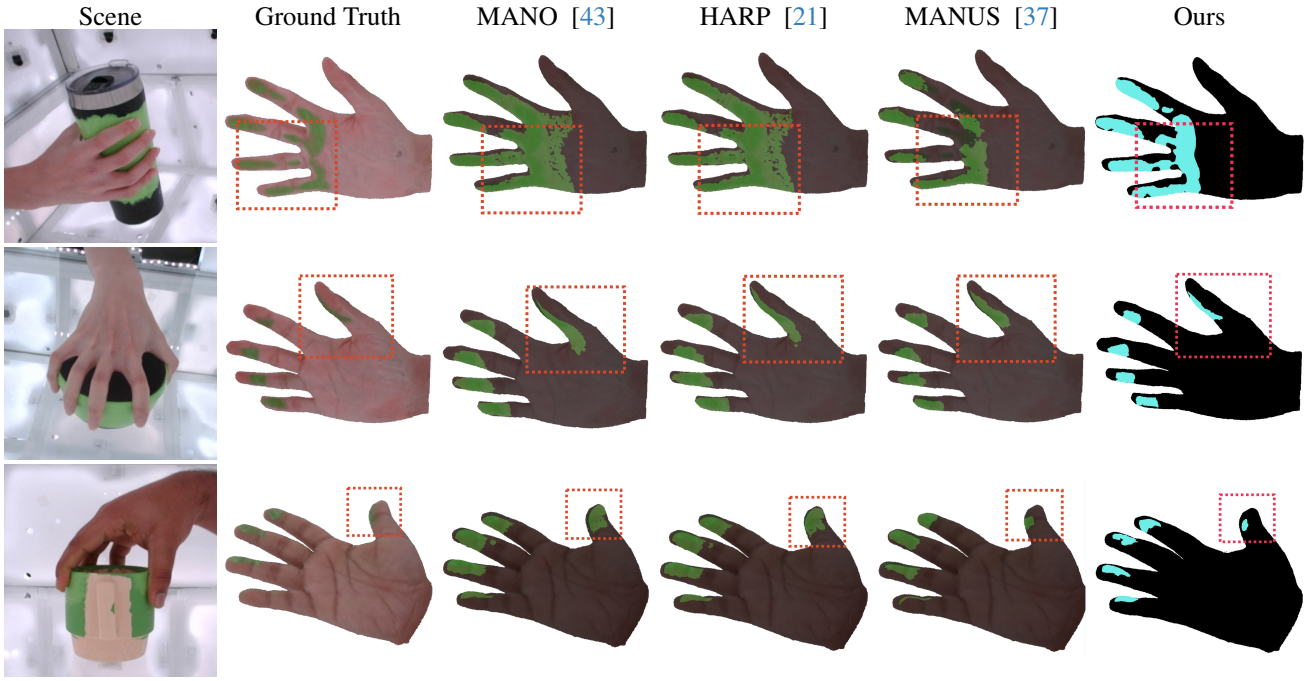

Figure 3. **Contact Comparisons**: We compare accumulated contacts of our method with MANUS, MANO, and HARP on ground truth contacts from MANUS-Grasps dataset. Our method provides more accurate contact estimation.

5. Experiments

5.1. Dataset

Benchmarking dynamic hand-object contact estimation is challenging because most real-world hand–object manipulation datasets [2, 9, 10, 24, 30–32, 35, 47, 59] lack ground-truth contact annotations. Although a few datasets provide ground truth via thermal imaging [4] or wet-paint transfer techniques [37], these methods are limited to grasping tasks with static objects rather than dynamic manipulation scenarios. To address this gap, we introduce the new dataset, which features diverse real-world hand–object manipulation sequences with ground-truth contacts. Specifically, our dataset integrates data from three multi-view hand–object interaction datasets: GigaHands [10], DiVa-360 [32], and MANUS-Grasps [37], which offers bimanual complex manipulation, long-range bimanual hand object interaction, and unimanual dynamic grasps, respectively. As these datasets do not include ground-truth contact labels for bimanual manipulations, we additionally captured six new complex bimanual manipulation sequences using an improved paint residue method [20, 37]. In total, there are 37 hand-object manipulation sequences, with 21 sequences containing ground truth contacts.

5.2. Baselines

For dynamic scene reconstruction, we compare our method with five state-of-the-art approaches: 4DGaussians [50], Deformable-3DGS [56], Realtime4DGS [57], 3DGStream [45], and AT-GS [6]. However, these methods do not naturally support dynamic contact estimation, as they do not differentiate between hands and objects. Therefore, to evaluate contact estimation, we select another group of baselines that are state-of-the-art analytical methods for contact estimation: MANO [43], HARP [21], and MANUS [37].

5.3. Evaluation on Dynamic Reconstruction

Qualitative Comparisons. Figure 2 presents qualitative comparisons of dynamic reconstruction on our dataset. Compared to baselines, our method presents higher reconstruction quality with clearer and more detailed reconstruction, especially in contact regions. Our approach accurately reconstructs both low-frequency non-interactive areas and high-frequency hand–object interactions. Moreover, while other unstructured Gaussian representations methods tend to generate floating artifacts, our template-based Gaussian representation enforces a strong inductive bias that reduces such artifacts with fewer Gaussian primitives, resulting in better 3D consistency.

Quantitative Comparisons. Table 1 presents quantitative comparisons on dynamic reconstruction with PSNR, SSIM and LPIPS [61] metrics. As indicated in Table 1, our method outperforms state-of-the-art baselines in all metrics across all scenes.

Dataset	GigaHands [10]				DiVa-360 [32]				MANUS-Grasps [37]			
Method—Metric	SSIM$^{\uparrow}$	PSNR$^{\uparrow}$	LPIPS$^{\downarrow}$	Mem	SSIM$^{\uparrow}$	PSNR$^{\uparrow}$	LPIPS$^{\downarrow}$	Mem	SSIM$^{\uparrow}$	PSNR$^{\uparrow}$	LPIPS$^{\downarrow}$	Mem
Deformable3DGS [56]	0.970	26.37	0.039	5MB	0.978	29.20	0.036	5MB	0.970	29.58	0.042	6MB
4DGaussians [50]	0.963	25.82	0.045	136MB	0.975	28.35	0.041	136MB	0.967	28.60	0.044	136MB
Realtime4DGS [57]	0.969	26.14	0.044	168MB	0.957	21.99	0.062	103MB	0.974	29.85	0.044	291MB
AT-GS [6]	0.972	26.62	0.038	380MB	0.979	28.41	0.033	160MB	0.972	29.40	0.039	122MB
3DGStream [45]	0.96	28.12	0.061	15MB	0.960	29.34	0.043	15MB	0.946	26.41	0.084	16MB
Ours	0.982	30.06	0.018	13MB	0.983	32.18	0.020	11MB	0.974	32.67	0.021	8MB

Table 1. **Reconstruction Quality Comparison.** We compare our dynamic reconstruction results against other Gaussian-based approaches on multiple datasets, showing that our method achieves superior performance in SSIM, PSNR, and LPIPS metrics with the highest memory efficiency.

	MANO	HARP	MANUS	Ours
mIoU$^{\uparrow}$	0.168	0.182	0.211	0.226
F1 score$^{\uparrow}$	0.279	0.299	0.343	0.378

Table 2. **Contact Accuracy Comparison.** We evaluate contact accuracy against other methods and demonstrate consistent improvements across all metrics.

5.4. Evaluation on Dynamic Contact Estimation

Qualitative Comparisons. Figure 3 provides qualitative comparisons on accumulated contact estimation with other methods. It shows that our method yields more accurate dynamic contact estimates that closely match the ground truth, unlike the over-segmentation seen in baseline methods. This improvement stems from our template-based Gaussian representation which enforces a strong inductive bias that reduces noise and removes floating artifacts around contact regions.

Quantitative Comparisons. Table 2 quantitatively evaluates contact estimation accuracy using Intersection over Union (IoU) and F1-score metrics [37] by comparing estimated and ground truth contact maps. As shown in Table 2, our method consistently outperforms all baselines on the dataset, which aligns with the visual results in Figure 3. Notably, while other Gaussian-based hand models[37] use approximately 300k Gaussians per sequence, achieves superior results with only about 10k Gaussians per sequence on average, underscoring its efficiency in dynamic contact estimation.

6. Conclusion

We introduced a novel markerless capture method that captures dynamic contacts for bimanual hand-object manipulations. The method leverages a dynamic articulated representation based on 2D Gaussian surfels to capture complex manipulations. Additionally, it takes advantage of the inductive bias from template models through binding surfels to MANO [43] meshes, which efficiently stabilizes and speeds up the optimization process. To evaluate dynamic contacts, we curate a new hand-object manipulation dataset with ground truth contacts. Extensive experiments on our dataset demonstrate that our method achieves state-of-the-art dynamic reconstruction quality and significantly improves the accuracy of dynamic contacts capturing.

Limitations & Future Work. While our primary focus in this paper is accurate dynamic contact estimation, we acknowledge that the complexity of hand and object dynamics in everyday life extends far beyond our current exploration. Our work has concentrated on modeling two hands manipulating rigid objects, without addressing the challenges posed by articulated or more general objects. We also see potential for enhancing the evaluation metrics for dynamic contacts in future works.

References

[1] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12, pages 640–653. Springer, 2012. 1, 2

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. arXiv preprint arXiv:2406.09598, 2024. 4

[3] Cristian Camilo Beltran-Hernandez, Damien Petit, Ixchel Georgina Ramirez-Alpizar, Takayuki Nishi, Shinichi Kikuchi, Takamitsu Matsubara, and Kensuke Harada. Learning force control for contact-rich manipulation tasks with rigid position-controlled robots. IEEE Robotics and Automation Letters, 5(4):5709–5716, 2020. 1

[4] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8709–8719, 2019. 1, 2, 4

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9044–9053, 2021. 2

[6] Decai Chen, Brianne Oberson, Ingo Feldmann, Oliver Schreer, Anna Hilsmann, and Peter Eisert. Adaptive and

temporally consistent gaussian surfels for multi-view dynamic reconstruction. arXiv preprint arXiv:2411.06602, 2024. 3, 4, 5

[7] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11, 2024. 1, 2

[8] Íñigo Elguea-Aguinaco, Antonio Serrano-Muñoz, Dimitrios Chrysostomou, Ibai Inziarte-Hidalgo, Simon Bøgh, and Nestor Arana-Arexolaleiba. A review on reinforcement learning for contact-rich robotic manipulation tasks. Robotics and Computer-Integrated Manufacturing, 81: 102517, 2023. 1

[9] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12943–12954, 2023. 2, 3, 4

[10] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities, 2024. 3, 4, 5

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 409–419, 2018. 2

[12] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations, 2018. 2

[13] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3196–3206, 2020. 1, 2

[14] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11807–11816, 2019. 2

[15] Guido Heumer, Heni Ben Amor, Matthias Weber, and Bernhard Jung. Grasp recognition with uncalibrated data gloves-a comparison of classification methods. In 2007 IEEE virtual reality conference, pages 19–26. IEEE, 2007. 2

[16] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 2

[17] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In SIGGRAPH 2024 Conference Papers. Association for Computing Machinery, 2024. 1, 2, 3

[18] Xuan Huang, Hanhui Li, Wanquan Liu, Xiaodan Liang, Yiqiang Yan, Yuhao Cheng, and CHENQIANG GAO. Learning interaction-aware 3d gaussian splatting for one-shot hand avatars. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. 1, 2

[19] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4220–4230, 2024. 1, 2

[20] Noriko Kamakura, Michiko Matsuo, Harumi Ishii, Fumiko Mitsuboshi, and Yoriko Miura. Patterns of static prehension in normal hands. The American journal of occupational therapy, 34(7):437–445, 1980. 1, 4

[21] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12802–12813, 2023. 4

[22] 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. 1, 2, 3

[23] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In European Conference on Computer Vision, pages 252–269. Springer, 2025. 1, 2

[24] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10138–10148, 2021. 1, 2, 4

[25] Jaeseong Lee, Taewoong Kang, Marcel C Bühler, Min-Jung Kim, Sungwon Hwang, Junha Hyung, Hyojin Jang, and Jaegul Choo. Surfhead: Affine rig blending for geometrically accurate 2d gaussian surfel head avatars. arXiv preprint arXiv:2410.11682, 2024. 2

[26] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph., 36(6):194–1, 2017. 2

[27] Yun Lin and Yu Sun. Grasp planning based on strategy extracted from demonstration. In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 4458–4463. IEEE, 2014. 2

[28] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21136–21145, 2024. 1, 2

[29] Xueyi Liu, Kangbo Lyu, Jieqiong Zhang, Tao Du, and Li Yi. Parameterized quasi-physical simulators for dexterous manipulations transfer. In European Conference on Computer Vision, pages 164–182. Springer, 2025. 1

[30] 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, pages 21013–21022, 2022. 2, 4

[31] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. arXiv preprint arXiv:2401.08399, 2024.

[32] Cheng-You Lu, Peisen Zhou, Angela Xing, Chandradeep Pokhariya, Arnab Dey, Ishaan Nikhil Shah, Rugved Mavidipalli, Dylan Hu, Andrew I Comport, Kefan Chen, et al. Diva-360: The dynamic visual dataset for immersive neural fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22466–22476, 2024. 4, 5

[33] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713, 2023. 1, 2

[34] Akshay Mundra, Jiayi Wang, Marc Habermann, Christian Theobalt, Mohamed Elgharib, et al. Livehand: Real-time and photorealistic neural hand rendering. In Proceedings of the IEEE/CVF international conference on computer vision, pages 18035–18045, 2023. 1

[35] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12999–13008, 2023. 4

[36] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. IEEE transactions on pattern analysis and machine intelligence, 40(12):2883–2896, 2017. 2

[37] Chandradeep Pokhariya, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2197–2208, 2024. 1, 2, 3, 4, 5

[38] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20299–20309, 2024. 2

[39] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians, 2024. 3

[40] Aashish Rai, Hiresh Gupta, Ayush Pandey, Francisco Vicente Carrasco, Shingo Jason Takagi, Amaury Aubel, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Towards realistic generative 3d face models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3738–3748, 2024. 2

[41] Aashish Rai, Dilin Wang, Mihir Jain, Nikolaos Sarafianos, Arthur Chen, Srinath Sridhar, and Aayush Prakash. Uvgs: Reimagining unstructured 3d gaussian splatting using uv mapping. arXiv preprint arXiv:2502.01846, 2025. 2

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 2

[43] 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. 1, 2, 4, 5

[44] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610, 2022. 3

[45] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20675–20685, 2024. 3, 4, 5

[46] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. Nature, 569(7758):698–702, 2019. 2

[47] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pages 581–600. Springer, 2020. 1, 2, 3, 4

[48] Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion, 2024. 2

[49] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19795–19806, 2023. 2

[50] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20310–20320, 2024. 1, 2, 3, 4, 5

[51] Qingxuan Wu, Zhiyang Dou, Sirui Xu, Soshi Shimada, Chen Wang, Zhengming Yu, Yuan Liu, Cheng Lin, Zeyu Cao, Taku Komura, Vladislav Golyanik, Christian Theobalt, Wenping Wang, and Lingjie Liu. Dice: End-to-end deformation capture of hand-face interactions from a single image, 2024. 2

[52] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics, 2024. 1, 2

[53] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head

avatar: Ultra high-fidelity head avatar via dynamic gaussians. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 2

[54] Jinbo Yan, Rui Peng, Luyang Tang, and Ronggang Wang. 4d gaussian splatting with scale-aware residual field and adaptive optimization for real-time rendering of temporally complex dynamic scenes. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 7871–7880, 2024. 1, 2

[55] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642, 2023. 3

[56] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20331–20341, 2024. 1, 2, 3, 4, 5

[57] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In International Conference on Learning Representations (ICLR), 2024. 4, 5

[58] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis, 2017. 2

[59] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 445–456, 2024. 4

[60] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. ACM Transactions on Graphics (ToG), 40(4):1–14, 2021. 1

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 4

[62] Xiang Zhang, Changhao Wang, Lingfeng Sun, Zheng Wu, Xinghao Zhu, and Masayoshi Tomizuka. Efficient sim-to-real transfer of contact-rich manipulation skills with online admittance residual learning. In Conference on Robot Learning, pages 1621–1639. PMLR, 2023. 1