# DyTact: Capturing Dynamic Contacts in Hand-Object Manipulation

Xiaoyan Cong[1] Angela Xing[1] Chandradeep Pokhariya[2] Rao Fu[1] Srinath Sridhar[1*]
[1]Brown University [2]IIT Delhi

## Abstract

*Reconstructing dynamic hand-object contacts is essential for realistic manipulation in AI character animation, XR, and robotics, yet it remains challenging due to heavy occlusions, complex surface details, and limitations in existing capture techniques. In this paper, we introduce DyTact, a markerless capture method for accurately capturing dynamic contact in hand–object manipulations in a non-intrusive manner. Our approach leverages a dynamic, articulated representation based on 2D Gaussian surfels to model complex manipulations. By binding these surfels to MANO [66] meshes, DyTact harnesses the inductive bias of template models to stabilize and accelerate optimization. A refinement module addresses time-dependent high-frequency deformations, while a contact-guided adaptive sampling strategy selectively increases surfel density in contact regions to handle heavy occlusion. Extensive experiments demonstrate that DyTact not only achieves state-of-the-art dynamic contact estimation accuracy but also significantly improves novel view synthesis quality, all while operating with fast optimization and efficient memory usage. Project Page: https://oliver-cong02.github.io/DyTact.github.io/.*

## 1. Introduction

Skillful object manipulation is one of the most common, yet impressive, human physical abilities. Human manipulation of objects is highly dynamic, and often bimanual, involving 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 [17]. Contact not only provides a measure of spatial proximity [6, 74, 98] but also conveys object affordances [31, 84, 90]. Moreover, variations in contact over time influence both the kinematics and dynamics [5, 47, 101] of the interaction.

Despite its significance, accurately capturing and reconstructing dynamic contacts remains a challenge. Traditional hardware-based approaches find it challenging to capture dynamic contacts. For example, instrumented gloves [26, 45, 73] can be intrusive, as they might constrain natural movements and affect tactile feedback, which compromises the realism and accuracy of capturing. Thermal sensors [6] can only estimate the accumulated contact in an interaction sequence. Therefore, they cannot sensitively capture instantaneous contact during dynamic manipulation. Moreover, most hardware solutions are expensive and difficult to scale up, limiting their potential applications.

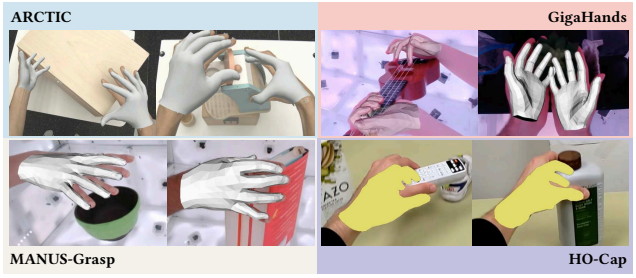


Figure 1. We demonstrate the **misalignment** between the parametric hand shape templates and actual hands by overlapping the rendering of actual hands with the annotated "Ground Truth" MANO meshes from four datasets, ARCTIC [18], GigaHands [20], MANUS-Grasp [58], HO-Cap [77]. Such misalignments exacerbate errors in dynamic contact estimation.

Recently, markerless capture [7, 10, 21, 25, 50, 58, 71] capable of estimating contact from visual inputs have emerged as a popular and affordable approach. These non-intrusive methods enable natural motion during capture and facilitate the subsequent reconstruction of geometry and appearance. Neural-hand [33, 54, 58] provides a data-driven markerless contact capture approach that requires vast amounts of diverse data to train a personalized hand avatar, which is inefficient and not user-friendly. Another prominent class of solutions leverages parametric hand shape templates, exemplified by MANO [66]. Although MANO-based approaches [3, 10, 18, 20, 49, 77, 90, 94, 97] exhibit strong generalizability across most scenarios. However, they frequently suffer from noticeable misalignments between the parametric template and the actual hand geometry, as shown in Fig. 1, which undermines the fidelity of fine-grained contact estimation. We argue that this limitation stems from the fact that MANO provides a generic, low-dimensional representation of hand shape, whereas ac-

*Corresponding author

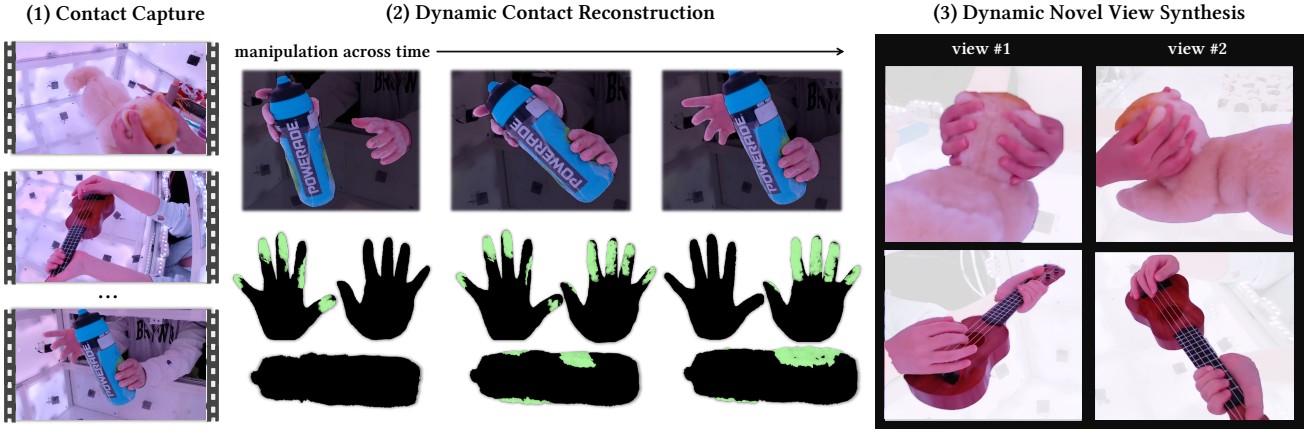


Figure 2. **DyTact** is a markerless Dynamic conTact capture method for complex hand-object manipulations. DyTact employs markerless contact capture (left) and uses a dynamic articulated representation based on 2D Gaussians to accurately model hand-object contacts without misalignments. Experimental results demonstrate that DyTact achieves accurate dynamic contact estimations (middle), and high-fidelity novel view synthesis (right).

curate contact estimation necessitates a personalized representation. Additionally, MANO pose estimation often lacks sufficient robustness and accuracy due to the presence of frequent occlusions and rapid motion during manipulation, thereby exacerbating errors in dynamic contact estimation.

To address the aforementioned limitations, we introduce **DyTact**, which captures Dynamic ConTact for complex hand-object manipulations. Our method is non-intrusive, accurate, efficient, and effective with occlusion. It employs multi-view markerless capture for natural manipulations, representing hands and objects with 2D Gaussian surfels [28] which accurately model surfaces and appearance, enabling accurate contact analysis while inherently avoiding the misalignment issues associated with purely parametric templates. Specifically, for each hand, we bind 2D Gaussian surfels to a parametric hand mesh [66] and optimize a refinement module to handle time-dependent complex surface deformation, ensuring fine detail alignment. Unlike grasping scenarios where the object typically remains static [58], our approach supports dynamic manipulation involving both in-hand and between-hand object motions. To facilitate this, we initialize a model-free 2D Gaussian surfel representation for the object and track its pose over time. Our explicit Gaussian-based representation allows for the efficient computation of both *instantaneous contacts* (defined as contacts at specific frames) and *accumulated contacts* (contacts aggregated over sequences), based on surfel pair distances. Furthermore, to address frequent occlusions, we introduce a contact-guided adaptive density control strategy, which selectively prunes surfels while maintaining accurate alignment with a minimal number of surfels.

In summary, our contributions are:

- We introduce DyTact, a method for accurate Dynamic conTact capture in complex hand-object manipulation.
- DyTact reconstructs **both the hand and the object** with dynamic 2D Gaussian surfels [28], enabling high-fidelity surface modeling without misalignments. We propose a contact-guided adaptive density control strategy to effectively address self-occlusions and object occlusions, as well as a time-dependent refinement module that precisely captures complex surface deformations for accurate contact estimation and dynamic reconstruction.
- Experimental results demonstrate the superior performance of DyTact in accurate dynamic contact estimation and high-fidelity novel view synthesis, coupled with fast optimization and efficient memory usage. The code and benchmark will be made publicly available upon acceptance.

## 2. Related Work

**Capturing and Modeling Contact.** Accurately capturing, understanding and modeling dynamic hand-object contact is essential for analyzing and replicating complex hand-object manipulations, a longstanding focus in both the graphics [9, 13, 39, 49, 69, 86, 98] and robotics communities [16, 24, 38, 47, 62, 81, 84, 95, 99, 102]. However, the human hand's intricate skeletal structure and extensive soft tissue continue to pose significant challenges.

Traditional methods have relied on instrumented gloves [26, 45, 73], specialized sensors [23, 57, 96], or thermal imaging [6] to capture contacts. However, these hardware-based approaches face difficulties in capturing dynamic contacts effectively. For instance, instrumented gloves [26, 45, 73] can be intrusive, as they may constrain natural movements and affect tactile feedback, thereby compromising the realism and quality of the capture. Thermal sensors [6], while allowing for relatively natural interaction, cannot by design capture dynamic or instantaneous contact; they can only estimate contact from resid-

ual thermal information after an interaction sequence has concluded. Since this residual thermal information can fade rapidly, such strategies require prompt and meticulously designed post-processing to record the contacts. Moreover, a common bottleneck for most hardware solutions is their high cost and difficulty to scale, limiting their widespread application.

Recent markerless motion capture approaches [7, 10, 21, 25, 50, 58, 71] enable more natural interactions during capture and are capable of estimating contacts from visual inputs. Nevertheless, the accuracy and efficiency of modeling and reconstructing dynamic contacts remain significant hurdles. Neural hand avatars [33, 54, 58], for example, require extensive training time to fit a personalized model from vast amounts of diverse dynamic sequences, offering a data-driven markerless contact estimation approach. However, this process is inefficient and not user-friendly for scalable deployment.

Another prominent class of solutions leverages parametric hand shape templates, exemplified by MANO [66], which is a generic, low-dimensional abstraction of human hands. Although MANO-based approaches [3, 10, 18, 20, 49, 77, 90, 94, 97] exhibit good generalizability in many scenarios, they often suffer from significant misalignments between the parametric template and the actual hand (as shown in Fig. 1). We argue that such misalignments are primarily caused by the generic nature of template models, as accurate alignment necessitates a personalized hand representation. Additionally, the MANO pose estimation process itself may lack sufficient robustness and accuracy, particularly when dealing with rapid motion and frequent self-occlusions or occlusions during hand-object interaction, thereby introducing further misalignments. Consequently, these misalignments hinder the capture of fine-grained contact details, compromising the accuracy of contact estimation. Benefiting from the pixel-wise supervision offered by the Gaussian Splatting-based representation, DyTact faithfully reconstructs both hands and objects without misalignments, achieving significant improvements in both dynamic contact capture accuracy and novel view synthesis quality.

**Dynamic Scene Representation.** Neural representations have recently become a prominent paradigm for scene modeling, achieving considerable progress in novel view synthesis. Neural Radiance Fields (NeRFs) [53], for instance, produce photorealistic renderings through differentiable volume rendering. Numerous methods extend NeRF to dynamic scenarios [1, 8, 14, 19, 22, 30, 35, 41, 43, 44, 56, 59, 68, 70, 76, 78, 79, 88]. DyNeRF [43] leverages time-conditioned latent codes to model dynamics. Tensor4D [68], K-Planes [19], and HexPlane [8] exploit planar factorization in the 4D spacetime domain. MixVoxels [76] adopts a mixture of static and dynamic voxel representations for faster rendering. Despite impressive results, most of these volumetric methods still require extensive sampling, which can hinder real-time rendering [4, 11, 53].

3D Gaussian Splatting (3D-GS) [34] achieves photorealistic novel view synthesis with rapid training speeds and real-time rendering, emerging as a compelling alternative for static scene modeling. Building on 3D-GS, a significant body of work on dynamic scenes has subsequently emerged [15, 29, 36, 46, 52, 64, 82, 85, 89, 91, 92]. Dynamic3DGS [52] learns transformations of Gaussian primitives over time, facilitating dynamic tracking [80]. Deformable3DGS [92] and 4DGS [82] define a deformation field mapping canonical Gaussian primitives to specific time steps. However, this approach faces challenges with new content appearing or disappearing. RealTime4DGS [91] and 4D-Rotor-GS [15] introduce 4D Gaussian primitives, improving flexibility for a variety of dynamic scenes. While these techniques perform well in general scenarios, few offer a dedicated approach for articulated objects such as hands. Their unrestricted Gaussian placement makes it difficult to incorporate kinematic structures or generalize to new sequences and environments. Consequently, driving these methods with a predefined hand model or employing them in complex, real-world manipulations remains a significant challenge.

Efforts like GaussianAvatars [60] and SurFhead [40] apply Gaussian Splatting to rig parametric models (e.g., FLAME [42]) for head avatar reconstruction [27, 63, 75, 83, 87], highlighting the potential of template-driven Gaussian Splatting. However, exploiting dynamic reconstruction for complex hand-object manipulations with template models such as MANO [66] remains largely unexplored. This paper introduces the first MANO-based 2D Gaussian hand representation, achieving accurate dynamic contact capture and high-fidelity novel view synthesis for complex manipulation scenarios.

## 3. Preliminary

**Gaussian Splatting.** Given multi-view images and a sparse point cloud of a scene, 3D Gaussian Splatting (3D-GS) [34] represents the scenes as a collection of anisotropic 3D Gaussians primitives $G = \{\mathbf{g}_i\}_{i=1}^{N}$. The geometry of each 3D Gaussian primitive is characterized by a mean position $\mathbf{x}_i \in \mathbb{R}^3$ and a 3D covariance matrix $\mathbf{\Sigma}_i$:

$$\mathbf{g}_i(\mathbf{x}) = \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T\mathbf{\Sigma}_i(\mathbf{x}-\mathbf{x}_i))$$

where the covariance matrix $\mathbf{\Sigma}_i$ is decomposed into a rotation matrix $\mathbf{R}_i$ and a scaling matrix $\mathbf{S}_i$ as $\mathbf{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^T\mathbf{R}_i^T$. $\mathbf{R}_i$ and $\mathbf{S}_i$ are stored as a rotation quaternion $\mathbf{r}_i \in \mathbb{R}^4$ and a scaling factor $\mathbf{s}_i \in \mathbb{R}^3$ respectively for independent optimization. The appearance of each 3D Gaussian is modeled by an opacity value $\sigma_i \in \mathbb{R}$ and a color $\mathbf{c}_i \in \mathbb{R}^k$

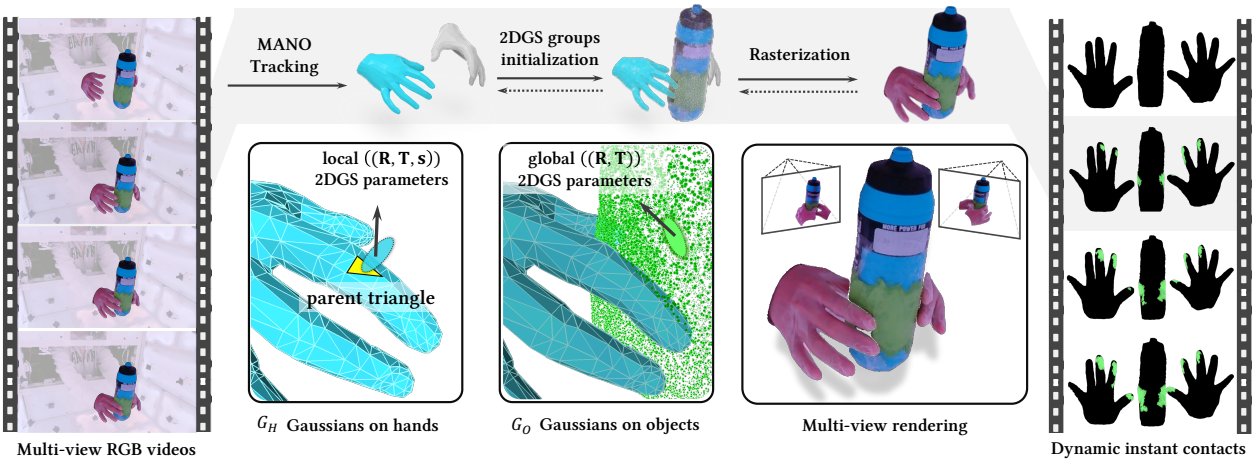


Figure 3. **DyTact** captures dynamic contacts with a markerless system using a surface-aware dynamic articulated Gaussian representation. Given multi-view RGB videos, it initializes Gaussian surfels on the hand by binding them to the tracked MANO mesh locally, which remain rigged throughout optimization. For objects, Gaussians are initialized by placing a coarse point cloud in the global coordinate space. A refinement module addresses time-dependent high-frequency deformations. A contact-guided adaptive sampling strategy selectively refines surfel density in contact regions to handle heavy occlusion, and further optimization of the surfels' geometry and appearance parameters ensures high-fidelity reconstructions that enable accurate contact estimation.

defined via $k$ spherical harmonics (SH) coefficients. Overall, $G = \{\mathbf{g}_i = \{\mathbf{x}_i, \mathbf{r}_i, \mathbf{s}_i, \sigma_i, \mathbf{c}_i\}\}_{i=1}^{N}$. The properties of Gaussians are optimized by minimizing the rendering loss between the synthesized and reference images. To render an image, a tile-based differentiable rasterizer projects Gaussians onto the image plane and applies standard $\alpha$-blending.

2D Gaussian Splatting (2D-GS) [28] extends 3D-GS to model more accurate geometry reconstruction by changing 3D Gaussian primitives to 2D Gaussian surfels. Each 2D Gaussian surfel is geometrically defined by its mean position $\mathbf{x}$, scaling $\mathbf{s} = (s_u, s_v)$, and rotation $\mathbf{r} = (r_u, r_v)$, where $r_u$ and $r_v$ and two principal tangential vectors and $s_u$ and $s_v$ control the corresponding variances. Ray-splat Intersection is used in the splatting process to enhance surface modeling quality. In this paper, we adopt 2DGS as our preferred representation to more accurately model geometry surfaces for contact analysis.

# 4. DyTact

Given multi-view videos $\mathcal{V} = \{\mathcal{V}_i : \{\mathbf{I}_i^j\} | 1 \leq i \leq N, 1 \leq j \leq T\}$ with $N$ views, $T$ frames, and their camera parameters, DyTact accurately reconstructs the geometry, appearance, and contacts between hands and objects in a manipulation sequence. We achieve this by representing the dynamic scene using time-varying surface geometry and analytically estimating contacts. Specifically, we optimize separate sets of 2D Gaussian surfels for both hands and objects to enable detailed and efficient modeling. Figure 3 presents our pipeline. In this section, we first describe our data pre-processing and initialization steps in Section 4.1. We then introduce a template-based Gaussian representation for hands in Section 4.2 with time-dependent deformation refinement. Section 4.3 describes how we represent the object and how we compose the hand and object to model the entire scene. Given our hand and object representations, Section 4.4 describes the dynamic contact estimation process. Finally, we detail our optimization strategy with contact-guided adaptive density control handling occlusions in Section 4.5.

## 4.1. Initialization

To accurately capture hand–object manipulations, we extract clear and consistent hand and object masks from the input multi-view images. We employ Segment Anything V2 [65] to obtain the foreground segmentation masks $\mathcal{M} = \{\mathcal{M}_i^j | 1 \leq i \leq N, 1 \leq j \leq T\}$, which include both hands and objects. Facilitating the precise geometry and appearance modeling, we initialize a coarse hand surface representation using the MANO model [67]. A fully automated pipeline [20] estimates a sequence of MANO meshes $\mathcal{T} = \{\mathcal{T}^j : \{\theta^j, \beta^j, R^j, T^j\} | 1 \leq j \leq T\}$ from the input multi-view videos $\mathcal{V}$ to initialize the hand(s), where $\theta$, $\beta$, $R$, and $T$ denote the pose, shape, relative rotation, and translation, respectively. Similarly, 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. In summary, our inputs consist of $\{\mathcal{V}, \mathcal{M}, \mathcal{T}, \mathcal{O}\}$ .

## 4.2. 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. This decoupling introduces an inductive bias that significantly accelerates the convergence of Gaussian fitting.

Following the approach in [61], 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 anisotropic scaling $\mathbf{s}$. These attributes are transformed to the world coordinates as follows:

$$(\mathbf{r}',\, \mathbf{x}',\, \mathbf{s}') = \big(\mathbf{R}\mathbf{r},\; s\,\mathbf{R}\mathbf{x} + \mathbf{T},\; s\,\mathbf{s}\big), \tag{1}$$

where $s$ is an isotropic scale representing the triangle's area. This scale factor is crucial for speeding up and stabilizing the optimization process, as it adjusts the step size relative to the triangle's area, where larger triangles naturally allow for larger step sizes.

As one single surfel per triangle is insufficient to capture the subtle interactions between hands and objects, to improve contact estimation capability, we randomly initialize $k$ 2D Gaussian surfels within each triangle's local coordinate system by sampling from a Gaussian distribution centered at $\mathbf{T}$ with variance $v$. Empirically, we set $k = 5$ and $v = 0.5$. The left hand and right hand are represented by two separate groups of such 2D Gaussian surfels as $\mathcal{G}_H = \{\mathcal{G}_H^{\text{left}}, \mathcal{G}_H^{\text{right}}\}$. With the binding relationship, our template-based Gaussian representation can be driven by any sets of MANO parameters.

**Time-dependent Hand Deformation Refinement.** Although such template-based Gaussian representation can model approximate dynamics, some time-dependent high-frequency deformations might be challenging to capture accurately, such as the complex deformations of the hand skin near the contacting region. To address this, we introduce a refinement module $\mathcal{R}_\theta$ that compensates these time-dependent high-frequency deformations. Taking the attributes $\{\mathbf{x}', \mathbf{r}', \mathbf{s}'\}$ of a Gaussian surfel in the world coordinate, and the corresponding timestamp $j$ as inputs, the refinement module $\mathcal{R}_\theta$ outputs the offsets of that Gaussian surfel's attributes $\{\delta\mathbf{x}', \delta\mathbf{r}', \delta\mathbf{s}'\}$ in the world coordinate:

$$\begin{aligned}(\delta\mathbf{x}', \delta\mathbf{r}', \delta\mathbf{s}') = \mathcal{R}_\theta(&\gamma(SG(\mathbf{x}'), L_x), \gamma(SG(\mathbf{r}'), L_r),\\ &\gamma(SG(\mathbf{s}'), L_s), \gamma(j, L_j)),\end{aligned} \tag{2}$$

where $\mathcal{R}_\theta$ is parameterized as an 8-Layer MLP with hidden dimension 256, $SG(\cdot)$ represents the step-gradiant operation, $\gamma(\cdot, L)$ is the same positional encoding as [53]. Empirically, we set $L_x = 8$, $L_r = 4$, $L_s = 4$, $L_j = 4$. Then the mean position, rotation and scaling in the world coordinate can be updated as $\{\mathbf{x}' + \delta\mathbf{x}', \mathbf{r}' + \delta\mathbf{r}', \mathbf{s}' + \delta\mathbf{s}'\}$ respectively.

### 4.3. Object Representation and Scene Composition

Given a sparse point cloud as initialization, we represent an object by a group of 2D Gaussian surfels $\{\mathbf{g}_i = \{\mathbf{x}_i, \mathbf{r}_i, \mathbf{s}_i, \sigma_i, \mathbf{c}_i\}\}_{i=1}^N$ in the world coordinate. We introduce a learnable parameter $\mathbf{P} = \{\mathbf{q}, \mathbf{t}\}$ to track the object's pose along the sequences, where the quaternion $\mathbf{q}$ and vector $\mathbf{t}$ indicate the rotation and translation respectively. Thus, all the objects can be represented as $\mathcal{G}_O = \{\mathcal{G}_O^o | o = 1, 2, \cdots\}$, where $\mathcal{G}_O^o = \{\{\mathbf{g}_i\}_{i=1}^N, \{\mathbf{P}^j\}_{j=1}^T\}$. The entire dynamic scene can be composed as $\mathcal{G} = \{\mathcal{G}_H, \mathcal{G}_O\}$. At timestep $j$, hands and objects are driven to the deformed space by the corresponding MANO parameters $\mathcal{T}^j$ and pose parameters $\mathbf{P}^j$. During rendering, all the 2D Gaussian surfels are projected onto an image plane and rendered by a differentiable tile-based rasterizer. During training, we adopt adaptive density control with binding inheritance [61] for hands and regular adaptive density control [28] for objects.

### 4.4. 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 analytical methods [18, 58, 74]. Specifically, for each 2D Gaussian surfel on the hand, we find the closest 2D Gaussian surfel on the object. This pair is considered to be in contact if the distance is less than a pre-defined threshold $\tau$. Given a group of hand Gaussian surfels $\mathcal{G}_H$ and object Gaussian surfels $\mathcal{G}_O$, the contact map between them is defined as:

$$\mathrm{C} = \begin{cases} \mathbf{D}(\mathbf{g}_H, \mathbf{g}_O) & \text{if } \mathbf{D}(\mathbf{g}_H, \mathbf{g}_O) < \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\mathbf{g}_H \in \mathcal{G}_H$, $\mathbf{g}_O \in \mathcal{G}_O$, and $\mathbf{D}$ represents the Euclidean distance between positions of two 2D Gaussian surfels.

### 4.5. Optimization

**Contact-Guided Adaptive Density Control.** To model regions with varying complexity more efficiently, adaptive density control strategy [34] provides a general guidance: intuitively, it allocates more Gaussian primitives to regions with higher complexity and vice versa. In hand-object manipulation scenes, self-occlusions and hand-object occlusions where interactions happen are more complicated and harder to fit. We observe that Gaussian surfels with weird shapes such as abnormal narrow long disks cause artifacts in the rendered contact maps. To capture more accurate contacts, we propose a contact-guided adaptive density control strategy with motivations of allocating more

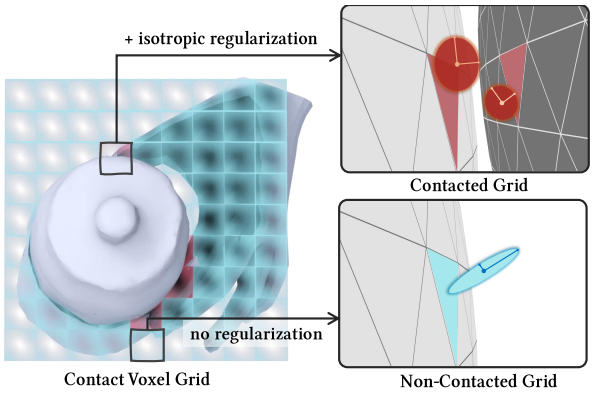


Figure 4. **Contact-guided adaptive sampling strategy** ensures Gaussians remain well-regularized in occluded areas. The entire space was voxelized and labeled based on contact status. An isotropic regularization is applied to Gaussian surfels within contacting regions, effectively preserving and accumulating informative gradients around contacting regions.

Gaussian surfels around the contacting regions and making them as isotropic as possible. As shown in Figure 4, We first voxelize the whole scene with a voxel size $\tau_v$. We set $\tau_v = \tau/\sqrt{3}$ and $\tau$ is the pre-defined contact threshold in Section 4.4. A voxel is identified as *contact-voxel* when it contains both hand Gaussian surfels and object Gaussian surfels. Then we introduce an isotropic regularization term $\mathcal{L}_i$ to constrain the variances of two orthogonal axes of Gaussian surfels within *contact-voxels*:

$$\mathcal{L}_i = ||\frac{\min_{\mathbf{s}}}{\max_{\mathbf{s}}} - \tau_{\mathbf{s}}||_2 \tag{4}$$

where $\min_{\mathbf{s}} \in \mathbb{R}^3$ and $\max_{\mathbf{s}} \in \mathbb{R}^3$ are the minimum and maximum scales and $\tau_{\mathbf{s}} = 0.4$ is the ratio threshold of scales.

**Overall Loss Terms.** We supervise the rendered images by photometric loss $\mathcal{L}_C$ combining $\mathcal{L}_1$ with a D-SSIM term [28, 34] $\mathcal{L}_{\text{D\_SSIM}}$:

$$\mathcal{L}_c = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D\_SSIM}} \tag{5}$$

where $\lambda = 0.2$. Following 2DGS [28], we use the depth distortion term $\mathcal{L}_d$ that encourages the concentration of the 2D Gaussian surfels by adjusting the intersection depth and the normal consistency loss $\mathcal{L}_n$ that encourages 2D Gaussian surfels locally approximate the surface by aligning their normals with the estimated surface normals. Following GaussianAvatars [61], we use two rigging regularization terms $\mathcal{L}_p$ and $\mathcal{L}_s$ to restrict the position and scale of hands' Gaussian surfels for better alignment with their parent triangles. The overall loss function is:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1\mathcal{L}_d + \lambda_2\mathcal{L}_n + \lambda_3\mathcal{L}_p + \lambda_4\mathcal{L}_s + \lambda_5\mathcal{L}_i, \tag{6}$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, and $\lambda_5$ are $100, 0.005, 0.01, 1$ and $0.1$ respectively.

# 5. Experiments

## 5.1. *DyTact*-21 Dynamic Contact Benchmark

Evaluating dynamic hand-object contacts remains a significant challenge due to the difficulty of obtaining reliable ground truth data. Existing real-world hand–object manipulation datasets [2, 18, 20, 37, 48, 50, 51, 55, 74, 97] lack ground-truth contact annotations, limiting their utility for quantitative evaluation. To address this gap, we introduce *DyTact*-21, a new benchmark featuring diverse real-world hand–object manipulation sequences with accurate ground-truth accumulated contacts. *DyTact*-21 includes 15 single-hand grasping sequences from MANUS-Grasps [58], along with 6 newly captured complex bimanual manipulation sequences.

These additional sequences were collected using an enhanced version of the wet-paint residue method [32, 58], redesigned to support scalable capture of complex interactions. We defer the details of our capture procedure to the supplementary material.

In total, *DyTact*-21 provides 21 real-world tabletop manipulation sequences with ground-truth accumulated contact, benchmarking dynamic contact estimation methods.

## 5.2. Dynamic Reconstruction Datasets

To evaluate dynamic reconstruction, we compare with baselines the performance of novel view synthesis on three public datasets: DiVa-360 [51], MANUS-Grasps [58], and GigaHands [20], all of which provide high-quality, multi-view recordings of real-world tabletop manipulation tasks. From *DiVa-360*, we select 3 daily-life, bi-manual tabletop manipulation sequences, each averaging 400 frames. We choose 20 tabletop grasping sequences with an average of 200 frames per sequence from *MANUS-Grasps* [58], and 8 complex bimanual hand-object interaction sequences, each averaging 300 frames from *GigaHands* [20]. In total, we evaluate dynamic reconstruction performance on 31 sequences. For each sequence, we use two camera views for testing and the remaining views for training. All metrics are calculated as the average over all testing frames.

## 5.3. Baselines

For dynamic scene reconstruction, we compare our method with five state-of-the-art approaches: 4DGaussians [82], Deformable-3DGS [92], Realtime4DGS [93], 3DGStream [72], and AT-GS [12]. However, these methods do not naturally support dynamic contact estimation, as they do not differentiate between hands and objects as DyTact. Therefore, to evaluate contact estimation, we select another group of baselines that are state-of-the-art analytical methods for contact estimation: MANO [66], HARP [33], and MANUS [58].

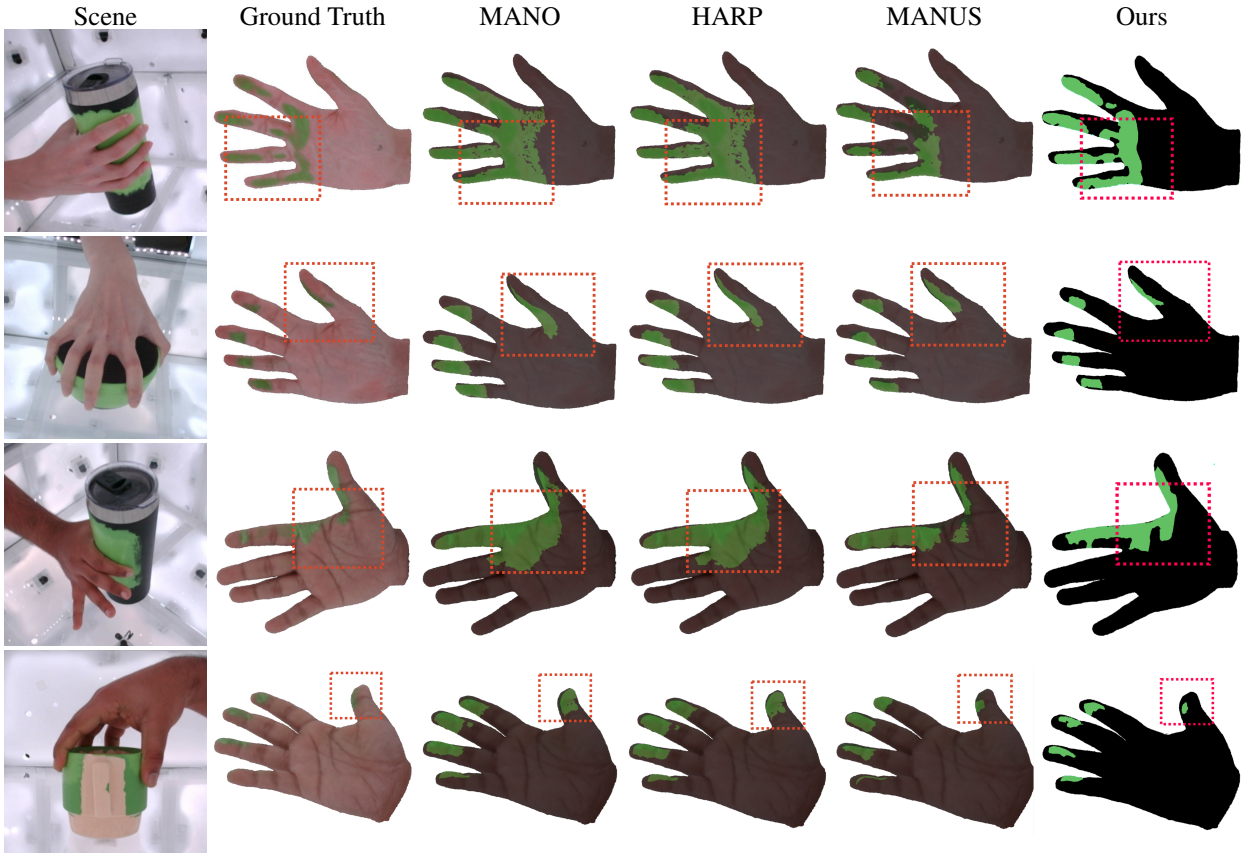


Figure 5. **Qualitative Comparisons on Contact Estimation.** We compare the accumulated contact estimation of DyTact against MANUS [58], MANO [66], and HARP [33] on sequences from *DyTact*-21 benchmark. DyTact yields more accurate and coherent accumulated contact estimation that closely aligns with the ground truth, contrasting sharply with the over-segmentation and noisy artifacts observed in baselines.

| Dataset | GigaHands | | | | DiVa-360 | | | | MANUS-Grasps | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method\Metric | SSIM$^{\uparrow}$ | PSNR$^{\uparrow}$ | LPIPS$^{\downarrow}$ | Mem$^{\downarrow}$ | SSIM$^{\uparrow}$ | PSNR$^{\uparrow}$ | LPIPS$^{\downarrow}$ | Mem$^{\downarrow}$ | SSIM$^{\uparrow}$ | PSNR$^{\uparrow}$ | LPIPS$^{\downarrow}$ | Mem$^{\downarrow}$ |
| Deformable3DGS | 0.970 | 26.37 | 0.039 | 5MB | 0.978 | 29.20 | 0.036 | 5MB | 0.970 | 29.58 | 0.042 | 6MB |
| 4DGaussians | 0.963 | 25.82 | 0.045 | 136MB | 0.975 | 28.35 | 0.041 | 136MB | 0.967 | 28.60 | 0.044 | 136MB |
| Realtime4DGS | 0.969 | 26.14 | 0.044 | 168MB | 0.957 | 21.99 | 0.062 | 103MB | 0.974 | 29.85 | 0.044 | 291MB |
| AT-GS | 0.972 | 26.62 | 0.038 | 380MB | 0.979 | 28.41 | 0.033 | 160MB | 0.972 | 29.40 | 0.039 | 122MB |
| 3DGStream | 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. **Quantitative Comparisons on Dynamic Reconstruction.** We compare DyTact with other Gaussian-based approaches regarding novel view synthesis quality, demonstrating the superior performance of DyTact across SSIM, PSNR, and LPIPS metrics with efficient memory usage.

| | 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. **Quantitative Comparisons on the Accumulated Contacts Estimation.** We evaluate the accuracy of the accumulated contacts estimation against other methods and demonstrate consistent improvements across all metrics.

### 5.4. Evaluation on Dynamic Contact Estimation

**Qualitative Comparisons.** Figure 5 presents qualitative comparisons of accumulated contact estimation on *DyTact*-21 benchmark against baseline methods. Our method produces more accurate and coherent dynamic contact maps that closely align with the ground truth, in contrast to the over-segmentation and noisy artifacts observed in other approaches. This improvement is attributed to three key factors: (1) our template-based Gaussian representation enforces a strong inductive bias that reduces noise and removes floating artifacts around contact regions; (2) our deformation refinement module effectively captures time-dependent surface deformation; (3) our contact-guided adaptive density control strategy allocates additional Gaus-

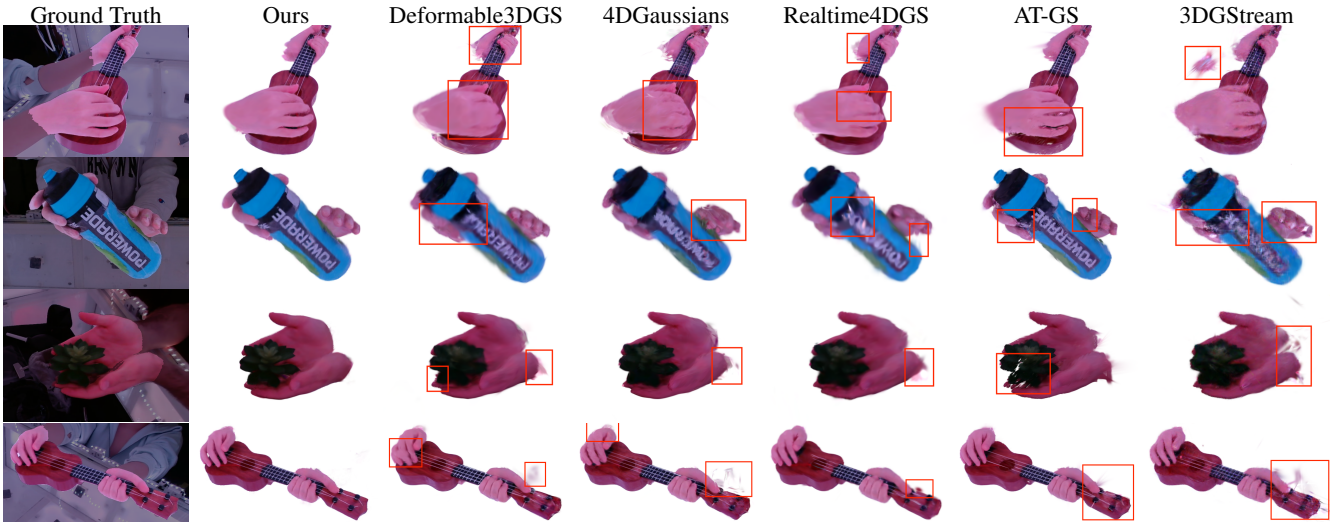


Figure 6. **Qualitative Comparisons on Novel View Synthesis.** DyTact produces superior reconstruction quality with sharper novel view synthesis renderings. Compared to baselines, DyTact delivers more fine-grained details, particularly in occluded regions and around edges, whereas baseline methods exhibit artifacts and blurriness. Please zoom in for better views.

sian primitives to high-variation regions, allowing for fine-grained reconstruction of contact details.

**Quantitative Comparisons.** Table 2 quantitatively evaluates contact estimation accuracy using Intersection over Union (IoU) and F1-score metrics [58] by comparing estimated and ground truth contact maps. As shown in Table 2, our method consistently outperforms all baselines on the *DyTact*-21 dataset, which aligns with the visual results in Figure 5. Notably, while other Gaussian-based hand models[58] use approximately 300k Gaussians per sequence, DyTact achieves superior results with only about 10k Gaussians per sequence on average, underscoring its efficiency in dynamic contact estimation.

### 5.5. Evaluation on Dynamic Reconstruction

**Qualitative Comparisons.** Figure 6 presents qualitative comparisons of novel view synthesis. 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, thanks to our contact-guided adaptive density control strategy. 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 [100] metrics. As indicated in Table 1, our method outperforms state-of-the-art baselines in all metrics across all scenes, demonstrating the superior performance and generalizability of DyTact. Notably, we achieve these improvements while maintaining fast optimization and efficient memory usage, underscoring the efficiency of our approach in modeling challenging hand–object manipulation scenes.

## 6. Conclusion

This paper introduces DyTact, a novel markerless method for capturing dynamic contacts in complex hand-object manipulations. The method utilizes a dynamic articulated representation based on 2D Gaussian surfels to effectively capture complex manipulations. DyTact takes advantage of the inductive biases from template models by binding surfels to MANO [66] meshes, thereby efficiently stabilizing and accelerating the optimization process. A dedicated refinement module addresses time-dependent high-frequency deformations, while a contact-guided adaptive sampling strategy selectively refines surfel density around contacting regions.

To evaluate dynamic contacts, we have curated a new benchmark, *DyTact*-21, which provides ground-truth accumulated contact data for 21 complex and diverse real-world multi-view manipulation sequences. Extensive experiments demonstrate that DyTact achieves state-of-the-art accuracy in dynamic contact capture and significantly improves dynamic reconstruction quality, leading to high-fidelity novel view synthesis. The source code and benchmark will be made publicly available upon paper acceptance.

## References


[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16610–16620, 2023. 3

[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. 6

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025. 1, 3

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 3

[5] 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

[6] 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

[7] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 1, 3

[8] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 3

[9] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. 2

[10] 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. 1, 3

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 3

[12] 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. 6

[13] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[14] Xiaoyan Cong, Haitao Yang, Liyan Chen, Kaifeng Zhang, Li Yi, Chandrajit Bajaj, and Qixing Huang. 4drecons: 4d neural implicit deformable objects reconstruction from a single rgb-d camera with geometrical and topological regularizations. *arXiv preprint arXiv:2406.10167*, 2024. 3

[15] 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. 3

[16] Ryan Elandt, Evan Drumwright, Michael Sherman, and Andy Ruina. A pressure field model for fast, robust approximation of net contact force and moment between nominally rigid objects. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8238–8245. IEEE, 2019. 2

[17] 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

[18] 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. 1, 3, 5, 6

[19] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 3

[20] 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. 1, 3, 4, 6

[21] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. Ieee, 2009. 1, 3

[22] Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4d: Voxel for 4d novel view synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 3

[23] 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

[24] Abhishek Gupta, Clemens Eppner, Sergey Levine, and Pieter Abbeel. Learning dexterous manipulation for a soft robotic hand from human demonstrations. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3786–3793. IEEE, 2016. 2

[25] 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, 3

[26] 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. 1, 2

[27] 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. 3

[28] 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. 2, 4, 5, 6

[29] 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. 3

[30] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 3

[31] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14713–14724, 2023. 1

[32] 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. 6

[33] 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. 1, 3, 6, 7

[34] 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. 3, 5, 6

[35] Seoha Kim, Jeongmin Bae, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Sync-nerf: Generalizing dynamic nerfs to unsynchronized videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2777–2785, 2024. 3

[36] 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. 3

[37] 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. 6

[38] Arjun Lakshmipathy, Dominik Bauer, and Nancy S Pollard. Contact tracing: A low cost reconstruction framework for surface contact interpolation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5165–5172. IEEE, 2021. 2

[39] Arjun Sriram Lakshmipathy, Nicole Feng, Yu Xi Lee, Moshe Mahler, and Nancy Pollard. Contact edit: Artist tools for intuitive modeling of hand-object interactions. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 2

[40] 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. 3

[41] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems*, 35:13485–13498, 2022. 3

[42] 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. 3

[43] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 3

[44] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–9, 2023. 3

[45] 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. 1, 2

[46] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaus-

sian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 3

[47] 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, 2

[48] 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. 6

[49] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. *arXiv preprint arXiv:2411.14280*, 2024. 1, 2, 3

[50] 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. 1, 3, 6

[51] 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. 6

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

[53] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3, 5

[54] 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, 3

[55] 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. 6

[56] Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Representing volumetric videos as dynamic mlp maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4252–4262, 2023. 3

[57] 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

[58] 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, 5, 6, 7, 8

[59] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

[60] 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. 3

[61] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians, 2024. 5, 6

[62] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4):10873–10881, 2022. 2

[63] 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. 3

[64] 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. 3

[65] 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. 4

[66] 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, 3, 6, 7, 8

[67] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 4

[68] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 3

[69] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori.

Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 2

[70] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3

[71] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 294–310. Springer, 2016. 1, 3

[72] 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. 6

[73] 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. 1, 2

[74] 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, 5, 6

[75] Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion, 2024. 3

[76] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19706–19716, 2023. 3

[77] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction, 2025. 1, 3

[78] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13524–13534, 2022. 3

[79] Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. Neural residual radiance fields for streamably free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 76–87, 2023. 3

[80] 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. 3

[81] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 2

[82] 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. 3, 6

[83] 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. 3

[84] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning*, pages 618–629. PMLR, 2023. 1, 2

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

[86] Pei Xu and Ruocheng Wang. Synchronize dual hands for physics-based dexterous guitar playing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[87] 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. 3

[88] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20040, 2024. 3

[89] 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. 3

[90] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 1, 3

[91] 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

[92] 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. 3, 6

[93] 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. 6

[94] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 1, 3

[95] Jessica Yin, Paarth Shah, Naveen Kuppuswamy, Andrew Beaulieu, Avinash Uttamchandani, Alejandro Castro, James Pikul, and Russ Tedrake. Proximity and visuotactile point cloud fusion for contact patches in extreme deformation. *arXiv preprint arXiv:2307.03839*, 2023. 2

[96] 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

[97] 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. 1, 3, 6

[98] 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, 2

[99] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024. 2

[100] 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. 8

[101] 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

[102] Yunbo Zhang, Alexander Clegg, Sehoon Ha, Greg Turk, and Yuting Ye. Learning to transfer in-hand manipulations using a greedy shape curriculum. In *Computer graphics forum*, pages 25–36. Wiley Online Library, 2023. 2