arXiv:2312.02137v1 [cs.CV] 4 Dec 2023

MANUS: Markerless Hand-Object Grasp Capture using Articulated 3D Gaussians

Chandradeep Pokhariya[1] * Ishaan N Shah[1] ** Angela Xing[2] ** Zekun Li[2]
Kefan Chen[2] Avinash Sharma[1] Srinath Sridhar[2]
[1]IIIT Hyderabad, [2]Brown University
ivl.cs.brown.edu/research/manus

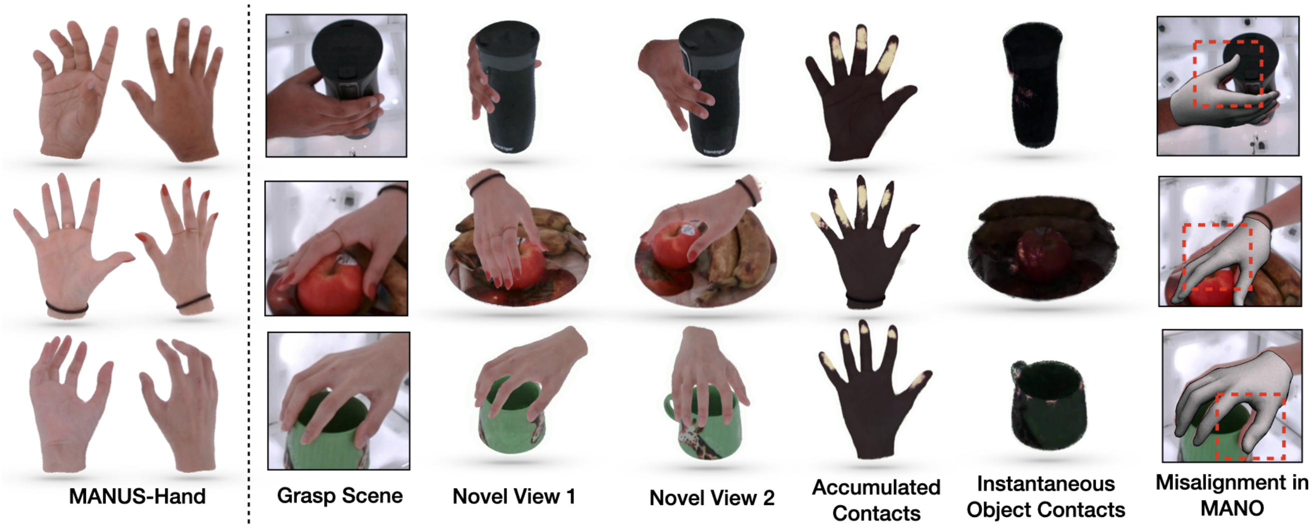

Figure 1. We present **MANUS**, a markerless grasp capture method, utilizing articulated 3D Gaussian neural representation to model the personalized hand model instead of using the parameter shape model. With more accurate geometry details, our method can access similar contact results with real-world capture compared with other template-based hand approaches.

Abstract

Understanding how we grasp objects with our hands has important applications in areas like robotics and mixed reality. However, this challenging problem requires accurate modeling of the contact between hands and objects. To capture grasps, existing methods use skeletons, meshes, or parametric models that can cause misalignments resulting in inaccurate contacts. We present MANUS, a method for ***Ma****rkerless Ha****n****d-Object Grasp Capture* ***u****sing Articulated 3D Gau****s****sians. We build a novel articulated 3D Gaussians representation that extends 3D Gaussian splatting [33] for high-fidelity representation of articulating hands. Since our representation uses Gaussian primitives, it enables us to efficiently and accurately estimate contacts between the hand and the object. For the most accurate results, our method requires tens of camera views that current datasets do not provide. We therefore build MANUS-Grasps, a new dataset that contains hand-object grasps viewed from 53 cameras across 30+ scenes, 3 subjects, and comprising over 7M frames. In addition to extensive qualitative results, we also show that our method outperforms others on a quantitative contact evaluation method that uses paint transfer from the object to the hand.*

1. Introduction

Each day, the average person effortlessly grasps more than a hundred different objects [80, 82]. This seemingly routine act of grasping poses a significant challenge for machines, as is evident from the extensive research on this topic in computer vision [18] and robotics [4, 5]. High-fidelity capture of natural human grasps could unlock new applications in areas like robotics and mixed reality, but this challenging problem first requires us to accurately **estimate the contact**

** Equal Contribution
* Work was done while at Brown University

between the hand and the object [6].

Previous work has addressed this problem by using gloves or special sensors [23, 54], but these devices are cumbersome and restrict hand movement. Therefore, a large body of work has focused on **markerless grasp capture** using one or more cameras [2, 7, 11, 24, 65].

Most of these methods use skeletons [24], meshes [2], or parametric models [30, 58] to model the hand and object. Although these representations are flexible and easy to use, they often cannot accurately model hand shape resulting in reduced contact accuracy (see Figure 1). Recently, articulated neural implicit representations [16, 45, 50] have been proposed as alternatives, but modeling contact in implicit representations is challenging and requires expensive sampling.

To overcome these limitations, we introduce **MANUS**, a method for **Ma**rkerless Ha**n**d-Object Grasp Capture **u**sing Articulated 3D Gau**s**sians. The key component of MANUS is a 3D Gaussian splatting [33] approach to build **MANUS-Hand**, an articulated hand model composed of 3D Gaussians that make it faster to train and infer than many implicitly-represented models. Similarly, we also capture the object using static 3D Gaussians. Since both MANUS-Hand and the object are modeled using Gaussians primitives with explicit positions and orientations, we can efficiently compute both *instantaneous* and *accumulated* contacts between them (see Section 4.2). When trained on datasets with tens of camera views, our method can accurately capture grasps since 3D Gaussians promote accurate pixel-level alignment resulting in more precise shape and contact estimation compared to existing methods.

Previous datasets [6, 20, 24, 25, 27, 41, 66, 81] have been instrumental in addressing the grasp capture problem but (1) they use specialized hardware (heat-sensitive cameras [6], or markers [66]) to capture hand-object grasps, making it hard to scale, (2) RGB camera-only datasets [7, 11, 20, 36], contain only a few views with occlusions making it hard to learn accurate contacts, and (3) they rely on the parametric models or skeletons to estimate contacts resulting in inaccurate contacts. **Our main insight is that accurate contact modeling is much easier with a large number of camera views that reduce the effect of (self-)occlusions.** Therefore, we curated a one-of-a-kind kind real-world multi-view RGB dataset, **MANUS-Grasps**, comprising over **$\tilde{7}$M frames** captured using 53 high-framerate cameras, providing a full 360-degree coverage and encompassing 400+ grasp sequences occurring in over 30 diverse everyday scenarios. In addition, this dataset contains 15 evaluation sequences that employ wet paint on objects to leave a contact residue on the hand [31] providing a natural way to evaluate contact quality without additional equipment or annotation. We show extensive experiments ablating and justifying different components of MANUS-Hand, as well as the MANUS grasping method. In addition, we also provide a new metric of contact quality to assess the performance of MANUS against template-based methods. While our method is not designed for photorealism, we observe that the captured grasping sequences are comparable in visual quality to the best implicit hand models.

To summarize, our contributions include:

- **MANUS-Hand**, a new efficient representation for articulated hands that uses 3D Gaussian splatting for accurate shape and appearance representation.
- **MANUS**, a method that uses MANUS-Hand and a 3D Gaussian representation of the object to accurately model contacts.
- **MANUS-Grasps**, a large real-world multi-view RGB grasp dataset with over 7M frames from 53 cameras, providing full 360-degree coverage of 400+ grasps in over 30 diverse everyday life scenarios.
- A unique and novel approach to validate contact accuracy using **paint transfer** between the object and the hand.

2. Related Work

Representations: Skeletons and collections of shape primitives were some of the first representations to be used for hand–object interaction modeling [54, 65], but these representations are often not accurate enough for contact estimation. Meshes [2] and parametric models [30, 58] are currently the most popular alternatives but can also be misaligned with observations due to their lower-dimensional representation (see Figure 1).

Coordinate-based implicit neural networks, or neural fields [74], have shown great promise in accurately modeling shape and appearance in static scenes [12, 14, 33, 42, 44, 45, 49, 51, 63, 70, 76, 78] as well as dynamic scenes [22, 38, 43, 69, 75, 77]. Several methods specifically address articulated shapes [37] like human bodies [37, 40, 52, 53, 72], or hands [16, 32, 39, 50, 55]. However, they use representations that are inefficient for sampling and contact estimation. In contrast, we propose a new articulated neural field representation that extends 3D Gaussian splatting [33] to hands enabling efficient training/inference and contact estimation.

Hand-Object Interaction Capture: Previous work has attempted to model hand-object interactions using skeletons [24, 36], or customized meshes [2] as the hand representation without explicitly estimating contacts. Most other work [11, 20, 27, 41, 66] uses MANO in combination with mocap, or one or more camera views. While it becomes easier to estimate contact with a parametric mesh model, misalignments are still common (see Figure 1). To overcome the difficulty of accurate contact estimation, some methods resort to physical simulation [15, 68, 79], but these are limited to synthetic grasps only. In contrast, we propose a template-free articulated 3D Gaussian splatting model that provides a natural way to estimate accurate contacts.

Grasp Datasets: Datasets for human grasps are challenging to obtain because they need specialized hardware, extensive annotation, and significant post-processing to make them useful. Some datasets use markers or special gloves to track the hand and object [3, 17, 23, 67] but this hinders natural hand motion and introduces changes in image appearance. Synthetic datasets [27, 47, 48] suffer from a domain gap that makes it challenging to generalize to real data. Therefore, work has focused on manual annotations [2, 8, 57, 65], optimization [24], or automatic annotation [11, 62] from RGB or depth. Many of these datasets provide only 3D hand poses and lack information about contacts. Other datasets like InterHand2.6M [46, 81] are limited to hands only without any objects, while others [61] focus on 2D understanding only. Addressing these limitations, HOnnotate [24] introduces a markerless system for automatically annotating frames across 77K frames. However, the variety of objects and grasps in this dataset is somewhat limited. ContactDB [6] and ContactPose [7] address this limitation targets a broader variety of grasps. While ContactDB is captured using thermal imaging, ContactPose uses multi-view RGB-D data. Nonetheless, both methods are restricted to 3D hand poses, use non-realistic objects, and lack sufficient views for neural fields.

In contrast, we introduce MANUS-Grasps that includes over 400+ grasps views from 53 cameras capturing at 120 FPS specifically to support neural field methods. In total, we provide over 7M frames with ground truth camera poses, segmentation, and estimated contacts.

3. Background

We briefly summarize recent advances in modeling radiance fields of static and dynamic scenes using 3D Gaussians [33, 43, 73]. Our method (see Section 4) extends the 3D Gaussians representation to articulated objects like the hand, and for grasp capture.

Static 3D Gaussians: Given multi-view images and a sparse point cloud of the scene, a set of 3D Gaussian primitives can be defined across world space $x \in \mathrm{R}^{3\times 1}$ as,

$$G(x) = e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

where each Gaussian primitive has 3D position (μ), opacity, anisotropic covariance matrix (Σ), and spherical harmonic (SH) coefficients. During the training of the radiance field, the properties of the initial 3D Gaussians are optimized together with a tile rasterizer [33] with the objective of minimizing pixel loss.

Dynamic 3D Gaussians: The 3D Gaussians approach has recently been extended to dynamic scenes [33, 73]. [73] introduces a deformation field that tracks changes in the position of the Gaussians and shape changes. Similarly, [43] enable Gaussians to move and rotate over time while maintaining their color, opacity, and size. While these methods can capture dynamic and deformable scenes, they do not provide a way to control dynamic motion, *e.g.*, using a skeleton. Furthermore, in these methods, Gaussians are free to move within the scene without any restrictions, which isn't suitable for representing hands due to their kinematic structure. An articulated 3D Gaussians representation would be advantageous for grasp capture since it would enable low-dimensional skeleton-based control of the hand.

Dataset	#N Images (Views)	Annot. Type
w/o Contacts Annotation		
H2O-3D [25]	76k (5)	multi-kinect
FHPA [23]	105k (1)	magnetic
HOI4D [41]	2.4M (1)	single-manual
FreiHand [81]	37k (8)	semi-auto
HO3D [24]	78k (1-5)	multi-kinect
DexYCB [11]	582k (8)	multi-manual
ARCTIC [20]	2.1M (9)	mocap
w/ Estimated Contacts Annotation		
ContactPose [7]	2.9M (3)	multi-kinect
GRAB [66]	- (-)	mocap
H2O [36]	571k (5)	multi-kinect
w/ Ground-Truth Contacts Annotation		
MANUS-Grasps (Ours)	7M (53)	multi-auto

Table 1. Dataset Comparison of existing Real World Datasets. The hands in previous datasets are represented by skeleton and MANO. Different from other works, we use Gaussian to model the hand. The keyword "single/multi-manual" denotes whether single or multiple views being used to annotate manually.

4. Method

MANUS aims to perform markerless capture of human hand grasps by accurately estimating the shape, appearance, and precise contact between the hand and the object from multi-view RGB videos. We achieve this by combining MANUS-Hand with an object model, both represented as 3D Gaussians, enabling us to compute contacts more efficiently than sampling-based implicit representations. Figure 2 provides an overview of the our method.

4.1. MANUS-Hand

Our template-free, articulated hand model MANUS-Hand adopts 3D Gaussian splatting as the representation for accurate shape and appearance modeling of hands. Our model can be trained on sequences from any multi-view dataset to build an articulable hand model at any novel pose.

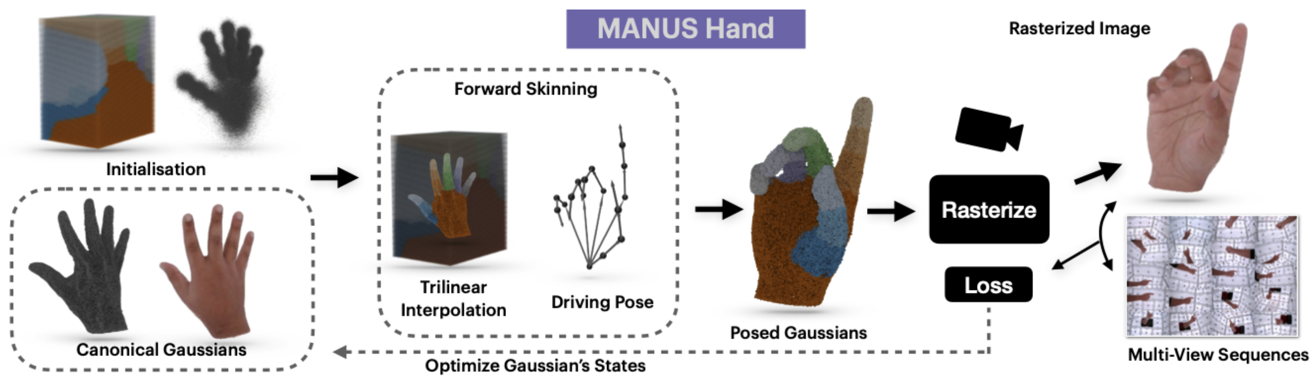

Figure 2. **MANUS-Hand** is a template-free, articulable hand model learned from multi-view hand sequences which utilizes 3D Gaussian splatting representation for accurate modelling of the shape and appearance of hands.

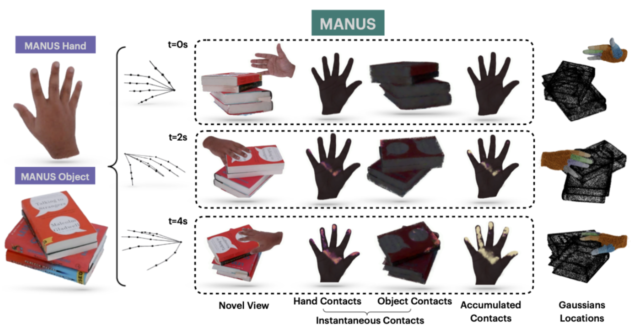

Figure 3. **MANUS** leverages MANUS-Hand to drive the hand in the grasp scene and an object to model both instantaneous and accumulated contacts between the two.

Representation: MANUS-Hand (see Figure 2) is composed of a skeleton with 21 bones and has 26 degrees of freedom. We built a custom pose estimation pipeline that uses AlphaPose [21] to estimate the 3D joint positions followed by an inverse kinematics fit (see appendix). Since bone lengths can vary among different individuals, we estimate these lengths from the dataset and adjust the skeleton accordingly. The unique shape and appearance of a person's hand in a canonical pose are determined by the states of 3D Gaussians, *i.e.*, positions μ, covariances Σ, opacities α, and spherical harmonics coefficients ϕ. The covariance of each Gaussian in the canonical space is further defined as $\Sigma = RSS^TR$, where R and S denote the rotation and scaling of the Gaussians.

Optimization: A unique MANUS-Hand is optimized separately for each hand from a dense multi-view dataset containing approx 20 hand poses. To initialize Gaussian states in MANUS-Hand, we set their means to be points on a normal distribution centered at the midpoint of each bone in a *canonical* hand pose, with the distribution's standard deviation adjusted to match the bone's length. We follow a similar protocol as [33] to initialize the covariances, opacity, and SH coefficients.

To get the Gaussian positions in the posed space, forward kinematics and linear blend skinning is applied to the canonical Gaussians. One way to obtain skinning weights is to assign MANO weights [58] directly to the closest Gaussians. However, this approach results in artifacts because Gaussians could move in unpredictable ways during training leading to mismatched skinning weights (refer to supp.) To address this, we create a canonical grid inspired by Fast-SNARF [13]. Skinning weights are then allocated to grid voxels using the nearest neighbor method, termed as grid weights. Now to obtain the skinning weights for the queried Gaussians W in the canonical space, trilinear interpolation of these grid weights is performed. We calculate the transformed Gaussian positions using a per-bone transformation matrix, denoted as T_b and linear blend skinning: $T_g = WT_b$, $\mu_p = T_g\mu$, where μ_p represents the location of Gaussians in the posed space, and T_g represents the transformation matrix for each Gaussian. To compute the covariance of the Gaussians in the posed space, it is transformed using a rotation matrix R_g, derived from from T_g. This is expressed as $\Sigma_p = R_g \Sigma R_g^T$. Regarding the appearance, we optimize spherical harmonics coefficients for each Gaussian ϕ_g in the canonical space. To derive the colors in the transformed or posed space, the view direction from posed space ν_p^g is first converted to the canonical space ν_c^g as $\nu_c^g = T_g^{-1}\nu_p^g$, using T_g for each Gaussian. After this step, we use these transformed view directions μ_c^g to query the spherical harmonics coefficients in canonical space and get corresponding RGB colors for each posed Gaussian. To get the final image rendering, all Gaussian states currently in the posed space are used as inputs to a differentiable rasterizer [33], denoted as $\mathcal{R}$

$$\mathcal{I} = \mathcal{R}(\mu_p, \nu_c, \Sigma_p, \alpha, \phi), \tag{1}$$

where $\mathcal{I}$ is the rendered image. During optimization, the Gaussian states are optimized using to minimize pixel loss on the posed hand. To optimize all Gaussian states, we impose a rendering loss $\mathcal{L}_1 = \|\hat{\mathcal{I}} - \mathcal{I}\|$ and structural similarity [71] loss $\mathcal{L}_{SSIM}$ between synthesized image $\mathcal{I}$ and ground truth image $\hat{\mathcal{I}}$ of the posed hand. To further improve the perceptual quality of the synthesized images, we add an additional perceptual loss $\mathcal{L}_{perc}$ [29].

To avoid highly anisotropic Gaussians that could cause geometric artifacts in the contact estimation, we incorporate an isotropic regularizer which ensures optimized Gaussians remain as isotropic as possible. If $\min_s \in R^3$ and $\max_s \in R^3$ are the minimum and maximum scale of the optimized Gaussians, then isotropic regularizer $\mathcal{L}_{iso}$ is defined as

$$\mathcal{L}_{iso} = (\frac{\min_s}{\max_s} - s)^2, \tag{2}$$

where s is set to be 0.4. Our final loss function is $L_h = \alpha\mathcal{L}_1 + \beta\mathcal{L}_{SSIM} + \gamma\mathcal{L}_{perc} + \delta\mathcal{L}_{iso}$.

Inference: Once the Gaussian states are optimized in the training phase, we can drive MANUS-Hand using a skeleton estimated using our pose estimation pipeline. Given a novel pose during the inference, MANUS-Hand outputs the transformed Gaussians as well as the rendered image from a particular view.

4.2. MANUS: Grasp Capture

While MANUS-Hand enables high-fidelity articulated hand modeling, it is not designed for capturing grasps and contacts. To capture grasps, we need a representation of the object as well as a method to estimate contacts.

Object Representation: For accurate representation of objects, we build a non-articulated Gaussian representation following Section 4.1 with some improvements to maintain geometric consistency and accuracy. To prevent floaters commonly found in implicit represntations, we prune outlier Gaussians during training by projecting them to the image and culling if they lie outside the object mask.

Grasp Capture: To capture the grasp in a particular sequence, we first articulate MANUS-Hand using the estimated hand pose. We then construct the object model as described above. Next, we combine both hand and object Gaussians. More specifically, if G_h and G_o are the hand Gaussians and object Gaussians in the grasp scene, we simply concatenate the Gaussians $G_f = \{G_o, G_h\}$. Because we use Gaussian Splatting, it allows such a concatenation operation naturally – this would not be possible with implicit representations [16, 37, 50]. As the rasterization module only requires a set of Gaussians and their states, we can seamlessly merge hand and object Gaussians for every frame. The final grasp image is given by a rasterized composition of these Gaussians using Equation (1).

Contact Estimation: The contact map is calculated based on the proximity in 3D space between hand and object Gaussian positions. For each Gaussian on the hand, we find the closest Gaussian on the object. This pair is considered to be in contact if their distance is less than a certain threshold, and the same applies when assessing contact from the object's perspective. Specifically, if G_h represents the Gaussians on the hand and G_o those on the object in the posed space, then the 3D contact map between them is defined as:

$$C = \begin{cases} d(G_h, G_o), & \text{if } d(G_h, G_o) < \tau \\ 0, & \text{otherwise} \end{cases},$$

where d represents the pairwise Euclidean distance between the Gaussian locations. A contact is considered to have occurred if this distance is less than τ (τ = 0.004 in our experiments), which is the predefined threshold for contact. We then use this method to estimate two kinds of contact maps on the hand and object: (1) an **instantaneous contact map** that denotes contact at a specific timestep, and (2) an **accumulated contact map** that denotes contact after the grasping has concluded. To get the accumulated contact map C_{acc} we simply add the previous frame's accumulated contact map to current frame. For rendering contact maps, we employ Equation (1) using the contacting distance as the color value of each Gaussian.

4.3. MANUS-Grasps

For our grasp capture method to work well, a key requirement is a multi-view RGB dataset with tens of camera views that help resolve (self-)occlusions. Many prior datasets (see Section 2) contain multi-view images or video of hand grasps [24, 62, 67], but none have the large number of views needed to support neural field representations or are limited to hands only [46]. We therefore present MANUS-Grasps, a large real-world multi-view RGB grasp dataset with over 7M frames from 53 cameras, providing full 360-degree coverage of 400+ grasps in over 30 diverse everyday scenarios.

Capture System: Our customized data capture setup consists of 53 RGB cameras uniformly located inside a cubical capture volume with each cube face consisting of 9 cameras. The sides of the cube are illuminated evenly using LED lights. Each RGB camera records at 120 FPS with a resolution of 1280 × 720. The cameras are software synchronized with a frame misalignment error of no more than 3 ms. The multi-view system is calibrated for camera intrinsics and extrinsics using COLMAP [59, 60] with fiducial markers on the walls.

Capture Protocol: Our capture protocol consists of four steps. First, we recorded multi-view videos of a subject's right hand as they performed a brief articulating movement. Next, we capture only the object without the hand. Then, without moving the object, we record multi-view videos of the subject's hand grasping the object. We repeat this process 30 times per subject with 2-5 grasps per object. For evaluation sequences, we additionally capture a canonical pose at the end to record accumulated contacts seen in the transferred paint (see below).

Ground Truth Contact: A unique feature of our dataset is the capture of 15 sequences where the object has wet paint during the grasp [31]. As a result, paint is transferred to the hand resulting in visual evidence of contact. This contact mark is a physically accurate representation of the true (accumulated) contact between the hand and the object making it the true ground truth (even methods like [6] suffer from heat dissipation). We chose a bright green paint to enable automatic segmentation thereby creating a **gold standard** for contact evaluation.

Data Annotation: MANUS-Grasps also provides 2D and 3D hand joint locations along with hand and object segmentation masks. We obtain the joint locations from AlphaPose [21] followed by 3D triangulation and inverse kinematics [64]. We impose constraints to limit the degrees of freedom and joint angles for the rotation of the bones. To achieve temporal smoothness for the sequence, we apply the 1€ Filter [9] on the estimated parameters. To segment the hand and object from the background, we use the Segment Anything Model (SAM) [35] followed by fitting an Instant-NGP model [49] to extract a binary mask to ensure multi-view consistency.

5. Experiments and Results

In this section, we show qualitative and quantitative results from our method. Our goal is to evaluate both the MANUS-Hand and the MANUS grasp capture method, compare with existing methods, and ablate key design choices.

5.1. Evaluating MANUS-Hand

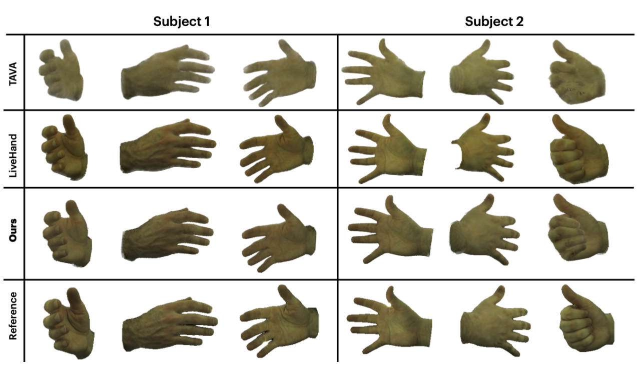

Figure 4. Qualitative comparison of MANUS-Hand with LiveHand [50] and TAVA [37]. It's noteworthy that our renderings closely resemble those of LiveHand and surpass TAVA in quality, even in the absence of any components designed to enhance photorealism.

We first show results and experiments related to MANUS-Hand only. We quantitatively as well qualitatively assess the visual quality of our hand model with the current state-of-the-art method LiveHand [50] and TAVA [37].

Metrics, Dataset & Setup: We assess the visual quality of our hand model using PSNR, SSIM, and LPIPS metrics (where higher scores indicate better performance) on the Interhand2.6M dataset, as shown in Table 3. While these metrics focus on visual quality, we note that they are proxies for geometric quality as well, which is important for grasp capture. For optimization, we used two subjects from Interhand2.6M (Capture0 and Capture1), focusing on the "ROM07-RT-Finger-Occlusions" sequence from the test set. We allocate 75% of the data for training and use the remainder for evaluation.

Quantitative Evaluation: MANUS-Hand is not specifically designed for photorealism since we leave out ambient occlusion and shadow mapping and focus only on geometric accuracy. Despite this, our method is superior to TAVA and on par with LiveHand. Furthermore, we evaluated our method on our MANUS-Grasps dataset , as detailed in Table 4, using a similar training and testing split of 3:1. Our dataset is better lit than InternHand2.6M with fewer harsh shadows resulting in significantly better performance of our method. In conclusion, despite not being tailored for photorealism, our method demonstrates substantial potential for application in photorealistic contexts.

Qualitative Evaluation: We conducted a qualitative comparison of our MANUS-Hand with TAVA [37] and LiveHand [50], as shown in Fig. 4. The quality of our renderings is superior to TAVA [37] and is on par with that of LiveHand.

5.2. Evaluating Grasp Capture

Next, we evaluate our MANUS method for grasp capture. In this paper, we assume that direct contact between the hand and the object is the primary mode of grasping (we ignore indirect grasping through tools). Therefore, the goal of grasp evaluation is to objectively measure the accuracy of contacts. We compare three methods: (1) a method that uses MANO [58] to model hand grasps, (2) HARP [32], and (3) our MANUS model. Please see the appendix for implementation details.

Metric, Dataset & Setup: In our experiments, we use the paint transfer method [31] to accurately record ground truth accumulated contacts (see Section 4.3). After grasp completion, users are instructed to return to a canonical post-grasp pose. In this pose, the green paint residue in the grasping hand is automatically segmented and 2D contact maps are rendered from 10 different views using [49]. We then assess the quality of grasps estimated by different methods using the Intersection over Union (IoU) and F1-score metrics. All experiments use 15 sequences of our wet paint dataset. For fairness, all methods used the same distance threshold $\tau = 0.004$ for contact estimation. For a fair comparison, we subdivide the meshes of MANO and HARP from 778 to 49,000 vertices before estimating contact.

For generating contact masks in all methods, we utilize the 'gray' color map [28] on the distance map. The contact masks for MANUS are rendered using [33], while for the other two frameworks, they are rendered using the emission shader in Blender. It's noteworthy that MANUS **consistently outperforms** the others in the contact metric across all three subjects.

Method	Subject1	Subject2	Subject3
mIoU ↑			
MANO	0.161	0.135	0.208
HARP	0.173	0.148	0.224
Ours	**0.206**	**0.152**	**0.275**
F1 score ↑			
MANO	0.270	0.228	0.338
HARP	0.28875	0.2474	0.361
Ours	**0.335**	**0.251**	**0.424**

Table 2. Comparison of MANUS grasp capture approach with MANO and HARP on contact metric. Note that, we perform consistently better in terms of both metrics.

Qualitative Evaluation: Figure 6, we present a qualitative comparison of our contact results against those obtained using MANO and HARP. Our method shows a more accurate representation of the contact area, closely matching the actual contact masks, unlike the over-segmentation observed in MANO and HARP methods. Although our method outperforms others, we note that there is still significant room for improvement on our dataset for future methods to address. We also show qualitative results in in Fig. 5 and benchmark our MANUS-Hand and object model on our dataset in Tab. 4.

5.3. Ablation Study

In Table 5 it is demonstrated how the quality of contacts diminishes as the number of camera views decreases. This finding is significant as it confirms our hypothesis that dense camera views are essential for accurate contact representation, helping to prevent occlusion scenarios. Please see the appendix for other ablations.

6. Conclusion

In conclusion, our method, MANUS, introduced a novel articulated 3D Gaussians representation, which successfully bridge the gap between the accurate modeling of contacts in hand-object interactions and the limitations of current data capturing techniques. The creation of MANUS-Grasps, with its extensive multi-view data from 53 cameras, offers an unprecedented level of detail and accuracy, covering a wide range of scenes, subjects, and frames. Overall, MANUS demonstrates remarkable potential for advancing the fields of robotics, mixed reality, and activity recognition, offering new possibilities for the development of more agile and accurate robotic systems and enhanced virtual interaction experiences.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Test time (s) ↓
TAVA	22.85	0.983	0.099	11.00
LiveHand	**31.16**	0.9818	**0.0278**	**0.022**
Ours	26.32	**0.9872**	0.068	0.049

Table 3. Comparison of MANUS-Hand on InterHand2.6M [46] dataset with LiveHand [50].

Categories	PSNR ↑	SSIM ↑	LPIPS ↓
Mugs	43.08	0.999	0.002
Bottles	38.17	0.997	0.008
Fruits	39.57	0.998	0.005
Utensils	38.25	0.994	0.009
Misc	38.79	0.995	0.008
Colored	42.38	0.999	0.004
Bags	38.44	0.994	0.011
Jars	40.66	0.999	0.005
Books	36.17	0.998	0.015
Tech	38.81	0.995	0.007
Hand1	28.34	0.995	0.031
Hand2	29.94	0.998	0.029
Hand3	29.71	0.997	0.027

Table 4. Benchmarking MANUS-Hand and MANUS-Object on MANUS Grasp dataset.

Camera Views	Subject1	Subject2	Subject3
mIoU ↑			
5	0.147	0.140	0.214
10	0.164	0.145	0.256
20	0.176	0.142	0.261
Ours (53)	**0.206**	**0.152**	**0.275**
F1 score ↑			
5	0.244	0.235	0.343
10	0.266	0.242	0.401
20	0.271	0.240	0.410
Ours (53)	**0.335**	**0.251**	**0.424**

Table 5. Ablation of how contact metric degrades as camera views are reduced.

Limitations: While our focus in this paper was on accurate contact estimation, we did not focus on multiple hands or non-rigid objects. We also observe that there is room for improvement in the metrics we propose for future work.

Acknowledgements: This work was supported by NSF CAREER grant #2143576, a gift from Meta Reality Labs, NSF CloudBank, and an AWS Cloud Credits award.

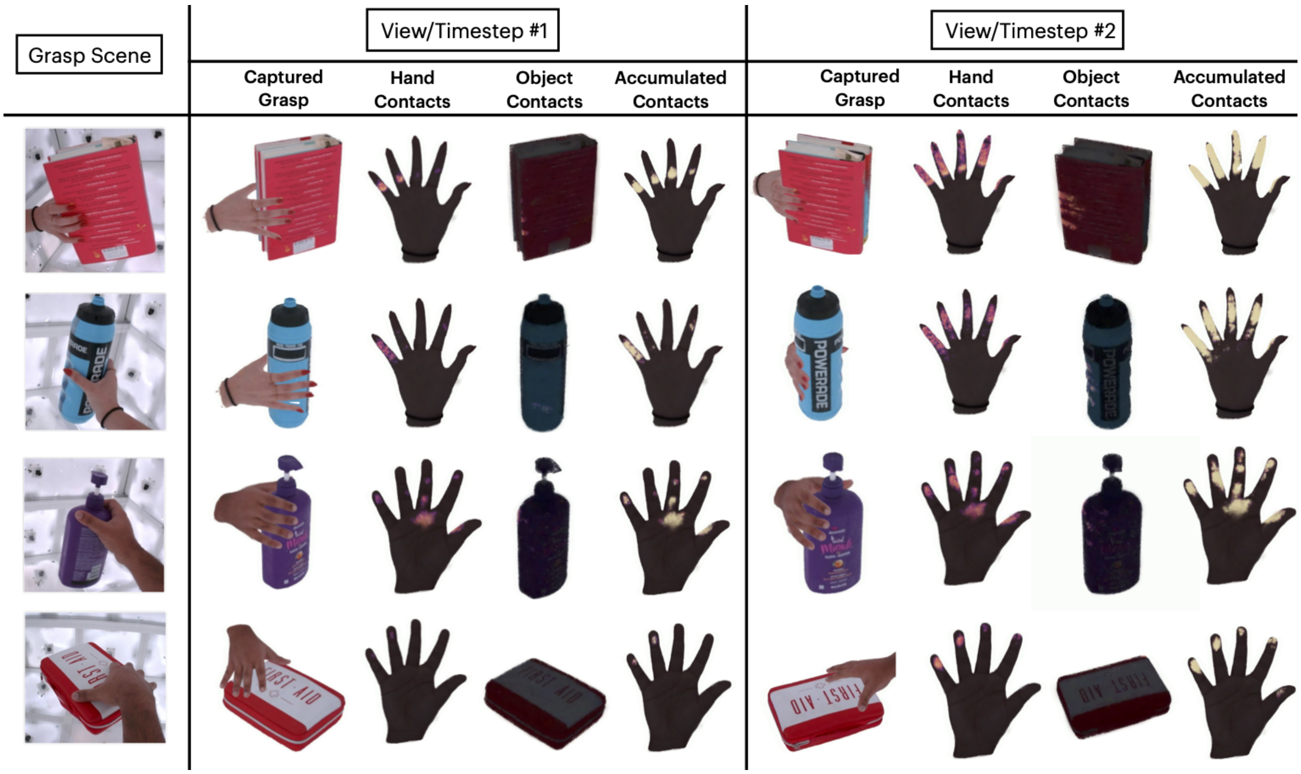

Figure 5. Here we show our contact estimation results on novel views for a variety of objects. We show both instantaneous and accumulated contacts for the hand in a canonical pose. Best viewed zoomed.

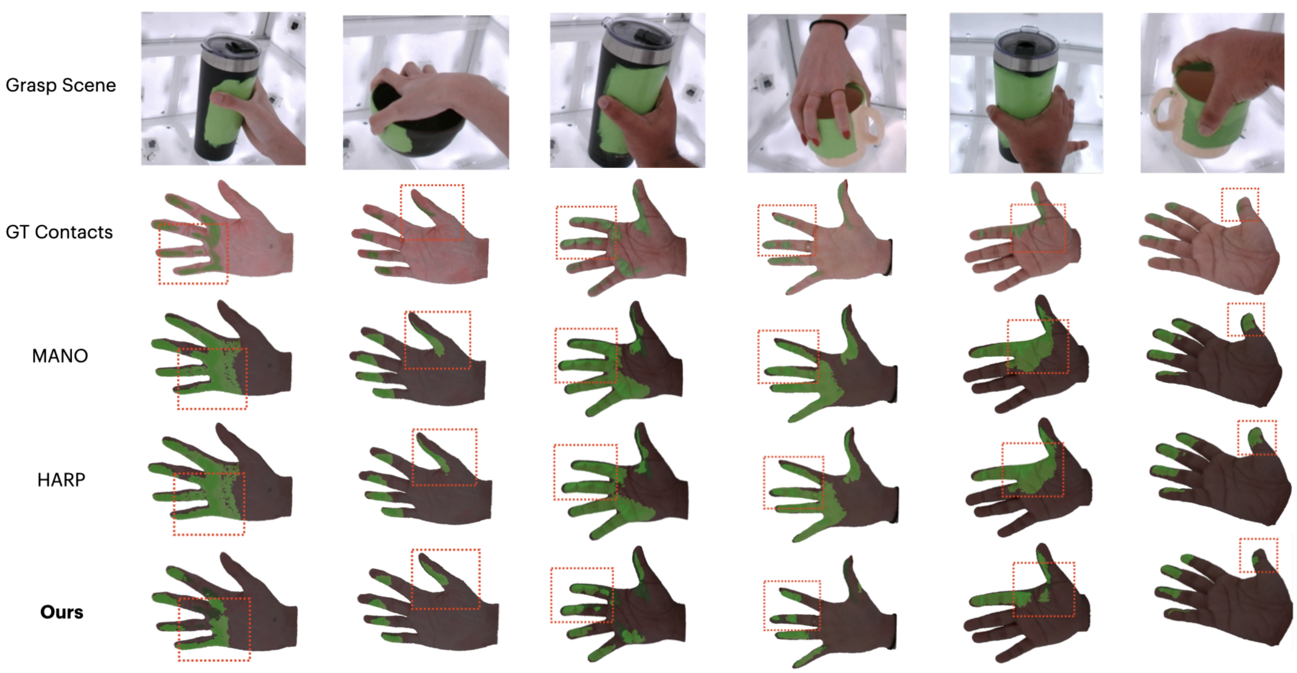

Figure 6. **Contact Comparisons**: We compare accumulated contacts of MANUS with that of MANO and HARP on ground truth contacts from MANUS Grasps dataset. It's visible that our contacts are far more accurate and closer to the actual ground truths.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 12

[2] 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. 2, 3

[3] Keni Bernardin, Koichi Ogawara, Katsushi Ikeuchi, and Ruediger Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *IEEE Transactions on Robotics*, 21(1):47–57, 2005. 3

[4] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, pages 348–353. IEEE, 2000. 1

[5] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics*, 30(2):289–309, 2013. 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. 2, 3, 6

[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 *European Conference on Computer Vision*, pages 361–378. Springer, 2020. 2, 3

[8] Ian M Bullock, Thomas Feix, and Aaron M Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *The International Journal of Robotics Research*, 34(3):251–255, 2015. 3

[9] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012. 6, 13

[10] Umberto Castiello. The neuroscience of grasping. *Nature Reviews Neuroscience*, 6:818–818, 2005. 14

[11] 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, 3

[12] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 2

[13] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[14] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[15] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[16] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *CVPR*, 2022. 2, 5

[17] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionnet: A multimodal dataset for human activities using wearable sensors in a kitchen environment. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[18] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108 (1-2):52–73, 2007. 1

[19] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. 12

[20] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[21] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4, 6, 12, 13

[22] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 2

[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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2, 3

[24] 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. 2, 3, 5

[25] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *IEEE Computer Vision and Pattern Recognition Conference*, 2022. 2, 3

[26] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learn-

ing joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 13

[27] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR 2019 - IEEE Conference on Computer Vision and Pattern Recognition*, pages 11799–11808, Long Beach, United States, 2019. IEEE. 2, 3

[28] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. 7

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[30] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[31] 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. 2, 6

[32] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12802–12813, 2022. 2, 6, 13, 14

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 1, 2, 3, 4, 7

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 13

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 6, 13

[36] 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. 2, 3

[37] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 2, 5, 6

[38] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *CoRR*, abs/2103.02597, 2021. 2

[39] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2

[40] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2

[41] 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, 3

[42] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, 2019. 2

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

[44] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[45] 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. 2

[46] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020. 3, 5, 7

[47] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 3

[48] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[49] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 6, 13

[50] Akshay Mundra, Jiayi Wang, Marc Habermann, Christian Theobalt, Mohamed Elgharib, et al. Livehand: Real-time and photorealistic neural hand rendering. *arXiv preprint arXiv:2302.07672*, 2023. 2, 5, 6, 7

[51] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[52] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2

[53] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[54] 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

[55] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 54–71. Springer, 2020. 2

[56] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 13

[57] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 3889–3897, 2015. 3

[58] 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. 2, 4, 6, 12

[59] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[60] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[61] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 3

[62] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 3, 5

[63] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[64] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 6, 13

[65] 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 *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 2, 3

[66] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[67] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 3, 5

[68] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *ECCV*, 2022. 2

[69] 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. 2

[70] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2021. 2

[71] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

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

[73] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 3

[74] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. 2

[75] Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8285–8295, 2023. 2

[76] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[77] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 2

[78] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenox-

els: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 2

[79] 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. *arXiv preprint arXiv:2309.03891*, 2023. 2

[80] Joshua Z Zheng, Sara De La Rosa, and Aaron M Dollar. An investigation of grasp type and frequency in daily household and machine shop tasks. In *2011 IEEE international conference on robotics and automation*, pages 4169–4175. IEEE, 2011. 1

[81] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2, 3

[82] Paula Zuccotti. *Every Thing We Touch: A 24-hour Inventory of Our Lives*. Penguin UK, 2015. 1

A. Ablation Study

A.1. MANUS-Hand

Initialization of Skinning Weights: We observe that the choice of method used to initialize skinning weights significantly influences the performance of our hand model. As demonstrated in Fig. 8(a), initializing skinning weights directly onto Gaussians using a nearest neighbor approach, as opposed to using a voxel grid, leads to an issue. This approach causes the unstructured Gaussians to shift towards an unrelated bone. Consequently, this misalignment results in artifacts, where skinning weights are incorrectly allocated to the wrong bone, causing the position to be associated with the incorrect bone. The impact of this method of initialization is presented both quantitatively and qualitatively in Tab. 6 and Fig. 7.

Ablation on LPIPS loss: We observed that LPIPS loss helps a lot in realistic renderings and maintaining consistency across views. Figure 7 and quantitatively at Table 6, we demonstrate that LPIPS metric enhances the overall visual quality of our hand model.

A.2. MANUS Grasp Capture

Effect of the number of Gaussians in contact map rendering: We show in Figure Fig. 8(b) that the quality of accumulated 2D contact maps deteriorates when the number of Gaussians is reduced. Therefore, in our experiments, we make sure to densely initialize Gaussians for both objects and hands.

B. Implementation Details

Our method was implemented in Python using the PyTorch Lightning [19] framework. All experiments were conducted

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Test time (s) ↓
w/o grid	26.108	0.987	0.0729	**0.0082**
w/o lpips	25.92	0.986	0.074	0.043
Ours	**26.328**	**0.9872**	**0.0688**	0.043

Table 6. Ablation on weight initialization approach and choice of LPIPS loss. Our design approach improve all visual quality metrics.

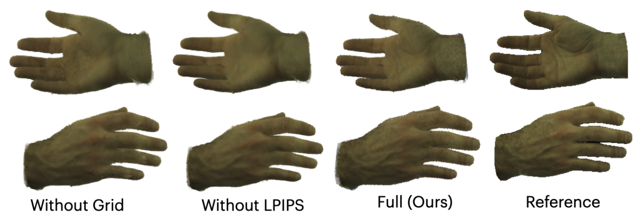

Figure 7. **Hand Ablation**: We perform ablation on the initialization of the skinning weights in the grid as well as the choice of using LPIPS loss function. Clearly our approach is better in terms of visual appearance.

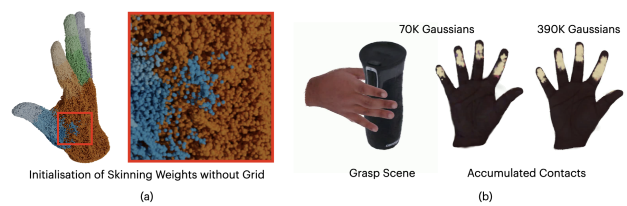

Figure 8. Here in **(a)** we show how initializing MANO weights without voxel grid allows the unstructured Gaussians to move erratically. In **(b)**, we show the affect on accumualated 2D contact renderings with change in the number of Gaussians.

using a single Nvidia RTX3090 GPU with gradient accumulation for 4 iterations. The weights of the different loss function terms - α, β, γ and δ - were experimentally determined and set at values of 0.7, 0.1, 0.1, and 0.1, respectively. In all our experiments, we chose a grid size of 256x160x142 around the canonical hand skeleton for storing the skinning weights initialized from MANO [58]. MANUS-Hand is initialized with 30K Gaussians per bone, amounting to 900K Gaussians in total. After training, this number is pruned and filtered down to approximately 300K.

C. Data capture and Pre-processing of MANUS Grasps

C.1. Bone length estimation

We first use the [21] to acquire 2D keypoints for every frame and view. These keypoints are then triangulated into 3D keypoints using the [1]. With these triangulated keypoints, we determine the bone lengths for each subject. Specifically, we average the 3D keypoints across all grasp sequences and then adjust the length of the skeleton accordingly.

C.2. Inverse Kinematics

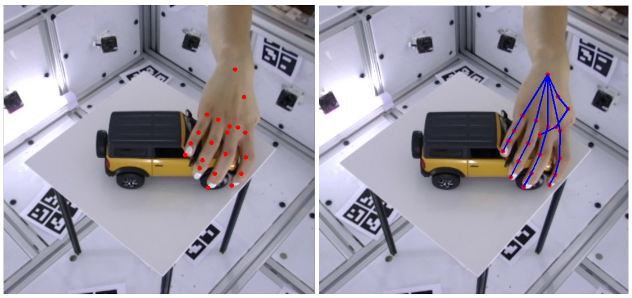

Figure 9. The left figure shows the backprojected 3D keypoints predicted by AlphaPose [21]. The right figure shows the fitted hand skeleton using inverse kinematics.

To obtain the joint angles of the hand and its global orientation we use an optimization-based approach inspired by [64]. Specifically, we treat the joint angles, global rotation and global translation as optimization parameters Θ. We then perform a forward kinematics ($Fk(\Theta)$) pass which takes the joint angles as input and outputs 3D joint locations. As the forward pass is differentiable, we apply gradient descent to obtain the optimal parameters that explain the given 3D joint positions. We minimize the L2 loss between predicted and target keypoints:

$$\mathcal{L}_{kyp} = ||Fk(\Theta) - x||^2 \quad (3)$$

where x are the 3D joint locations predicted by AlphaPose [21]. We also impose anatomical constraints (See Figure 10) and joint angle limits by applying a hinge loss as limit loss $\mathcal{L}_{lim}$ as follows:

$$\mathcal{L}_{lim} = \sum_{i=1}^{|\Theta|}((\max(0, ||\Theta^i - l_h^i||^2) + \max(0, ||l_l^i - \Theta^i||^2)) \quad (4)$$

where l_l and l_h are the lower and upper limits on joint angles, respectively. The final loss function is given by:

$$\mathcal{L} = \mathcal{L}_{kyp} + \lambda\mathcal{L}_{lim} \quad (5)$$

We use Adam [34] as our choice of optimizer with a learning of 0.001. We set the value of λ to be 1 in all our experiments. We also initialize the current frame based upon previous frame, this helps in faster convergence and helps in maintaining temporal consistency. Once we get the joint angles, we apply one euro filter [9] to the joint angles to smoothen any high-frequency jitter in the sequence. We show illustration of this process in Fig. 9.

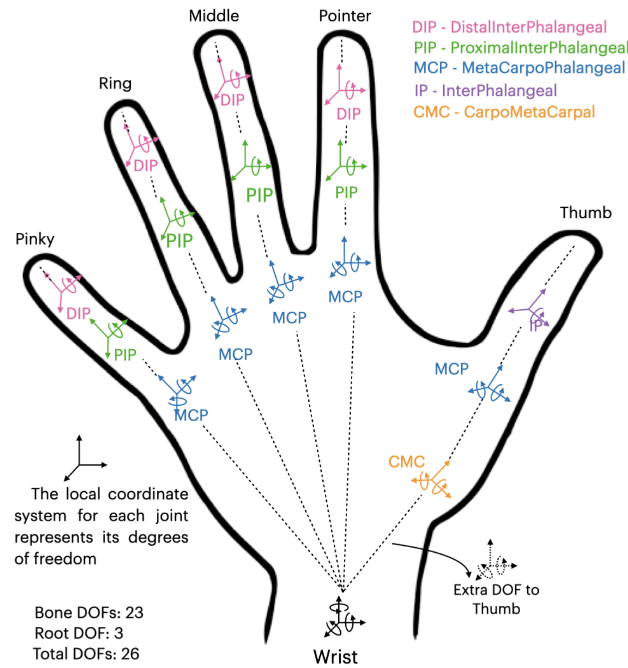

Figure 10. Figure showing the degrees of freedom of rotation for each of the joint.

C.3. Segmentation

For every segmentation task, we employ a combined approach utilizing InstantNGP [49] and SAM [35]. Initially, the scene is segmented using the text-based SAM technique. Following this, we obtain a segmentation mask that maintains consistency across multiple views using InstantNGP. If the segmentation masks are found to be inadequate due to inaccurate predictions from the text-based SAM, the process is repeated until satisfactory results are achieved.

D. MANO and HARP fitting

We begin by estimating the shape and scale parameters of the MANO model for each subject. This process starts with acquiring the mesh from [49], which we process using MeshLab and Blender software. Subsequently, we employ an optimization framework akin to that used in [26], focusing on optimizing all MANO parameters, including angle, translation, shape, and scale. This optimization incorporates both keypoint loss (3) and point-to-surface loss [56]. For subsequent sequences, we maintain the previously determined scale and shape parameters constant, while optimizing only the angles and translations, using only keypoint loss. We initialize the previously optimized parameters as our new parameters to improve the convergence rate.

To attain a more precise hand geometry based on multiview images, we follow HARP [32] using local displacement to model the personalized hand shape. Leveraging the differentiable render techniques, we optimize the MANO mesh based on the losses in [32].

E. MANUS Dataset Details

Dataset Release: We plan to release complete dataset in the future including the estimated contacts using MANUS Grasp Capture framework.

Grip Aperture: The grip aperture [10] refers to the distance between the thumb and fingers when grasping or holding an object. It's an important concept in fields like ergonomics, rehabilitation, and robotics. Here in Fig. 11, we plot the change of grip aperture with change in timestep for our dataset.

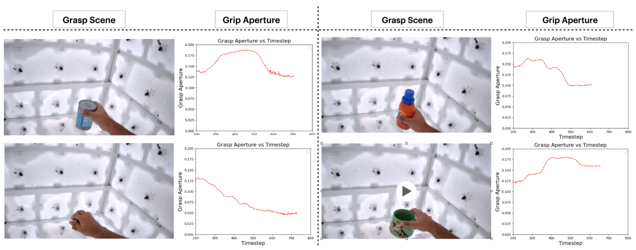

Figure 11. Variation of grip aperture with change in timestep while grasping.