

GHOST: Fast Category-agnostic Hand-Object Interaction Reconstruction from RGB Videos using Gaussian Splatting

Ahmed Tawfik Aboukhadra^{1,2}
Jameel Malik⁴

Marcel Rogge^{1,2}
Ahmed Elhayek³

Nadia Robertini²
Didier Stricker^{1,2}

Abdalla Arafa^{1,2}

¹RPTU ²DFKI-AV Kaiserslautern ³UPM Saudi Arabia ⁴NUST-SEECs Pakistan

firstname{secondname}.lastname@dfki.de

Abstract

Understanding realistic hand-object interactions from monocular RGB videos is essential for AR/VR, robotics, and embodied AI. Existing methods rely on category-specific templates or heavy computation, yet still produce physically inconsistent hand-object alignment in 3D. We introduce **GHOST** (**G**aussian **H**and-**O**bject **S**pl^Atting), a fast, category-agnostic framework for reconstructing dynamic hand-object interactions using 2D Gaussian Splatting. GHOST represents both hands and objects as dense, view-consistent Gaussian discs and introduces three key innovations: (1) a geometric-prior retrieval and consistency loss that completes occluded object regions, (2) a grasp-aware alignment that refines hand translations and object scale to ensure realistic contact, and (3) a hand-aware background loss that prevents penalizing hand-occluded object regions. GHOST achieves complete, physically consistent, and animatable reconstructions from a single RGB video while running an order of magnitude faster than prior category-agnostic methods. Extensive experiments on ARCTIC, HO3D, and in-the-wild datasets demonstrate state-of-the-art accuracy in 3D reconstruction and 2D rendering quality, establishing GHOST as an efficient and robust solution for realistic hand-object interaction modeling. Code is available at <https://github.com/ATABoukhadra/GHOST>.

1. Introduction

Hands enable nearly every form of physical interaction we perform, such as grasping, manipulating objects, operating devices, or expressing intent [3, 16, 21, 52, 55]. As a result, reconstructing realistic 3D hand-object interactions from monocular RGB videos has become fundamental for virtual reality (VR), augmented reality (AR), teleoperation, and embodied AI. Achieving this, however, remains challenging: mutual occlusions between hands and objects, large variation in object topologies, and inherent depth

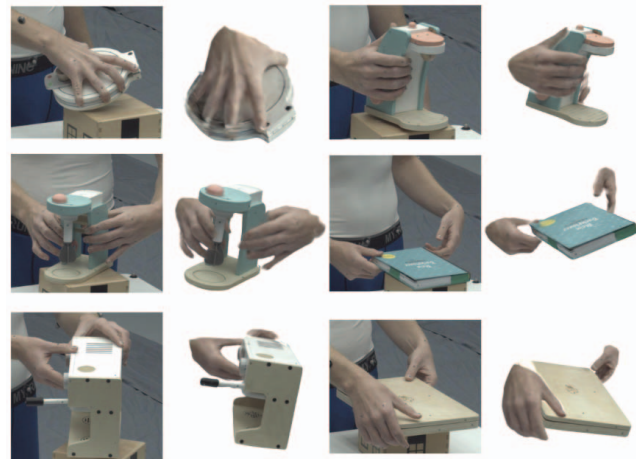


Figure 1. Our method reconstructs complete 3D hand-object interactions from a single monocular RGB video—recovering full object surfaces and realistic hand contact even under severe occlusions—while enabling fast, accurate, category-agnostic reconstruction and novel-view rendering.

ambiguities in monocular settings often lead to unstable scales and physically inconsistent contact.

Although significant progress has been made, existing hand-object reconstruction methods typically fall into two categories. Template-based approaches [6, 18, 19] can generate high-quality meshes, but are restricted to a small, fixed set of known object shapes, restricting their ability to generalize to novel or unconventional objects encountered in real-world scenarios. Category-agnostic methods remove this constraint but are often computationally expensive [14, 36], struggle to recover missing geometry under occlusion [1], or require multi-view supervision [40]. NeRF-based pipelines such as HOLD [14] achieve photorealistic rendering, yet demand hours of per-sequence optimization. Recent Gaussian-splatting frameworks improve runtime and provide explicit 3D representations [36, 40]. However, these methods still tend to produce unrealistic hand-object contact under severe occlusions, and their optimization

remains time-consuming, limiting practical applicability.

To address these challenges, we present **GHOST**, a fast, category-agnostic framework for reconstructing realistic bimanual hand-object interactions from monocular RGB videos. GHOST represents hands and objects as dense, view-consistent Gaussian discs, enabling complete and physically consistent reconstructions even under heavy occlusions. By integrating grasp-aware reasoning with geometric object priors to recover hidden surfaces, GHOST achieves high-quality, photorealistic rendering and produces animatable reconstructions. The framework operates in three stages: (1) Preprocessing, which initializes hand meshes and retrieves object priors; (2) hand-object alignment, which refines scale and hands translation through grasp-aware reasoning; and (3) Gaussian-splatting optimization, which jointly reconstructs hands and objects with occlusion-aware consistency. Together, these components enable GHOST to achieve state-of-the-art performance in both 3D accuracy and 2D rendering fidelity (See Fig. 1) while running over an order of magnitude faster than previous category-agnostic methods. Our contributions are summarized as follows:

- A fast, category-agnostic framework for reconstructing animatable bimanual hand-object interactions from monocular RGB sequences.
- A prior-aware reconstruction strategy that fills occluded object regions using geometric consistency from retrieved priors.
- A grasp-aware alignment that ensures realistic and stable hand-object contact.
- A hand-aware background loss that preserves valid object regions despite persistent occlusion.
- State-of-the-art performance on both 3D reconstruction and 2D rendering metrics, while achieving over 13× faster runtime than prior category-agnostic approaches.

We conduct an extensive evaluation across three datasets: ARCTIC, HO3D, and in-the-wild videos, covering both 3D geometric and 2D photometric metrics. GHOST consistently surpasses previous methods in hand-object reconstruction accuracy and rendering quality, establishing a new benchmark for fast and realistic category-agnostic hand-object reconstruction.

2. Related Work

The problem of realistic hand-object reconstruction spans multiple subdomains, including monocular hand reconstruction under occlusion [39], animatable hand avatars [25], and category-agnostic interaction reconstruction [14, 36].

3D Hand Reconstruction from RGB under Occlusions.

Monocular hand reconstruction has been extensively studied in [1, 12, 39, 41]. Recent Transformer-based approaches [1, 30, 38, 39, 41] significantly improve robustness

under occlusions, achieving accurate 3D hand meshes. HaMeR [39], in particular, provides a strong baseline trained on a large-scale dataset and generalizes well to in-the-wild scenarios, making it suitable for initializing hands in interaction pipelines.

Neural and Gaussian-based Avatars. Neural implicit representations such as NeRFs [34], and the more recent explicit formulations like Gaussian Splatting [26], have become standard for photorealistic, animatable human avatars [5, 37, 42]. Recent works [28, 35, 37, 50] initialize a set of 3D Gaussians on the surface of a canonical SMPL [32] or FLAME [29] mesh and animate them through parameter-controlled skeletal motion. Rendering is performed by projecting the Gaussians onto the image plane, where they are blended to produce smooth, photorealistic images of the avatar.

Hand-Object Interaction Reconstruction. Objects, unlike hands, lack consistent topology or parametric structure and exhibit wide variations in shape, material, and appearance. Early approaches [17–19] jointly reconstructed hand and object meshes from RGB inputs, enforcing contact through attraction–repulsion losses and leveraging photometric or silhouette consistency. However, these methods rely on fixed object templates, limiting generalization. Template-free representations such as THOR-Net [1] and ShapeGraformer [2] addressed this limitation by modeling objects as spherical deformations with shared topology, enabling category-agnostic learning.

Category-agnostic Hand-Object Reconstruction. Building on implicit neural fields [34], several category-agnostic pipelines [14, 53, 54, 56] reconstruct hand-object interactions directly from monocular sequences. HOLD [14] uses off-the-shelf hand pose estimators [30, 39], SAM-based object segmentation [9, 27], and HLoc-based SfM [46–48] to recover geometry via NeRF optimization, but requires hours per sequence. BIGS [36] replaces the NeRF stage with Gaussian Splatting [26], slightly improving runtime and explicitness of the representation. However, these methods still produce unrealistic contact under occlusions and high computational cost, motivating our efficient and physically consistent Gaussian-based framework.

3. Method

Our method reconstructs hand-object interactions from monocular RGB videos. It consists of three main stages: a preprocessing pipeline that extracts geometric and motion cues, a HO alignment stage that uses tracking and grasp-aware reasoning to align hands with the object, and a Gaussian Splatting optimization stage that jointly reconstructs

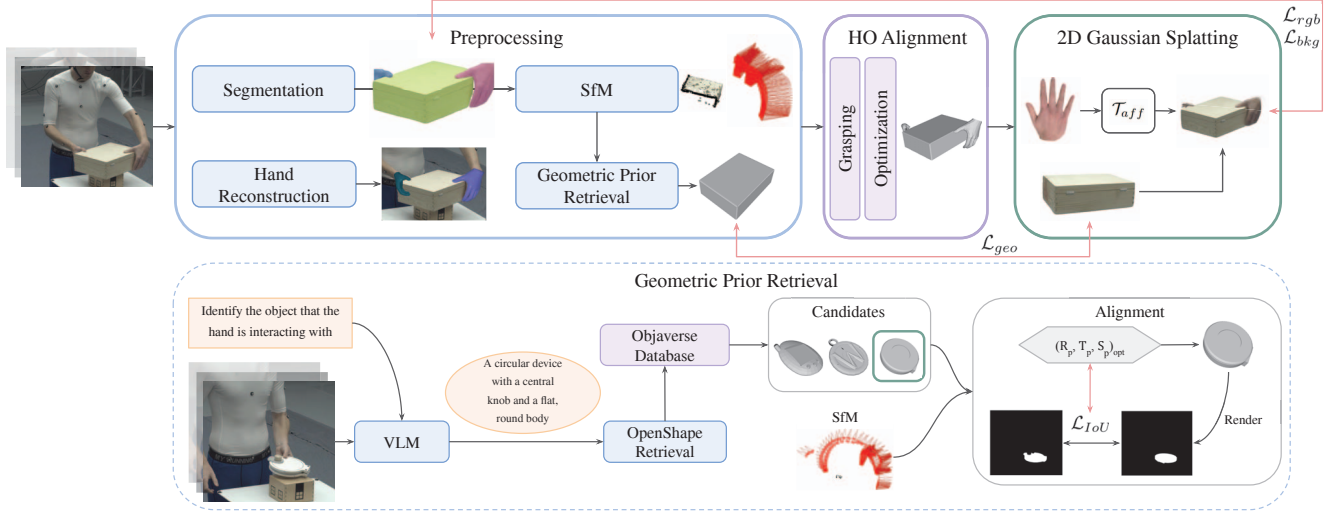


Figure 2. (Top) Overview of our pipeline, which consists of three stages. In **preprocessing**, we extract hand meshes, camera poses, and object information (i.e. mask, point cloud, and geometric prior). During **hand-object alignment**, object’s scale and hand translations are optimized using grasp-aware and temporal reasoning. In the **Gaussian Splatting** stage, hands and objects are jointly reconstructed with occlusion-aware losses. (Bottom) Retrieval and alignment pipeline used to obtain the object’s geometric prior.

the photorealistic hands and object in 3D. Fig. 2 summarizes the stages of our pipeline. In the following subsections, we describe each component in detail.

3.1. Preprocessing Pipeline

Given an input video $V = \{I_t\}_{t=1}^T$, with T frames, we first segment the object using SAM2 [43], obtaining per-frame masks $\mathcal{M}_t^o$. Using the object’s segmentation, we apply Structure-from-Motion (SfM) to estimate camera intrinsics $\mathbf{K}$, relative camera poses $\{(\mathbf{R}_t^c, \mathbf{T}_t^c)\}_{t=1}^T$, and a sparse object point cloud $\mathcal{P}_{sfm}$. For each frame t , the full camera projection function is denoted as: $\Pi_t(\mathbf{x}) = \mathbf{K}[\mathbf{R}_t^c | \mathbf{T}_t^c][\mathbf{x}]$. Given the critical influence of SfM on subsequent stages, we compare two SfM approaches. In the first, following HOLD [14], we employ HLoc [46] with COLMAP [48, 49], enhanced by a temporal-window pairing strategy to improve candidate matching. In parallel, we evaluate the recent VGGSfM [51] video-based pipeline and demonstrate its benefits for the ARCTIC [13] dataset.

3.1.1. Object Geometric Prior

Reconstructing objects under hand manipulation is challenging due to frequent hand and self-occlusions, which often leave unseen regions incomplete. To address this, we leverage open-source large-scale 3D object databases such as Objaverse [10, 11], which provide extensive coverage of everyday objects. These models serve as geometric priors to refine and complete partial reconstructions, thereby improving both object accuracy and hand-object interaction quality. Fig. 2 (Bottom part) shows our geometric prior retrieval and alignment algorithm.

Retrieval. A textual description d of the hand-held object is obtained from a few sampled video frames using a vision-language model (e.g., InternVL [7, 8]) or provided directly (e.g., ‘Box’). The text embedding $\phi_{txt}(d)$ is computed using OpenShape’s [31] CLIP model, which maps text and object meshes into a shared embedding space. Since all Objaverse objects are pre-embedded, retrieval reduces to a nearest-neighbor search between $\phi_{txt}(d)$ and stored embeddings $\{\phi_{obj}(\mathcal{O}_p)\}$, producing k candidate meshes $\{\mathcal{O}_1, \dots, \mathcal{O}_k\}$. We use a ray casting algorithm to simplify these structures and extract their outer surface.

Prior-Mask Alignment. The retrieved 3D meshes are not aligned with the object’s point cloud $\mathcal{P}_{sfm}$; they may be rotated, shifted, or scaled differently. Therefore, we estimate an affine transformation that includes quaternion rotation $\mathbf{R}_p$, 3D translation $\mathbf{T}_p$, and 3D scale $\mathbf{S}_p$ for each retrieved candidate mesh $\mathcal{O}_p$. We optimize previous variables such that, after transforming the mesh and projecting it into the camera view, the rendered mesh silhouette matches the object’s mask $\mathcal{M}_t^o$. This alignment is measured using the Intersection-over-Union loss $\mathcal{L}_{IoU}$ between the rendered silhouette and the ground-truth mask. More formally:

$$\mathcal{L}_{IoU} = 1 - \text{IoU}(\Pi_t([\mathbf{R}_p | \mathbf{T}_p][\mathbf{S}_p \cdot \mathcal{O}_p]), \mathcal{M}_t^o). \quad (1)$$

For all subsequent optimization stages, we select the object’s geometric prior $\mathcal{O} \in \mathbb{R}^{N \times 3}$ with N vertices as the transformed candidate $[\mathbf{R}_p | \mathbf{T}_p][\mathbf{S}_p \cdot \mathcal{O}_p]$ with the lowest $\mathcal{L}_{IoU}$. If the retrieved prior has a low number of vertices, we densify its point cloud by adding more points on its surface.

3.1.2. Hand Reconstruction Initialization

We obtain initial 3D meshes of both hands using an off-the-shelf hand reconstruction model (i.e., HaMeR [39]), guided by bounding-box detections from RTMPose [22–24]. For each frame I_t , HaMeR produces MANO [45] parameters consisting of pose θ_t , shape β_t , and global hand rotation $\mathbf{R}_t^h$. In practice, strong occlusions from the manipulated object often result in jittery or implausible hand reconstructions. To address this issue, we apply a post-processing step that detects and removes unreliable frames. Frames whose MANO parameters deviate beyond predefined thresholds from their neighboring frames are discarded, and their parameters are subsequently recomputed using linear interpolation for translations and spherical interpolation for rotations¹. This process produces temporally consistent 3D hand vertices $\mathcal{V}_t^h$, 3D joints $\mathcal{J}_t^h$, and 2D joints $\mathcal{J}_{2D,t}^h$ for left and right hands $h \in \{r, l\}$ throughout the entire sequence. In addition, we segment and track the hand masks using SAM2 [43] and crop it using the hand’s bounding boxes to exclude the forearm, producing $\mathcal{M}_t^h$.

3.2. Hand Translation and Object Scale Optimization (HO Alignment)

Initial hand reconstructions and the object’s point cloud $\mathcal{P}_{sfm}$ may be misaligned in 3D, due to scale inconsistencies from SfM or drifting hand translations estimated by HaMeR. To correct this, we jointly refine global hand translations $\mathbf{T}_t^h$ and global object scale $\mathbf{S}_o$. We describe the details of this step in the following subsections. The optimized scale is then applied to the point cloud $\mathcal{P}_{sfm}$ and camera translations $\mathbf{T}_t^c$.

3.2.1. Grasping Detection

In bimanual hand-object interactions, changes to object poses are the result of hand interaction, and hence, the hand and object should be in close contact when the object is moving. To identify frames where the hands interact with the object, we compare the motion trajectories of both hands relative to the object’s motion. The object’s motion $\Delta\mathcal{T}^o$ is computed as the change in its point cloud center across frames. For each hand, we compute its translation change $\Delta\mathcal{T}^h$ and measure the cosine similarity between its motion and the object’s motion, projected to the x, y plane:

$$c_h = \Delta\hat{\mathcal{T}}_{xy}^o \cdot \Delta\hat{\mathcal{T}}_{xy}^h, \quad (2)$$

where $\Delta\hat{\mathcal{T}}$ denotes normalized translation direction vectors. If both hands have $c_h > \tau_{sim}$, the frame is labeled as a two-hand grasp; otherwise, the hand with the higher score is considered the interacting one. We empirically choose $\tau_{sim} = 0.5$.

¹See the supplementary (Appendix. A) for more details.

3.2.2. Optimization Objective

Once grasping frames are identified, we optimize a joint objective consisting of three complementary loss terms.

Contact Loss ($\mathcal{L}_{contact}$). Applied only during grasping frames, the contact loss enforces the interacting hand to remain close to the object. It is defined as the Chamfer Distance (CD) between the translated hand mesh vertices $\mathcal{V}_t^h$ and the scaled object geometric prior $\mathcal{O}$. If only one hand is grasping, the loss is applied to that hand’s vertices. If both hands grasp, both sets of vertices are included.

$$\mathcal{L}_{contact} = \text{CD}(\mathbf{T}_t^h + \mathcal{V}_t^h, [\mathbf{R}_t^c \mid \mathbf{T}_t^c][\mathbf{S}_o \cdot \mathcal{O}]). \quad (3)$$

Hand 2D Projection Loss ($\mathcal{L}_{proj}$). This loss aims to use the camera intrinsics $\mathbf{K}$ to compute reprojection error between translated hand joints $\mathcal{J}_t^h$ and the HaMeR-predicted 2D hand joints and it is defined as:

$$\mathcal{L}_{proj} = \|\Pi_t(\mathbf{T}_t^h + \mathcal{J}_t^h) - \mathcal{J}_{2D,t}^h\|_1. \quad (4)$$

In practice, small bounding boxes typically indicate hands that are far from the camera and thus less reliable. Such low-confidence predictions are omitted when computing $\mathcal{L}_{proj}$. The purpose of this term is to maintain 2D consistency of the hands.

Temporal Consistency Loss ($\mathcal{L}_{temp}$). Enforces smooth hand translation across time defined as:

$$\mathcal{L}_{temp} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\mathbf{T}_{t+1}^h - \mathbf{T}_t^h\|_2^2. \quad (5)$$

Combined together, they form:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{contact} + \lambda_2 \mathcal{L}_{proj} + \lambda_3 \mathcal{L}_{temp}. \quad (6)$$

We empirically choose $\lambda_1 = 10^3$, $\lambda_2 = 10^{-1}$, and $\lambda_3 = 10$ to balance the loss term values¹.

3.3. Hand-Object Gaussian Splatting

3.3.1. Preliminary

Gaussian Splatting (GS) [26] represents static 3D scenes from calibrated multi-view images and enables rendering from novel viewpoints. Instead of optimizing a neural volume, the scene is modeled as a set of 3D Gaussians, each defined by its center, rotation, scale, opacity, and color encoded in spherical harmonics. During optimization, a differentiable rasterizer renders RGB views that are compared against ground-truth views using a photometric loss term consisting of an L1 and D-SSIM term, denoted in Fig. 2 as $\mathcal{L}_{rgb}$. The number of Gaussians adaptively changes through *densification* (adding Gaussians in sparse regions) and *pruning* (removing excessively large or low-opacity ones).

3.3.2. Object Optimization

Rogge *et al.* [44] applied 2D Gaussian Splatting [20] to reconstruct segmented objects from RGB videos. To suppress Gaussians rendered outside object masks, they applied a background loss $\mathcal{L}_{bkg}$ that penalizes background-projected Gaussians during training. Following their setup, we fit Gaussians to the object in the first stage of our pipeline, where the set of all Gaussians is parameterized as,

$$\mathcal{G}_o = \{c_o, r_o, s_o, \alpha_o, SH_o\}. \quad (7)$$

Here, c_o are 3D centers, r_o are 3D rotations, s_o are 2D scales, α_o are opacities, and SH_o are spherical harmonics that represent the color of the Gaussians. We further introduce two losses tailored for hand-object interaction. We explain these losses and their rationale in the following paragraphs.

Hand-aware Background Loss ($\mathcal{L}_{bkg,h}$). While effective in single-object scenes, the background loss $\mathcal{L}_{bkg}$ is less suited for cases with strong occlusions, such as when a hand holds the object. In such cases, occluded regions may be incorrectly removed (see finger-shaped holes in Fig. 3a). To mitigate this, we combine both hands and object masks to more reliably distinguish foreground from background during object reconstruction creating an updated hand-aware background loss term ($\mathcal{L}_{bkg,h}$). Using hand masks ensures that occluded object regions are retained rather than mistakenly discarded as background. Although hand masks inevitably include some non-object areas, the temporal variation across frames -where the object and hands appear under different orientations- guarantees that all true background regions are eventually removed over time.

Geometric Consistency Loss ($\mathcal{L}_{geo}$). Using the geometric prior $\mathcal{O}$ from earlier stages, we introduce a novel geometric consistency loss $\mathcal{L}_{geo}$ to keep the reconstructed Gaussians consistent with the prior surface. The first component, $\mathcal{L}_{out}$, penalizes Gaussians whose centers are farther than a threshold τ_{out} from the prior, preventing them from drifting away from the object surface. The second component, $\mathcal{L}_{fill}$, measures the distance from each point on the prior surface to the closest Gaussian center and penalizes large gaps that exceed a threshold τ_{fill} , encouraging the model to fill holes caused by occlusions. The final loss is a weighted combination where:

$$\mathcal{L}_{out} = \sigma(\|c_o - \mathcal{O}\|_2 - \tau_{out})^2, \mathcal{L}_{fill} = \sigma(\|\mathcal{O} - c_o\|_2 - \tau_{fill})^2, \quad (8)$$

and

$$\mathcal{L}_{geo} = \lambda_{geo}(\mathcal{L}_{out} + \mathcal{L}_{fill}), \quad (9)$$

where σ is the ReLU operation and λ_{geo} is empirically set to 5. In essence, $\mathcal{L}_{geo}$ keeps Gaussians close to the prior while promoting completeness in the reconstructed surface.

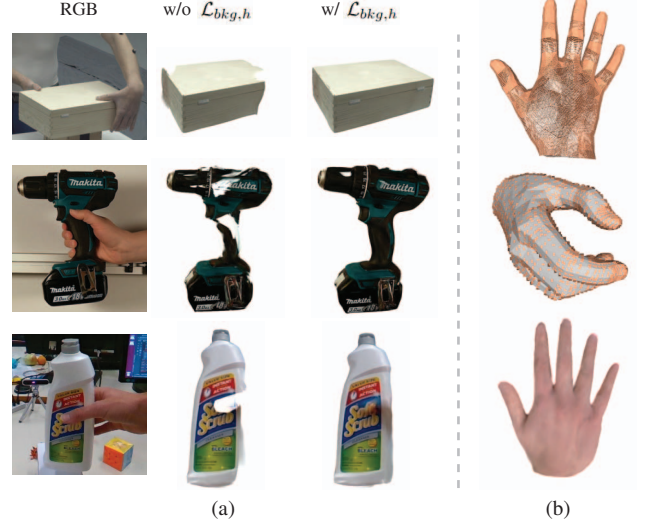


Figure 3. Qualitative comparison showing (a) the effect of our novel background loss $\mathcal{L}_{bkg,h}$ on object reconstruction quality. (b) **Top:** Gaussian centers on the canonical hand; **middle:** deformed hand mesh with aligned Gaussian centers after $\mathcal{T}_{aff}$; **bottom:** final animatable Gaussian hand after training.

3.3.3. Hand Optimization

Unlike rigid objects, hands deform over time, making Gaussian Splatting more challenging. To handle this, we follow GaussiansAvatars [42] and represent hands in a canonical space by attaching Gaussians to the MANO [45] mesh faces and, similar to the object, represent canonical hand Gaussians as:

$$\mathcal{G}_h = \{c_h, r_h, s_h, \alpha_h, SH_h\}. \quad (10)$$

Hand Gaussians Rigging. Let the canonical MANO mesh be defined by vertices $\mathcal{V} \in \mathbb{R}^{778 \times 3}$ and faces $\mathcal{F} \in \mathbb{R}^{1538 \times 3}$. For each face $f \in \mathcal{F}$, we attach m Gaussians with canonical centers $c_{h,f} \in \mathbb{R}^{m \times 3}$, rotations $r_{h,f} \in SO(3)^m$, and scales $s_{h,f} \in \mathbb{R}^{m \times 2}$. Given deformed hand meshes $\mathcal{V}_t^h$, we compute for each face a local affine transformation that maps from its canonical to deformed state. This transformation $\mathbf{M}_{f,t}$ is derived from basis matrices constructed over the face vertices in the canonical pose ($\mathbf{B}_f^c$) and the deformed pose ($\mathbf{B}_{f,t}^d$):

$$\mathbf{M}_{f,t} = \mathbf{B}_{f,t}^d \mathbf{B}_f^{c^{-1}}. \quad (11)$$

Each Gaussian i attached to face f then deforms according to the following equations:

$$c_{h,f,i}^d = \mathbf{M}_{f,t} c_{h,f,i}, \quad r_{h,f,i}^d = \text{Polar}(\mathbf{M}_{f,t}), \quad (12)$$

where $\text{Polar}(\cdot)$ extracts the rotation from $\mathbf{M}_{f,t}$. This formulation allows Gaussians to follow the mesh deformation of

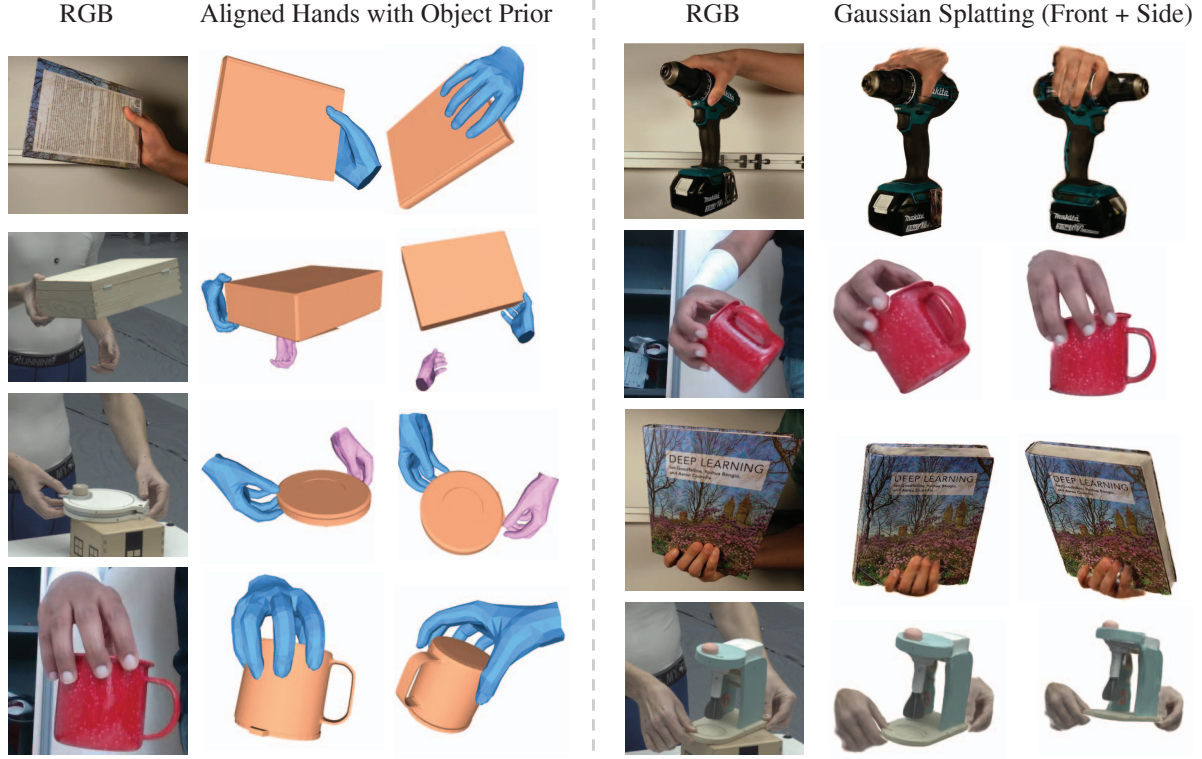


Figure 4. Qualitative results of GHOST on ARCTIC [13], HO3D [15], and in-the-wild examples: **Left**: shows aligned 3D hand meshes with the object’s geometric prior obtained during HO alignment. **Right**: presents photorealistic Gaussian Splatting renderings from original viewpoint and novel viewpoints. GHOST produces consistent hand–object alignment in 3D and maintains realistic appearance even under view changes, enabling physically plausible interaction reconstruction and high-fidelity rendering across viewpoints.

their corresponding faces, as shown in Fig. 3b. We refer to this operation by $\mathcal{T}_{aff}$.

Finally, we optimize the hand Gaussians by loading the pretrained object representation $\mathcal{G}_o$ into the 3D scene and optimizing only (s_h, α_h, SH_h) , while keeping mesh-driven parameters $(c_h$ and $r_h)$ fixed. We also allow the hand translation $(\mathbf{T}_t^h)$ to change during optimization to create more 2D consistency for the rendered hand. We disable the steps of pruning and densification in the joint hand-object optimization and study the effect of m on the rendering quality. This two-stage optimization—object first, then deformable hands—produces a coherent joint representation of hands and object renderable from arbitrary viewpoints.

4. Experiments and Results

In this section, we describe our experimental setup, including the datasets, evaluation metrics, ablation studies, and comparisons with state-of-the-art methods.

4.1. Datasets

We evaluate our method on three sources of data. First, we use the **ARCTIC Bi-CAIR** dataset [13, 14], which provides

9 video sequences of two hands interacting with diverse objects. Second, we evaluate sequences from the **HO3D** dataset [15]. This benchmark provides a controlled setup for assessing category-agnostic object reconstruction under severe hand occlusions. Finally, we capture two additional hand–object interaction sequences (Drill and Book) using a GoPro camera. These videos are used for qualitative evaluation to demonstrate the generalization of our method beyond existing datasets.

4.2. Metrics

All 3D evaluation metrics follow the official protocol released by HOLD [14] for the ARCTIC [13] Bi-CAIR challenge². For hands, we report the mean per-joint position error (MPJPE), computed on root-aligned hand joints. For objects, we evaluate multiple Chamfer-based metrics. Object’s Chamfer Distance relative to each hand’s root (CD_r , CD_l) quantify the interaction quality between the hand and object surfaces, while also being sensitive to the object’s absolute scale, as no rigid alignment is applied. Their average forms CD_h . We further compute CD_{ICP} ,

²<https://arctic-leaderboard.is.tuebingen.mpg.de/leaderboard>

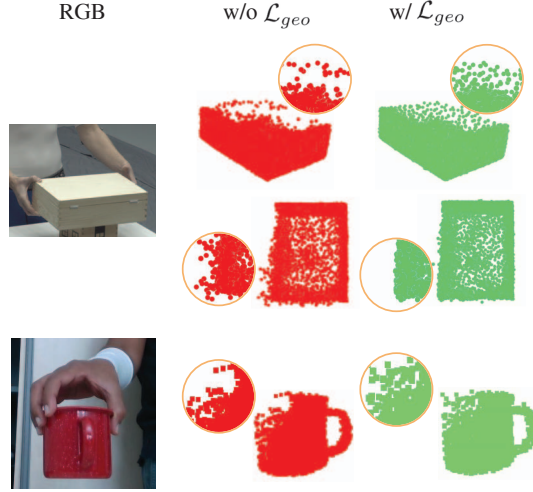


Figure 5. Qualitative comparison demonstrating the effect of the geometric loss $\mathcal{L}_{geo}$ on the quality of reconstructed object point clouds derived from Gaussian centers.

which aligns predicted and ground-truth object point clouds via iterative closest point (ICP) before computing the distance, isolating reconstruction quality from global scale or translation errors.

We also assess the photometric fidelity of rendered reconstructions. We compute three standard image-quality metrics between rendered outputs and ground-truth RGB: PSNR, SSIM, and LPIPS. Finally, we report the average optimization runtime on a single NVIDIA RTX A6000 GPU, using an ARCTIC sequence comprising 300 frames of bimanual hand-object interaction. This comparison highlights the relative efficiency of our Gaussian-based formulation compared to prior NeRF- and diffusion-based pipelines.

4.3. Ablation Study

Structure-from-Motion (SfM). In Table 1, we evaluate the impact of different SfM methods and the geometric consistency loss $\mathcal{L}_{geo}$ on object reconstruction metrics and interaction metrics. Category-agnostic object reconstruction

Table 1. Ablation study on the impact of SfM method and geometric consistency loss on object reconstruction and interaction scores for the ARCTIC dataset. τ_{fill} is reported in mm , τ_{out} is in cm , and CD metrics are in cm^2 .

SfM	$\mathcal{L}_{geo}$	τ_{fill}	τ_{out}	$CD_h \downarrow$	$CD_{ICP} \downarrow$
HLoc [46–48]	✗	-	-	-	8.18
HLoc [46–48]	✓	-	-	-	4.86
VGGSfM [51]	✗	-	-	18.32	2.78
VGGSfM [51]	✓	1.0	5.0	18.21	3.66
VGGSfM [51]	✓	0.1	2.0	18.42	2.33
VGGSfM [51]	✓	0.01	2.0	18.37	2.83
VGGSfM [51]	✓	0.01	7.0	18.4	2.26

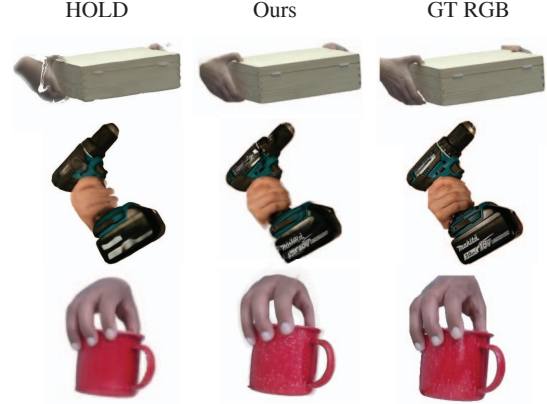


Figure 6. **Qualitative comparison of 2D rendered images:** We compare our results against *HOLD* [14] using representative examples from the evaluated datasets. As shown, *ours* produces higher quality and more consistent 2D renderings, whereas *HOLD* exhibits noise, blur, and degraded visual quality.

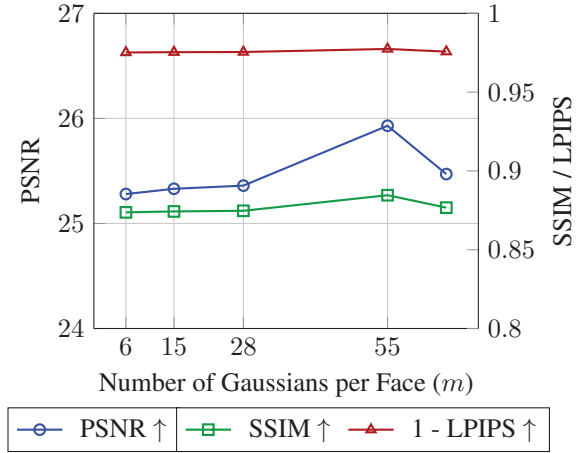


Figure 7. Rendering quality as a function of surface Gaussian density. Higher density improves reconstruction fidelity.

tion from videos is very sensitive to the object’s Structure-from-Motion (SfM) as noted in previous methods [14, 36]. Therefore, we compare two SfM pipelines (i.e., HLoc combined with COLMAP [46–48] and VGGSfM [51]). We observed significant improvements using VGGSfM on the ARCTIC data compared to the HLoc pipeline. However, VGGSfM did not show similar improvements on the HO3D dataset sequences [15] and our own recorded data.

$\mathcal{L}_{geo}$ and $\mathcal{L}_{bkg,h}$. The purpose of $\mathcal{L}_{geo}$ is to retrieve occluded regions of the object. Fig. 5 illustrates the impact of applying $\mathcal{L}_{geo}$ on the point cloud obtained from Gaussian centers c_o . It shows that point clouds become clustered around their priors completing missing regions of the object. Table 1 compares how $\mathcal{L}_{geo}$ and its hyperparameters (τ_{fill} and τ_{out}) affect the object reconstruction for

Table 2. Comparison of 3D metrics across HOLD, BIGS, and our method. All MPJPE metrics are in mm , and all CD metrics are in cm^2 .

Method	MPJPE _{RA,h} ↓	MPJPE _{RA,l} ↓	MPJPE _{RA,r} ↓	CD _{ICP} ↓	CD _r ↓	CD _l ↓	CD _h ↓	F _{10mm} (%)↑	F _{5mm} (%)↑
HOLD [14]	25.91	27.13	24.70	2.07	123.54	105.92	114.73	63.92	37.13
BIGS [36]	24.49	24.63	24.35	1.36	31.28	46.11	38.69	81.78	56.41
Ours	24.07	25.42	22.71	2.26	13.40	23.41	18.40	60.88	34.67

Table 3. Evaluation of 2D rendering quality on ARCTIC and HO3D datasets, comparing our method with prior works. In addition, we report significant runtime improvement.

Method	2D Rendering Quality (ARCTIC)			2D Rendering Quality (HO3D)			Runtime↓
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
HOLD [14]	12.83	0.66	0.32	16.20	0.74	0.21	16h
BIGS [36]	24.87	0.96	0.05	24.51	0.92	0.07	13h
Ours	25.93	0.88	0.02	21.37	0.75	0.03	1h

the ARCTIC dataset. Decreasing τ_{out} keeps the object’s point cloud closer to the prior, restricting the freedom of the gaussians to roam away from the prior where needed. Lower τ_{fill} results in more holes in the point cloud being filled. Note that we only apply $\mathcal{L}_{geo}$ on sequences where high-quality geometric priors were found for the object. Supplementary material (Fig. 11) also includes more evaluation on the impact of $\mathcal{L}_{geo}$ on a selected set of HO3D sequences. On the other side, applying $\mathcal{L}_{bkg,h}$ focuses on retrieving hand-covered object parts. Fig. 3a illustrates how $\mathcal{L}_{bkg,h}$ improved the reconstruction of the object under hand occlusion.

Number of Gaussians per hand (m). Furthermore, Fig. 7 shows the importance of the number of Gaussians (m) attached to hand mesh faces on the rendering quality of the final hand-object reconstructions. We observe that using 55 Gaussians per face and in total 84k Gaussians per hand produces the highest rendering quality. Because the number of training iterations is fixed (30k), adding more Gaussians makes them harder to fit properly. The optimization spends more time adjusting the larger set, and the shape becomes more complex, which eventually reduces accuracy.

4.4. Comparison with state-of-the-art

Our approach consistently outperforms previous methods on the ARCTIC dataset, achieving lower interaction metrics (CD_h, CD_r, and CD_l) as reported in Table 2. This improvement is attributed to the grasp-aware contact loss in HO alignment and the novel object reconstruction losses. Furthermore, we show the importance of jitter detection discussed in Section 3.1.2 on the improvement of hand reconstruction errors (MPJPE), knowing that previous methods also use HaMeR [39] for hand initialization. In addition, we observe an improvement in interaction score (CD_r) on 3 sequences and an improvement in CD_{ICP} on 2 sequences from the HO3D dataset [15] compared to HOLD [14].

To evaluate the rendering quality of our GS pipeline,

we report 2D rendering quality metrics in comparison with previous methods in Table 3. We observe an improvement in PSNR and LPIPS compared to BIGS [36] and a significant improvement compared to HOLD in all metrics on the ARCTIC dataset [13]. This can also be observed qualitatively in Fig. 6, where we show different rendering examples in comparison to HOLD. All this is obtained while achieving 13× improvement in runtime as reported in Table 3. More qualitative results are shown in Figs. 1 and 4.

5. Conclusion

In this paper, we introduced GHOST, a fast and category-agnostic framework for reconstructing realistic hand-object interactions from monocular RGB videos using 2D Gaussian Splatting. By combining geometric-prior retrieval, grasp-aware alignment, and hand-aware background reasoning, GHOST produces complete object surfaces, physically consistent hand-object contact, and animatable hand avatars. GHOST sets a new state-of-the-art baseline in 3D interaction reconstruction accuracy and 2D rendering quality on the ARCTIC Bi-CAIR benchmark, while running over 13× faster than existing category-agnostic baselines. Beyond quantitative gains, GHOST also produces animatable hand avatars and photorealistic novel-view renderings. In future work, we will extend GHOST to deformable and articulated objects, explore direct integration of geometric priors within the SfM pipeline, and investigate real-time inference from stereo or RGB-D streams. These extensions can enable practical deployment in teleoperation, interactive AR/VR systems, robotics manipulation, and in-the-wild motion capture settings where speed and physical plausibility are critical. **Acknowledgments:** This work was partially funded by the European Union’s Horizon Europe programme through the SHARESPACE project (grant no. 101092889) and the LUMINOUS project (grant no. 101135724).

References

- [1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1001–1010, 2023. 1, 2
- [2] Ahmed Tawfik Aboukhadra, Jameel Malik, Nadia Robertini, Ahmed Elhayek, and Didier Stricker. Shapegraformer: Graformer-based network for hand-object reconstruction from a single depth map. *IEEE Access*, 12:124021–124031, 2024. 2
- [3] Ahmed Tawfik Aboukhadra, Nadia Robertini, Jameel Malik, Ahmed Elhayek, Gerd Reis, and Didier Stricker. Surgeonet: Realtime 3d pose estimation of articulated surgical instruments from stereo images using a synthetically-trained network. In *DAGM German Conference on Pattern Recognition (GCPR)*, pages 199–211. Springer, 2024. 1
- [4] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014. 3
- [5] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8683–8693, 2023. 2
- [6] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12890–12900, 2023. 1
- [7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 3
- [9] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 2
- [10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 35799–35813, 2023. 3, 1
- [11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 3, 1
- [12] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:2127–2160, 2024. 2
- [13] 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. 3, 6, 8
- [14] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J. Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 494–504, 2024. 1, 2, 3, 6, 7, 8
- [15] 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 (CVPR)*, pages 3196–3206, 2020. 6, 7, 8, 1
- [16] Guangyi Han, Wei Zhai, Yuhang Yang, Yang Cao, and Zheng-Jun Zha. Touch: Text-guided controllable generation of free-form hand-object interactions. *arXiv preprint arXiv:2510.14874*, 2025. 1
- [17] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kaleyvnykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 2
- [18] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 571–580, 2020. 1
- [19] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021. 1, 2
- [20] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference papers*, pages 1–11, 2024. 5
- [21] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. Hoigt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 7136–7146, 2025. 1
- [22] Tao Jiang. rtmllib. <https://github.com/Tau-J/rtmllib>, 2023. 4, 1

- [23] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RtmPose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 1
- [24] Tao Jiang, Xinchun Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024. 4
- [25] 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 (CVPR)*, pages 12802–12813, 2023. 2
- [26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG), Proceedings of SIGGRAPH*, 42(4):139:1–139:14, 2023. 2, 4
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2
- [28] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 505–515, 2024. 2
- [29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2
- [30] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [31] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 44860–44879, 2023. 3, 1
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866, 2023. 2
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 3
- [34] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [35] Arthur Moreau, Jifei Song, Helisa Dhano, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 788–798, 2024. 2
- [36] Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17437–17447, 2025. 1, 2, 7, 8
- [37] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024. 2
- [38] Joonkyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1496–1505, 2022. 2
- [39] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 2, 4, 8, 1
- [40] 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 (CVPR)*, pages 2197–2208, 2024. 1
- [41] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 12242–12254, 2025. 2
- [42] 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 (CVPR)*, pages 20299–20309, 2024. 2, 5, 1
- [43] 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. 3, 4, 1
- [44] Marcel Rogge and Didier Stricker. Object-centric 2d gaussian splatting: Background removal and occlusion-aware pruning for compact object models. In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 519–530, 2025. 5
- [45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG), Proceedings of SIGGRAPH Asia*, 36(6):245:1–245:17, 2017. 4, 5, 3

- [46] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#), [3](#), [7](#)
- [47] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [48] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [2](#), [3](#), [7](#)
- [49] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. [3](#)
- [50] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1606–1616, 2024. [2](#)
- [51] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21686–21697, 2024. [3](#), [7](#), [2](#)
- [52] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [1](#)
- [53] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What’s in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3895–3905, 2022. [2](#)
- [54] Yufei Ye, Poorvi Hebbbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19717–19728, 2023. [2](#)
- [55] 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 (CVPR)*, pages 445–456, 2024. [1](#)
- [56] Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, Fabian Manhardt, Bowen Fu, Federico Tombari, and Xiangyang Ji. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9992–10002, 2024. [2](#)