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

Supplementary Material

In our supplementary material, we give more illustrations of different steps in our preprocessing pipeline in Figs. 8 and 9. Furthermore, we compare different design choices and their qualitative gains in Fig. 10. Fig. 11 shows quantitative evaluation of applying $\mathcal{L}_{geo}$ on 5 sequences from the HO3D dataset [15]. We observe an improvement in the interaction distance relative to the right hand root (CD_r) on 3 sequences. Fig 12 shows limitations of $\mathcal{L}_{bkg,h}$ and $\mathcal{L}_{geo}$. In addition, Appendix A and B discuss different hyperparameters in our approach. Finally, we show examples of our animatable hand avatar visualization inherited from GaussianAvatars [42] viewer in Fig. 13.

A. Postprocessing HaMeR Predictions

For image I_t with area A_{img}, RTMPose [22, 23] provides hand keypoints for left and right hands with confidence for each keypoint producing right and left hand bounding boxes (B_t^r, B_t^l) with areas (A_t^r, A_t^l) and averaged keypoint confidence (c_t^l and c_t^r). Using the hand bounding boxes, HaMeR [39] generates initial hand reconstructions as stated in Section 3.1.2. For timestep t, our algorithm decides based on the combined rejection rule in Eq. 20 for each hand $h \in \{r, l\}$ if the predictions of the frames should be discarded and interpolated or not. The individual conditions are defined as follows:

1. Pose jitter condition:

$$\mathcal{C}_{\text{p}} = \left(\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2 > \tau_{\text{p}} \ \wedge \ \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 > \tau_{\text{p}}\right), \tag{13}$$

2. Orientation jitter condition:

$$\mathcal{C}_o = \left(\|\mathbf{R}_t^h - \mathbf{R}_{t-1}^h\|_2 > \tau_o \ \wedge \ \|\mathbf{R}_t^h - \mathbf{R}_{t+1}^h\|_2 > \tau_o\right), \tag{14}$$

3. Translation jitter condition (x–y plane):

$$\mathcal{C}_t = \left(\|\mathbf{T}_t^h - \mathbf{T}_{t-1}^h\|_2 > \tau_t \ \wedge \ \|\mathbf{T}_t^h - \mathbf{T}_{t+1}^h\|_2 > \tau_t\right), \tag{15}$$

4. Shape deviation condition:

$$\mathcal{C}_s = \frac{\left|\boldsymbol{\beta}_t - \text{median}(\boldsymbol{\beta})\right|}{\text{std}(\boldsymbol{\beta}) + \varepsilon} > \tau_{shape}. \tag{16}$$

5. Confidence threshold:

$$\mathcal{C}_c = c_t^h < \tau_{conf}, \tag{17}$$

6. Bounding-box area constraint:

$$\mathcal{C}_a = \left(A_t^h < A_{\text{min}} \ \vee \ A_t^h > A_{\text{max}}\right), \tag{18}$$

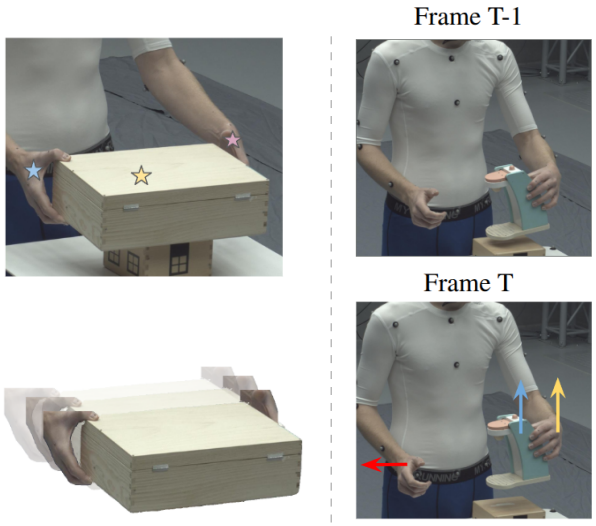

Figure 8. (a) SAM2 [43] is initialized with 3 seed pixels to segment and track the hands ($\mathcal{M}_t^h$) and the object ($\mathcal{M}_t^o$) in the scene. (b) During the grasping detection, the object's motion vector $\hat{\mathcal{T}}_{xy}^o$ (blue arrow) is compared with left hand's motion vector $\hat{\mathcal{T}}_{xy}^l$ (orange arrow) and right hand's motion vector $\hat{\mathcal{T}}_{xy}^r$ (red arrow). The example shows a left hand grasp based on the similarity between motion vectors.

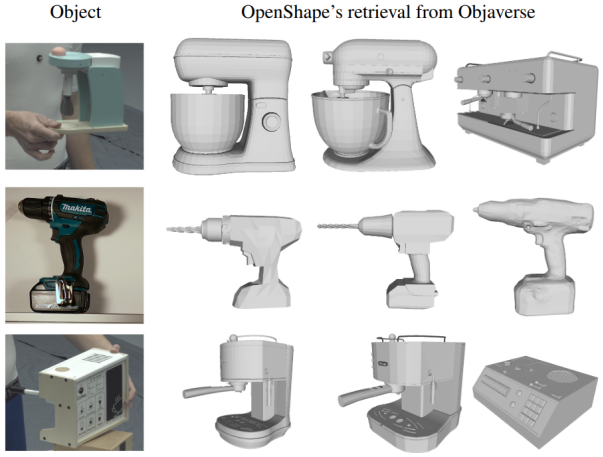

Figure 9. OpenShape [31] retrieves 3D models from Objaverse [10, 11], however, the retrieved 3D models do not always match the geometry of the desired object. Therefore, the final geometric prior $\mathcal{O}$ can be suboptimal.

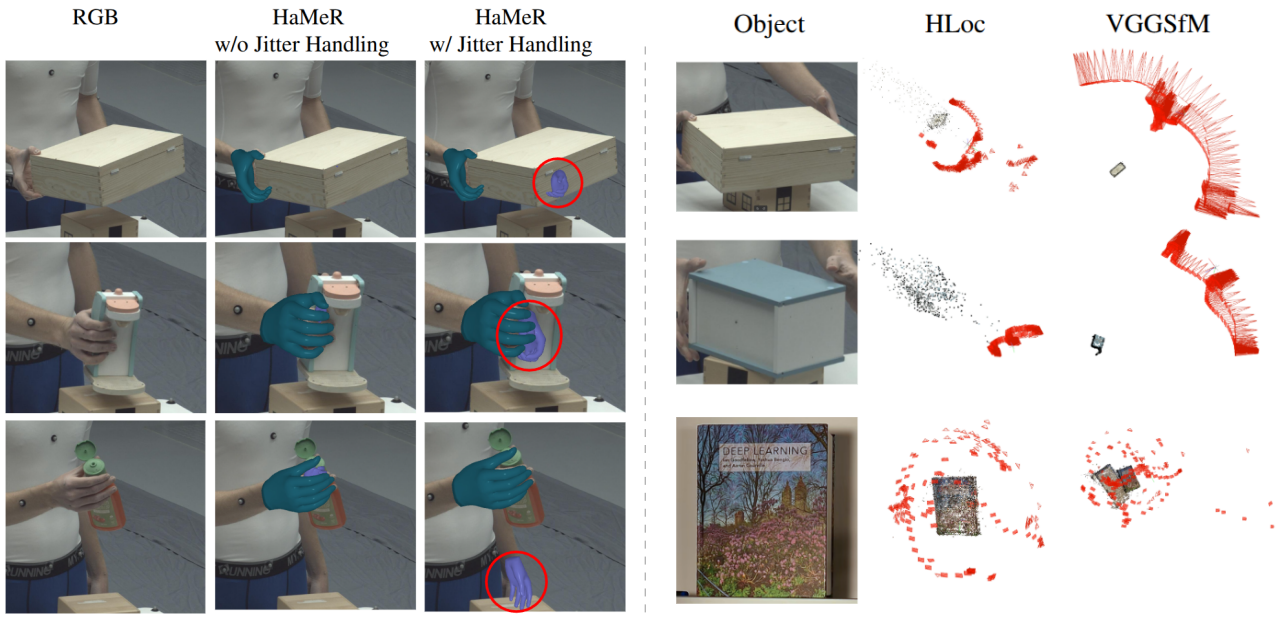

Figure 10. a) Initial hand reconstructions $\mathcal{V}_t^h$ obtained from HaMeR [39] suffer from Jitter under occlusion. Detecting jitter based on temporal cues, detection confidence, and interpolation results in improving initial hand meshes $\mathcal{V}_t^h$. b) Structure-from-Motion (SfM) has a large impact on subsequent steps. VGGSfM [51] improved SfM when applied to the arctic data compared to the HLoc+COLMAP [46, 48] pipeline. However, VGGSfM is sensitive to hyperparameter selection and does not always show similar improvements as seen in the last row.

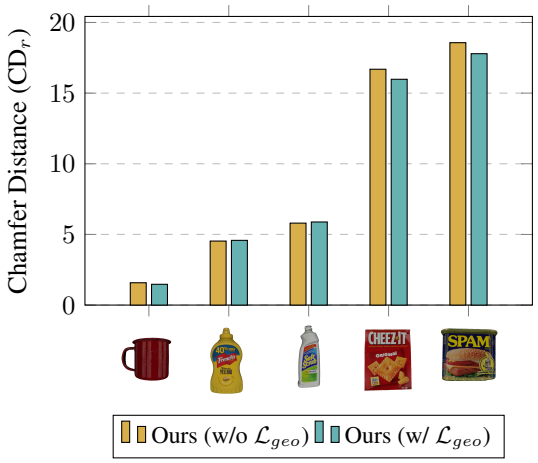

Figure 11. The influence of $\mathcal{L}_{geo}$ (CD$_r$, lower is better) across five HO3D sequences using distinct YCB objects.

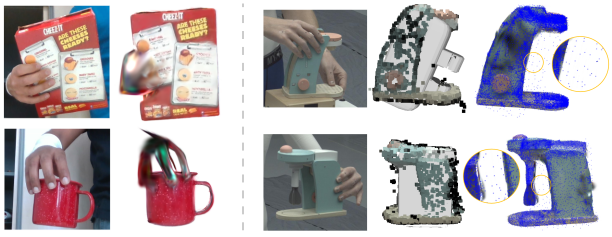

Figure 12. a) Applying $\mathcal{L}_{bkg,h}$ for object reconstruction fails when the hand never changes its contact point with the object. In that case, unwanted gaussians spawn in the hand region. b) In some cases, the retrieved geometric prior $\mathcal{O}$ does not align perfectly with the initial object's point cloud $\mathcal{P}_{sfm}$ (Middle column). The results of applying $\mathcal{L}_{geo}$ in this case will result in moving gaussian centers c_o towards unwanted regions (see blue point cloud).

7. Bounding-box overlap (IoU) constraint:

$$\mathcal{C}_{iou} = IoU(B_t^r, B_t^l), \tag{19}$$

where the thresholds are empirically chosen as: $\tau_\text{p} = 1.0$, $\tau_\text{o} = 1.0$, $\tau_\text{t} = 2.0$, $\tau_\text{s} = 4.0$, $\tau_\text{c} = 0.3$, $\tau_\text{iou} = 0.3$, $A_\text{min} = 0.006\, A_{img}$, and $A_\text{max} = 0.2\, A_{img}$.

Hand rejection rule:

$$\mathcal{F}_\text{reject} = \mathcal{C}_\text{p} \vee \mathcal{C}_\text{o} \vee \mathcal{C}_\text{t} \vee \mathcal{C}_\text{s} \vee \mathcal{C}_\text{c} \vee \mathcal{C}_\text{a} \vee \mathcal{C}_\text{iou}. \tag{20}$$

Fig 10a shows the importance of this rejection rule on hand meshes $\mathcal{V}_t^h$.

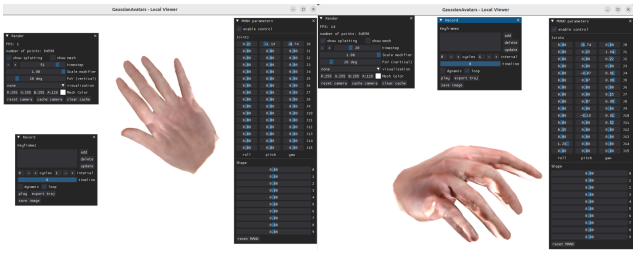

Figure 13. We also provide an interactive 3D viewer for Gaussian hand avatars. The interface visualizes and controls MANO-based [45] hand reconstructions, allowing users to adjust both pose and shape parameters in real time. Imported motion sequences can be played to animate the Gaussian hand avatar.

B. Experimental details

1. Prior Alignment Parameters. Optimizer: AdamW [33], LR: 10^{-2}, Betas: $0.9, 0.99$, eps:10^{-8}, Iterations: 1500.
2. HO Alignment. Optimizer: Adam [4], LR: 0.05, Iterations: 500.
3. Object Gaussian Splatting Optimization. Optimizer: Adam [4]. Iterations: 30000.
4. Hand-Object Gaussian Splatting Optimization. Optimizer: Adam [4]. Iterations: 30000. More details on the Gaussian Splatting hyperparameters are available in the code.