
Generalizable Hand-Object Modeling from Monocular RGB Images via 3D Gaussians (Supplementary Material)

Xingyu Liu^{*}, Pengfei Ren^{*}, Qi Qi, Haifeng Sun, Zirui Zhuang,
Jing Wang, Jianxin Liao, Jingyu Wang[†]
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{liuxingyu, rpf, qiqi8266, hfsun, zhuangzirui,
wangjing, liaojx, wangjingyu}@bupt.edu.cn

In this supplementary material, we provide:

- Details of mesh extraction (Section A).
- Optimization of hand-object Gaussians (Section B).
- Discussion of mesh alignment (Section C).
- More qualitative results (Section D).

A Mesh Extraction

Inspired by the impressive novel view synthesis capabilities of 3D Gaussian Splatting (3DGS), recent studies [4, 10] have focused on exploring the use of 3D Gaussian representations for surface reconstruction. To extract explicit meshes from hand-object Gaussians, we extend Gaussian Opacity Fields (GOF), a 3DGS-based surface reconstruction method tailored for unbounded scenes, into an interacting hand-object mesh reconstruction framework.

A.1 Preliminary: Gaussian Opacity Fields

GOF uses an explicit ray-Gaussian intersection instead of projection, which allows evaluating the opacity value or transmittance of any 3D point \mathbf{x} along the ray. At the most basic level, given a single 3D Gaussian \mathcal{G}_k , the opacity at any point along a ray can be defined as:

$$O_k(\mathcal{G}_k, \mathbf{o}, \mathbf{r}, t) = \begin{cases} \mathcal{G}_k^{1D}(t) & \text{if } t \leq t^* \\ \mathcal{G}_k^{1D}(t^*) & \text{if } t > t^* \end{cases} \quad (1)$$

where $\mathbf{x} = \mathbf{o} + t\mathbf{r}$. Intuitively, the opacity increases until it reaches its maximal value, and remains constant thereafter. Next, considering a set of 3D Gaussians \mathcal{G} , the opacity of point \mathbf{x} is given by:

$$O(\mathbf{o}, \mathbf{r}, t) = \sum_{k=1}^K \alpha_k O_k(\mathcal{G}_k, \mathbf{o}, \mathbf{r}, t) \prod_{j=1}^{k-1} (1 - \alpha_j O_j(\mathcal{G}_j, \mathbf{o}, \mathbf{r}, t)) \quad (2)$$

As a 3D point might be visible by any training view, the vanilla GOF defines the opacity field $O(x)$ of a 3D point x as the minimal opacity value among all training views or viewing directions:

$$O(x) = \min_{(\mathbf{o}, \mathbf{r})} O(\mathbf{o}, \mathbf{r}, t) \quad (3)$$

^{*} Equal contribution.

[†] Corresponding author.

A.2 Dynamic Hand-Object Reconstruction

However, for dynamic hand-object reconstruction, directly employing GOF faces two problems: (1) We observe the interacting hand-object from a single viewing direction, rather than scanning various views in an unbounded scene, making it impossible to rely on multiple input view directions to model the opacity field; (2) Unlike static unbounded scenes, the hand-object motions cause instability in the position and rotation of the 3D Gaussians, resulting in drastic changes in the local geometry. To address these issues, we first generate inward-oriented rays that converge toward the geometric core of the hand-object, enabling systematic surface interrogation. Specifically, we uniformly sample $K = 16$ points $\{\mathbf{o}_k\}_{k=1}^K$ from the surface of a bounding sphere closely encapsulating the hand-object. For each sampled point \mathbf{o}_k , a directed ray \mathbf{r}_k is parametrized as:

$$\mathbf{r}_k(t) = \mathbf{o}_k + t \cdot \frac{\mathbf{c} - \mathbf{o}_k}{\|\mathbf{c} - \mathbf{o}_k\|}, \quad t \geq 0 \quad (4)$$

Next, we adopt a smoother manner to integrate the opacity of each 3D point from different views. Specifically, we enforce hard opacity constraints while maintaining smoothness in visible regions. The conditional branch serves two critical purposes: (1) **Physical Constraint.** Zero opacity is strictly enforced when any viewing ray confirms full transparency, preventing phantom geometry artifacts; (2) **View Consensus.** The mean operation in non-transparent regions smooths out inconsistencies from sparse view sampling. The final opacity field can be defined as:

$$O(\mathbf{x}) = \begin{cases} 0 & \text{if } \exists (\mathbf{o}_k, \mathbf{r}_k) \in \Omega, \\ & O(\mathbf{o}_k, \mathbf{r}_k, t) = 0 \\ \frac{1}{|\Omega|} \sum_{(\mathbf{o}_k, \mathbf{r}_k) \in \Omega} O(\mathbf{o}_k, \mathbf{r}_k, t) & \text{otherwise} \end{cases} \quad (5)$$

Finally, following [10], we use the center and corners of 3D bounding boxes around the 3D Gaussian primitives as vertex sets for the tetrahedral mesh, and utilize the Marching Tetrahedra algorithm [9] for triangle mesh extraction upon assessing the opacity at tetrahedral points.

B Optimization

B.1 Pixel Color Loss

Beyond joint rendering for interaction modeling, we adopt independent color supervision for hands and objects, to address the color bleeding issue in close interaction regions where traditional unified rendering fails to disentangle component-specific appearances:

$$\begin{aligned} C &= \sum_{i \in \mathcal{N}_{\text{ho}}} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \\ C_{\text{hand}} &= \sum_{i \in \mathcal{N}_{\text{ho}}} \mathbf{c}_i^h \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad \mathbf{c}_i^h = \begin{cases} \mathbf{c}_i & i \in \mathcal{N}_{\text{hand}} \\ 0 & \text{otherwise} \end{cases} \\ C_{\text{obj}} &= \sum_{i \in \mathcal{N}_{\text{ho}}} \mathbf{c}_i^o \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad \mathbf{c}_i^o = \begin{cases} \mathbf{c}_i & i \in \mathcal{N}_{\text{obj}} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (6)$$

where $\mathcal{N}_{\text{ho}} = \text{DepthSort}(\mathcal{N}_{\text{hand}} \cup \mathcal{N}_{\text{obj}})$ ensures correct occlusion handling across components. The pixel-level RGB loss can be formulated as:

$$\mathcal{L}_{\text{rgb}} = \|C - C^{\text{gt}}\|_1 + \|C_{\text{hand}} - C_{\text{hand}}^{\text{gt}}\|_1 + \|C_{\text{obj}} - C_{\text{obj}}^{\text{gt}}\|_1 \quad (7)$$

Method	CD _o ↓	F _o @5↑	F _o @10↑
Hasson et al. [5]	1.94	0.383	0.642
Grasping Field [6]	2.06	0.392	0.660
AlignSDF [2]	1.83	0.410	0.679
gSDF [1]	1.55	0.437	0.709
Ours (not rotation-aligned)	<u>1.51</u>	<u>0.497</u>	<u>0.730</u>
Ours	0.24	0.785	0.918

Table S.1: Comparison of object reconstruction on DexYCB.

B.2 Mask Loss

To facilitate geometry supervision, we compute the opacity value by accumulating the alpha values, performed separately for hand, object, and jointly hand-object:

$$\begin{aligned}
O &= \sum_{i \in \mathcal{N}_{ho}} \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \\
O_{\text{hand}} &= \sum_{i \in \mathcal{N}_{ho}} \alpha_i^h \prod_{j=1}^{i-1} (1 - \alpha_j), \quad \alpha_i^h = \begin{cases} \alpha_i & i \in \mathcal{N}_{\text{hand}} \\ 0 & \text{otherwise} \end{cases} \\
O_{\text{obj}} &= \sum_{i \in \mathcal{N}_{ho}} \alpha_i^o \prod_{j=1}^{i-1} (1 - \alpha_j), \quad \alpha_i^o = \begin{cases} \alpha_i & i \in \mathcal{N}_{\text{obj}} \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{8}$$

The mask loss measures the L1 distance between the rendered opacity and the corresponding ground truth mask values:

$$\mathcal{L}_{\text{mask}} = \|O - M\|_1 + \|O_{\text{hand}} - M_{\text{hand}}\|_1 + \|O_{\text{obj}} - M_{\text{obj}}\|_1 \tag{9}$$

B.3 Perceptual Loss

To enhance high-frequency detail preservation and mitigate blurring artifacts in synthesized interactions, we extend the independent supervision paradigm to perceptual feature space, optimizing LPIPS as the perceptual loss with AlexNet [7] as the backbone. Following the RGB loss structure, we employ LPIPS metric for decomposed components. The final perceptual loss aggregates multi-component measurements:

$$\begin{aligned}
\mathcal{L}_{\text{perc}} &= \text{LPIPS}(\mathcal{R}(C), \mathcal{R}(C^{\text{gt}})) \\
&\quad + \text{LPIPS}(\mathcal{R}(C_{\text{hand}}), \mathcal{R}(C_{\text{hand}}^{\text{gt}})) \\
&\quad + \text{LPIPS}(\mathcal{R}(C_{\text{obj}}), \mathcal{R}(C_{\text{obj}}^{\text{gt}}))
\end{aligned} \tag{10}$$

where $\mathcal{R}(\cdot)$ denotes the rendering function from alpha-composited colors to RGB images.

B.4 Pose Loss

To ensure physically plausible hand-object interactions, we supervise the 6D pose parameters (rotation $\mathbf{R} \in \text{SO}(3)$, translation $\mathbf{t} \in \mathbb{R}^3$) and the corner positions \mathbf{P}_C of manipulated objects through the SmoothL1 loss:

$$\begin{aligned}
\mathcal{L}_{\text{rot}} &= \text{SmoothL1}(\hat{\mathbf{R}}, \mathbf{R}^{\text{gt}}) \\
\mathcal{L}_{\text{trans}} &= \text{SmoothL1}(\hat{\mathbf{t}}, \mathbf{t}^{\text{gt}}) \\
\mathcal{L}_{\text{corner}} &= \text{SmoothL1}(\hat{\mathbf{P}}_C, \mathbf{P}_C^{\text{gt}})
\end{aligned} \tag{11}$$

The overall pose loss $\mathcal{L}_{\text{pose}}$ is then computed as the weighted sum of these components:

$$\mathcal{L}_{\text{pose}} = \lambda_{\text{rot}} \mathcal{L}_{\text{rot}} + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}} + \lambda_{\text{corner}} \mathcal{L}_{\text{corner}} \tag{12}$$

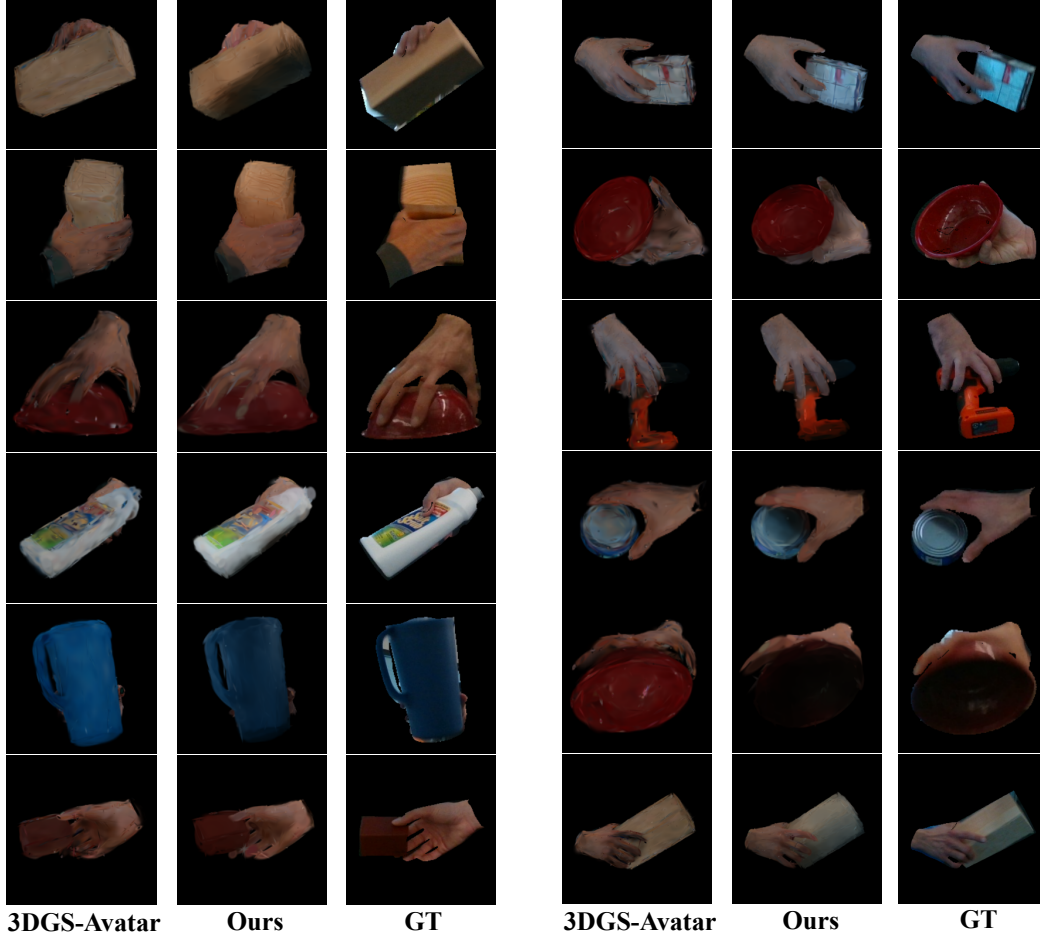


Figure S.1: Qualitative results on DexYCB dataset.

where the weight factors λ_{rot} , λ_{trans} , and λ_{corner} are set as 10, $1e4$, and $1e3$, respectively. In our framework, the ground truth object poses are optional. For example, in comparison on the HO3D dataset, since HOLD [3] does not use any pose annotations, our method does not access the ground truth poses in this experiment either.

B.5 Overall Loss

In addition to the aforementioned loss functions, we further introduce several regularization terms to enhance the robustness and physical plausibility of our framework. Specifically, we follow [8] to incorporate a Skinning Loss $\mathcal{L}_{\text{skin}}$ for regularizing the forward skinning network. To preserve local geometric consistency during deformation, we employ an as-isometric-as-possible constraint, which consists of $\mathcal{L}_{\text{iso-pos}}$ and $\mathcal{L}_{\text{iso-cov}}$ to restrict neighboring 3D Gaussian centers and covariance matrices, ensuring they maintain similar distances after deformation. Furthermore, we treat the centers of hand and object Gaussians as hand-object point clouds, and following [5], apply distance-based penetration loss \mathcal{L}_{pen} and contact loss $\mathcal{L}_{\text{cont}}$ between them to enforce physically plausible contact and prevent interpenetration. The overall loss can be formulated as:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} \\ & + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{skin}}\mathcal{L}_{\text{skin}} \\ & + \lambda_{\text{iso-pos}}\mathcal{L}_{\text{iso-pos}} + \lambda_{\text{iso-cov}}\mathcal{L}_{\text{iso-cov}}, \end{aligned} \quad (13)$$

where λ_{rgb} , λ_{mask} , λ_{perc} , λ_{pose} , λ_{cont} , λ_{pen} , λ_{skin} , $\lambda_{\text{iso-pos}}$, and $\lambda_{\text{iso-cov}}$ are set to 1, 0.1, 0.01, 1, 20, 10, 0.1, 1, and 100, respectively.

C Discussion of Mesh Alignment

Existing SDF-based methods [2, 6, 1] typically reconstruct object meshes within a unit bounding box with non-uniform scaling. To evaluate the pose-independent reconstruction performance, these methods perform ICP alignment with translation and scaling between the reconstructed mesh and the ground truth mesh, while this alignment strategy, which aims to solve the inconsistency of scale, is not applicable to our 6DoF pose-driven reconstruction process. Therefore, for the object branch, we adopt the ICP alignment with translation, scaling, and rotation in the main comparison. Nevertheless, we compare various alignment strategies in Table S.1, where our method outperforms the SDF-based approach under both alignment strategies.

D Qualitative Results

We present more qualitative comparisons of hand-object rendering on DexYCB in Fig. S.1. The rendering results of our method show more delicate colors and more robust hand-object poses, and can adapt to the light and shadow patterns between hand and object.

E Societal Impacts

This paper proposes a fine-grained hand-object interaction modeling framework from monocular RGB images, with the positive impact of broadening access to immersive technologies in AR/VR, HCI, and robotics applications. A possible negative impact may arise from privacy and fairness concerns due to implicit data collection and dataset bias, necessitating cautious and transparent deployment in real-world applications.

References

- [1] 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.
- [2] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 231–248. Springer, 2022.
- [3] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, 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*, pages 494–504, 2024.
- [4] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 2024.
- [5] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019.
- [6] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020.
- [7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [8] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024.
- [9] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.
- [10] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics*, 2024.