Towards Generalized 3D Reconstruction in Hand-Object Interaction

Kyeonghwan Gwak

UNIST
khgwak@unist.ac.kr

Abstract. The fundamental goal of 3D reconstruction in Hand-Object Interaction (HOI) is to recover both the geometry and appearance of the hand and the manipulated object from visual observations. While traditional methods rely on category-specific priors, recent advancements aim for category-agnostic reconstruction. In this report, we review the evolution of 3D reconstruction in HOI, starting with the theoretical foundations essential for understanding these methodologies and tracing how the field has advanced towards generalized solutions. In particular, we focus our technical review on two representative frameworks: HOLD and BIGS.

Keywords: 3D Reconstruction · Hand-Object Interaction · Signed Distance Function · 3D Gaussian Splatting

1 Introduction

3D reconstruction in Hand-Object Interaction (HOI) plays a fundamental role in immersive applications, ranging from Augmented and Virtual Reality (AR/VR) to intuitive Human-Computer Interaction (HCI). However, this task presents a unique challenge arising from the interaction itself: *mutual occlusion*. The hand dynamically occludes the object, while the object simultaneously blocks the view of the hand, complicating the recovery for both entities. To address this, various reconstruction methodologies have evolved, each proposing distinct mechanisms to resolve these occlusion ambiguities.

While earlier approaches relied on pre-scanned templates to compensate for such missing information, real-world scenarios necessitate *category-agnostic* capabilities—reconstructing arbitrary objects without relying on pre-scanned templates. This report provides a bottom-up survey organized as follows: We first establish the theoretical foundations in Sec. 2. Then, we review the progression from isolated hand reconstruction in Sec. 3 to joint hand-object systems in Sec. 4. Finally, in Sec. 5, we analyze state-of-the-art category-agnostic methods, specifically HOLD [8] and BIGS [14].

2 Preliminaries

This section introduces the core concepts of 3D representation and parametric hand modeling that serve as the building blocks for modern reconstruction methods.

2.1 3D Representation Methods

We categorize 3D representations into two primary streams: Explicit and Implicit.

Explicit Representations: Point Cloud, Mesh, and 3DGS. Explicit representations describe the scene using discrete geometric primitives.

- **Point Cloud:** A point cloud is the simplest representation, defined as an unordered set of points $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$. However, point clouds lack topological connectivity and do not define a continuous surface, limiting their ability to model solid geometry.
- **Mesh:** A mesh explicitly defines the surface topology via a graph structure $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$, consisting of vertices, edges, and faces. Unlike point clouds, meshes create a continuous surface, enabling the representation of solid geometry and texture. Crucially, they serve as the standard format for parametric models like MANO [18].
- **3D Gaussian Splatting (3DGS):** 3DGS [10] represents the scene as a collection of 3D Gaussians. The influence of a Gaussian at a query point $\mathbf{x} \in \mathbb{R}^3$ is defined as:

$$G(\mathbf{x}) = o \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{1}$$

  where $\mu \in \mathbb{R}^3$ is the mean position and $o \in [0,1]$ is the opacity. To facilitate convergence, the set of Gaussians is initialized using a sparse point cloud derived from Structure-from-Motion (SfM) [20]. To ensure the covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$ remains valid (positive semi-definite) during optimization, it is decomposed into a rotation matrix R and a scaling matrix S:

$$\Sigma = RSS^TR^T \tag{2}$$

  Here, $S \in \mathbb{R}^{3\times3}$ is a diagonal scaling matrix and $R \in SO(3)$ is the rotation matrix derived from a normalized quaternion $\mathbf{q} \in \mathbb{R}^4$.

  For rendering, these 3D Gaussians are projected onto the image plane. The resulting 2D covariance matrix $\Sigma^{2D} \in \mathbb{R}^{2\times2}$ is approximated using the Jacobian of the projective transformation $J \in \mathbb{R}^{2\times3}$ and the viewing transformation matrix $W \in \mathbb{R}^{3\times3}$:

$$\Sigma^{2D} = JW\Sigma W^TJ^T \tag{3}$$

Finally, the pixel color $C \in \mathbb{R}^3$ is computed via α-blending of $\mathcal{N}$ Gaussians sorted by depth (front-to-back):

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{4}$$

where $c_i \in \mathbb{R}^3$ is evaluated from the spherical harmonics based on the viewing direction derived from W, and $\alpha_i \in [0, 1]$ is obtained by multiplying the learned opacity parameter o with the 2D Gaussian probability density at the specific pixel coordinate.

The efficient tile-based rasterization pipeline enables real-time rendering, while its differentiable nature allows for end-to-end optimization of the Gaussian parameters. Furthermore, the explicit nature of the representation facilitates deformation, establishing 3DGS as the foundational framework for BIGS [14].

Implicit Representations: SDF and NeRF. Implicit methods define the scene as a continuous field learned by a neural network. Instead of storing discrete primitives, the network predicts geometric properties (*e.g.*, distance or density) for any query point in space.

- **Signed Distance Function (SDF):** An SDF maps a spatial query point $\mathbf{x} \in \mathbb{R}^3$ to a scalar value $f(\mathbf{x}) \in \mathbb{R}$, representing the signed distance to the nearest surface $\mathcal{S} \subset \mathbb{R}^3$:

$$f(\mathbf{x}) = s(\mathbf{x}) \cdot \min_{\mathbf{y} \in \mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2 \tag{5}$$

  where $s(\mathbf{x}) \in \{-1, 1\}$ is the sign function indicating whether $\mathbf{x}$ is inside (negative) or outside (positive). Consequently, the object surface is implicitly represented as the zero-level set of this function, defined as $\{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = 0\}$. Crucially, this continuous representation can model arbitrary topologies. This flexibility is key to achieving category-agnostic reconstruction, serving as the foundational framework for HOLD [8].
- **Neural Radiance Field (NeRF):** NeRF approaches scene representation by optimizing a continuous volumetric function (see Fig. 2 in Mildenhall *et al.* [13]), approximated by a Multilayer Perceptron (MLP). Mathematically, this network approximates a mapping F:

$$F : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma) \tag{6}$$

  where the input consists of a 3D coordinate $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{S}^2$, and the output includes the emitted color $\mathbf{c} \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}_{\geq 0}$.

An image pixel is rendered by integrating along a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt \tag{7}$$

Here, $C(\mathbf{r}) \in \mathbb{R}^3$ is the final predicted pixel color, and t_n, t_f denote the near and far bounds of the ray. $T(t)$ denotes the accumulated transmittance, defined as:

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\, ds\right) \tag{8}$$

which quantifies the probability that the ray travels from t_n to t without hitting any particles. In practice, this integral is numerically approximated via stratified sampling.

While NeRF achieves photorealistic rendering, this requirement for dense network evaluations along every ray results in high computational costs, contrasting with the real-time capabilities of explicit methods like 3DGS.

2.2 Parametric Hand Model

To obtain a robust and controllable 3D hand representation, the MANO model [18] is widely utilized in the literature. Constructed from a large-scale dataset of 3D hand scans, MANO captures the statistical variations of human hands using Principal Component Analysis (PCA). It maps low-dimensional shape parameters $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{16\times3}$ to a high-dimensional hand mesh $M_h \in \mathbb{R}^{778\times3}$ and 3D joint locations $J \in \mathbb{R}^{21\times3}$ via Linear Blend Skinning (LBS):

$$M_h(\beta, \theta) = W(T_P(\beta, \theta), J(\beta), \theta, \mathcal{W}) \tag{9}$$

Here, the joint locations J are linearly regressed from the shaped mesh vertices. W denotes the skinning function (LBS), which deforms the template mesh T_P according to the pose θ and the adapted joints J, utilizing fixed skinning weights $\mathcal{W}$.

Figures 6 and 7 in Romero *et al.* [18] show scanned hands alongside their corresponding MANO mesh representations, demonstrating the template's capability to capture diverse hand shapes and poses across different subjects.

The key advantage of this representation is that it reduces the complex problem of 3D hand reconstruction to the regression of compact parameters β and θ. By leveraging the learned statistical knowledge, biologically plausible dense mesh and joint locations can be recovered even from partial observations.

3 3D Hand Reconstruction

The most fundamental challenge in this domain is 3D hand reconstruction. This task aims to recover two primary attributes from a single input image I: *shape* and *pose*.

Mathematically, the reconstruction process can be formulated as a mapping function Φ_H that estimates the mesh vertices $M \in \mathbb{R}^{V\times 3}$ and the 3D joint locations $J \in \mathbb{R}^{K\times 3}$ from the image:

$$\Phi_H(I) \rightarrow (\hat{M}, \hat{J}) \tag{10}$$

where V denotes the number of mesh vertices representing the hand shape, and K denotes the number of joints representing the hand pose.

To address this task, contemporary approaches typically employ an encoder-decoder architecture. Powerful backbones, such as Vision Transformers (ViT) [7], are utilized to encode the input image into latent features, which are subsequently decoded to recover the target mesh vertices M and joint locations J. While sharing this general structure, different approaches utilize different encoding and decoding strategies. Methods like METRO [4,5,11,12,21] focus on *direct regression*, where the network directly predicts the 3D coordinates of the mesh vertices M from the latent representations. Subsequently, the 3D joint locations J are obtained from these vertices using the standard joint regressor $\mathcal{J}_{reg}$, formulated as $J = \mathcal{J}_{reg}M$. In contrast, methods like HaMeR [3,6,15,17,19] adopt a *parametric regression* strategy; they first estimate the MANO parameters (β, θ) and subsequently derive the dense mesh and joints via the differentiable MANO layer (see Fig. 2 in Pavlakos *et al.* [15]).

To quantitatively evaluate the reconstruction quality, two standard metrics are employed: MPJPE for the pose and MPVPE for the shape.

MPJPE (Mean Per-Joint Position Error) assesses the skeletal pose accuracy by calculating the average Euclidean distance between the estimated joints $\hat{J}$ and the ground truth joints J_{gt}:

$$\text{MPJPE} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{J}_k - J_{gt,k}\|_2 \tag{11}$$

MPVPE (Mean Per-Vertex Position Error) assesses the mesh surface quality by calculating the average Euclidean distance between the estimated vertices $\hat{M}$ and the ground truth vertices M_{gt}:

$$\text{MPVPE} = \frac{1}{V}\sum_{i=1}^{V} \|\hat{M}_i - M_{gt,i}\|_2 \tag{12}$$

In standard benchmarks, these metrics are often computed after Procrustes Analysis (PA). PA aligns the predicted structure to the ground truth by adjusting its rotation, translation, and scale. The resulting metrics, PA-MPJPE and PA-MPVPE, focus purely on the reconstruction quality by excluding errors from global misalignment and size differences.

4 3D Hand-Object Reconstruction

In daily life, hands are rarely observed in static isolation; their most natural state is dynamically interacting with the physical world. Consequently, the scope of 3D reconstruction naturally extends from isolated hands to the joint recovery of the hand and the object it manipulates.

Mathematically, this task aims to map a single input image I to the hand components (mesh M_h, joints J) and the object mesh M_{obj} simultaneously:

$$\Phi_{HOI}(I) \rightarrow (\hat{M}_h, \hat{J}, \hat{M}_{obj}) \tag{13}$$

To address this task, RHO [1] was proposed for rigid object reconstruction in in-the-wild scenarios. While RHO utilizes a template mesh, the variable nature of unconstrained in-the-wild environments necessitates estimating the object scale $s \in \mathbb{R}$ in addition to the standard 6D pose (rotation $R \in SO(3)$ and translation $T \in \mathbb{R}^3$). The reconstruction is achieved by transforming the template $\mathcal{T}_{obj}$ as follows:

$$\hat{M}_{obj} = s \cdot R \cdot \mathcal{T}_{obj} + T \tag{14}$$

Moving beyond rigid bodies, ARCTIC [9] addresses a more complex category: *articulated objects* (*e.g.*, scissors, laptops). Since ARCTIC operates with pre-defined, precisely scanned object models, scale estimation is omitted. Instead, the focus shifts to modeling the object's articulation. The framework predicts an articulation parameter $\omega \in \mathbb{R}$ (*i.e.*, rotation angle) along a specific axis to dynamically deform the template:

$$\hat{M}_{obj} = R \cdot \mathcal{T}_{obj}(\omega) + T \tag{15}$$

Figure 4 in Fan *et al.* [9] shows the ArcticNet-SF architecture. The network extracts features from an input image via a CNN backbone. It employs separate decoders to estimate the MANO parameters $\Theta = \{\theta, \beta\}$ and translation T for both hands. Simultaneously, it predicts the object state $\Omega \in \mathbb{R}^7$, which consists of the 1D articulation angle $\omega \in \mathbb{R}$ and the 6D rigid pose (rotation $R_o \in \mathbb{R}^3$ and translation $T_o \in \mathbb{R}^3$). Figure 1 in the supplementary material of Fan *et al.* [9] shows the 11 articulated objects included in the dataset. These objects were pre-scanned to obtain high-fidelity template meshes.

While leveraging strong geometric priors enables these methods to maintain high precision even under severe mutual occlusion, they inherently rely on the availability of pre-scanned object meshes, limiting generalization to unseen objects in the wild.

To quantitatively evaluate the quality of the reconstructed object, the Chamfer Distance (CD) is commonly employed. CD measures the geometric similarity between the estimated object mesh vertices $\hat{M}_{obj}$ and the ground truth vertices M_{gt}. It is defined as the average distance from each point in one set to its nearest neighbor in the other set:

$$\text{CD} = \frac{1}{|\hat{M}_{obj}|} \sum_{x \in \hat{M}_{obj}} \min_{y \in M_{gt}} \|x - y\|_2^2 + \frac{1}{|M_{gt}|} \sum_{y \in M_{gt}} \min_{x \in \hat{M}_{obj}} \|y - x\|_2^2 \tag{16}$$

A lower Chamfer Distance indicates that the reconstructed surface is geometrically closer to the ground truth.

5 Category-Agnostic 3D Hand-Object Reconstruction

Despite the high precision achieved by template-based methods, their reliance on pre-scanned object models fundamentally limits their scalability in in-the-wild scenarios where object geometries are unknown. Consequently, the field has shifted towards *category-agnostic* reconstruction, which aims to recover the 3D geometry and appearance of arbitrary objects from visual observations, without prior geometric knowledge. This transition requires solving a more complex challenge: simultaneously inferring the shape of an unseen object while resolving the severe mutual occlusions inherent in interaction.

To address this challenge in single-hand scenarios, HOLD [8] proposes a compositional implicit neural framework. A key advantage of this approach lies in leveraging a Signed Distance Function (SDF) for object representation. This continuous field allows for modeling arbitrary shapes, fundamentally enabling the category-agnostic reconstruction of unseen objects.

Technically, HOLD integrates separate volumetric fields for the hand (f_h), the object (f_o), and the background (f_b) (see Fig. 3 in Fan *et al.* [8]). To align the dynamic observation space with the canonical implicit representations, the framework employs specific coordinate transformations: Inverse Linear Blend Skinning (LBS) for the articulated hand and rigid transformations for the object. These elements are then composited via a NeRF-inspired differentiable volumetric rendering pipeline. By transforming signed distances into volume densities and optimizing for photometric consistency, the framework recovers high-fidelity geometry and texture from monocular video, while simultaneously refining poses via interaction constraints.

BIGS [14] extends this capability to bimanual interaction scenarios. It employs 3D Gaussian Splatting (3DGS) to flexibly represent the geometry of arbitrary objects in a category-agnostic manner, while significantly improving optimization and rendering efficiency compared to implicit representations. However, since 3DGS optimization fundamentally relies on minimizing the photometric error against input images, it inherently fails to reconstruct regions that are occluded or unobserved in the input sequence due to the lack of supervision. To bridge this gap, BIGS incorporates a generative prior via Score Distillation Sampling (SDS) loss [16]. This loss enforces object images rendered from random viewpoints—including perspectives completely unobserved in the input—to align with plausible images generated by a pre-trained 2D diffusion model, thereby reliably recovering the geometry and appearance of unseen regions. For hand reconstruction, BIGS initializes 3D Gaussians from MANO vertices, leveraging prior knowledge to ensure rapid convergence. A notable strategy is the optimization of a single shared canonical Gaussian set, treating the left hand as a mirrored

version of the right. By mapping both hands to a unified coordinate space, the framework accumulates visual cues from both instances, allowing visible features from one hand to effectively compensate for occlusions in the other.

The BIGS pipeline is built upon an explicit 3D Gaussian Splatting representation (see Fig. 2 in On *et al*. [14]). The process begins with *Pre-processing*, where coarse meshes derived from a MANO regressor and an object reconstructor are used to initialize the canonical Gaussians ($\mathcal{G}_H, \mathcal{G}_O$). The core architecture employs TriplaneNets [2] ($\mathcal{T}^H, \mathcal{T}^O$) combined with MLPs to decode Gaussian parameters (*e.g.*, opacity, color) and Linear Blend Skinning (LBS) weights. These weights are subsequently utilized to deform the canonical hand Gaussians into the target pose. The optimization proceeds in two stages: a *Single-subject optimization* first refines each shape individually—utilizing SDS loss to recover occluded regions—followed by an *Interacting-subjects optimization* that fine-tunes spatial alignment via contact regularization to ensure physical plausibility.

6 Conclusion

In this report, we have surveyed the evolution of 3D reconstruction in Hand-Object Interaction, tracing the trajectory from isolated hand recovery to complex, category-agnostic joint reconstruction. The progression has been fundamentally driven by the shift in underlying 3D representations. While early methods relied on pre-scanned templates to resolve ambiguities, recent advancements leverage the strengths of neural representations to handle unknown objects. We highlighted how HOLD exploits the continuity of implicit SDFs to ensure geometric consistency in single-hand grasping, whereas BIGS leverages the efficiency of explicit 3D Gaussians combined with diffusion priors to tackle the complex dynamics of bimanual interaction.

Despite these significant advancements, a critical limitation remains: current category-agnostic approaches predominantly assume that the manipulated object is *rigid*. Specifically, these methods optimize a single canonical geometry shared across the entire sequence of input frames, allowing only for rigid pose transformations per timestep. However, real-world interactions frequently involve deformable (*e.g.*, squeezing a sponge) or articulated (*e.g.*, using scissors) objects, where the intrinsic shape changes dynamically. Therefore, a promising direction for future research lies in extending category-agnostic reconstruction to *non-rigid* scenarios. Developing systems capable of recovering time-varying geometries and topological changes without relying on category-specific priors represents the next milestone in achieving truly generalized 3D reconstruction.

References

1. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV (2021)
2. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022)

3. Chen, R., Zhuo, L., Yang, L., WANG, Q., Bo, L., Zhang, B., Yao, A.: ExtPose: Robust and coherent pose estimation by extending ViTs. In: ICML (2025)
4. Chen, X., Song, Z., Jiang, X., Hu, Y., Yu, J., Zhang, L.: HandOS: 3D hand reconstruction in one stage. In: CVPR (2025)
5. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In: ECCV (2022)
6. Dong, H., Chharia, A., Gou, W., Vicente Carrasco, F., De la Torre, F.D.: Hamba: Single-view 3D hand reconstruction with graph-guided bi-scanning mamba. In: NeurIPS (2024)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
8. Fan, Z., Parelli, M., Kadoglou, M.E., Kocabas, M., Chen, X., Black, M.J., Hilliges, O.: HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In: CVPR (2024)
9. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR (2023)
10. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) (2023)
11. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021)
12. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: ICCV (2021)
13. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
14. On, J., Gwak, K., Kang, G., Cha, J., Hwang, S., Hwang, H., Baek, S.: BIGS: Bimanual category-agnostic interaction reconstruction from monocular videos via 3D gaussian splatting. In: CVPR (2025)
15. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: CVPR (2024)
16. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: DreamFusion: Text-to-3D using 2d diffusion. In: ICLR (2023)
17. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In: CVPR (2025)
18. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG) (2017)
19. Rong, Y., Shiratori, T., Joo, H.: FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In: ICCV Workshops (2021)
20. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
21. Zhou, Z., zhou, S., Lv, Z., Zou, M., Tang, Y., Liang, J.: A simple baseline for efficient hand mesh reconstruction. In: CVPR (2024)