MagicHOI: Leveraging 3D Priors for Accurate Hand-object Reconstruction from Short Monocular Video Clips

Shibo Wang[1] Haonan He[1] Maria Parelli[3,†] Christoph Gebhardt[1] Zicong Fan[3,4] Jie Song[1,2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]ETH Zürich, Switzerland [4]Max Planck Institute for Intelligent Systems, Tübingen, Germany

2025 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3315-8775-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCV51701.2025.00563

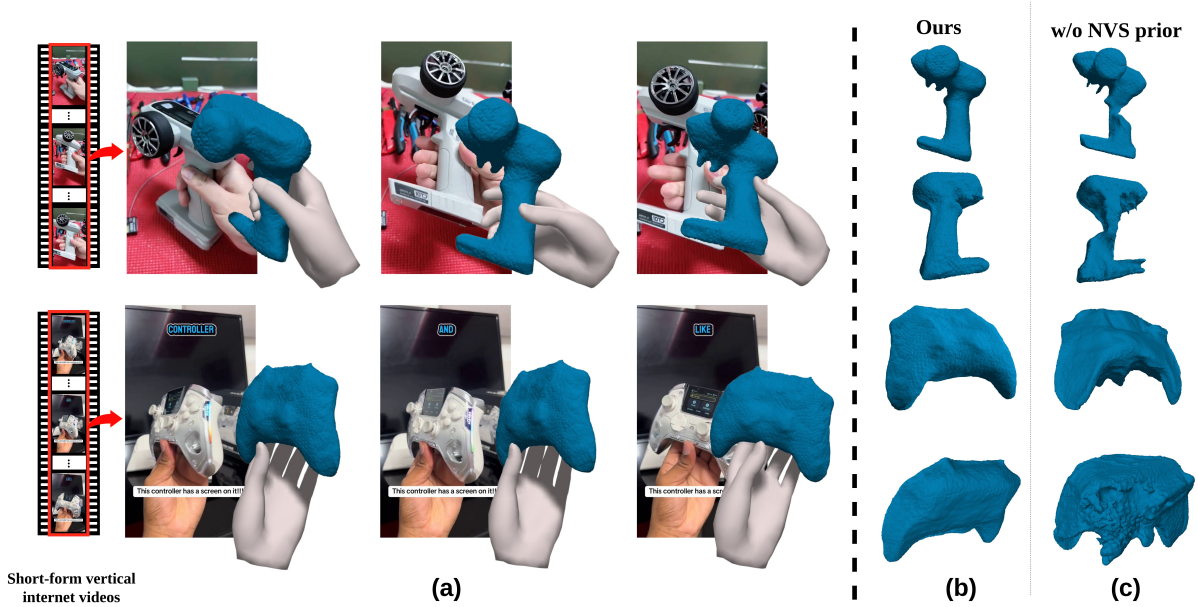

Figure 1. **MagicHOI:** Given a short-form monocular video capturing hand-object interaction, our method reconstructs high-quality 3D hand and object surfaces, including occluded object regions caused by hand interaction or object self-occlusion. **(a)** Input images and corresponding reconstructed surfaces of the hand and object. **(b, c)** Comparison of our method with and without a novel view synthesis (NVS) prior from the object's front and back views, demonstrating improved reconstruction in occluded areas. Best viewed in color.

Abstract

Most RGB-based hand-object reconstruction methods rely on object templates, while template-free methods typically assume full object visibility. This assumption often breaks in real-world settings, where fixed camera viewpoints and static grips leave parts of the object unobserved, resulting in implausible reconstructions. To overcome this, we present MagicHOI, a method for reconstructing hands and objects from short monocular interaction videos, even under limited viewpoint variation. Our key insight is that, despite the scarcity of paired 3D hand-object data, large-scale novel view synthesis diffusion models offer rich object supervision. This supervision serves as a prior to regularize unseen object regions during hand interactions. Leveraging this insight, we integrate a novel view synthesis model into our hand-object reconstruction framework. We further align hand to object by incorporating visible contact constraints. Our results demonstrate that MagicHOI significantly outperforms existing state-of-the-art hand-object reconstruction methods. We also show that novel view synthesis diffusion priors effectively regularize unseen object regions, enhancing 3D hand-object reconstruction.

† Prior to joining University of Tübingen and Tübingen AI Center

2380-7504/25/$31.00 ©2025 IEEE
DOI 10.1109/ICCV51701.2025.00563

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

1. Introduction

Humans interact effortlessly with a wide variety of objects every day, often without conscious thought. These complex interactions are frequently captured in casual videos, which are abundant on the internet (e.g., product reviews on TikTok, Instagram Reels, and YouTube Shorts). Such videos provide a valuable resource for scaling human demonstrations in robotic grasping [8, 88, 89]. A key challenge in reconstructing 3D hand-object interactions at scale is that most internet videos do not offer complete, dense views of the objects due to hand-induced occlusions and object self-occlusions, particularly when the objects are not fully rotated. This incomplete observation hinders the accurate reconstruction of hands and objects (see Figure 1).

To address this challenge, we introduce the task of *Partially-visible Hand-object Reconstruction*, which aims to reconstruct 3D hand and object geometries from videos where the objects are not fully visible from all angles. Specifically, we focus on short video clips of interactions, as these often contain under-observed regions of the objects (e.g., the action of picking up a cup of coffee captured from a static monocular view).

Existing methods for hand-object reconstruction typically rely on pre-scanned 3D object templates [3, 9, 15, 40, 71, 79], which limit their scalability in real-world scenarios. While emerging learning-based approaches are available, most depend on supervised training using a limited number of paired 3D hand-object annotations, resulting in poor generalization [23, 31, 82]. Recently, methods leveraging volumetric rendering have shown promise in generalizing to unseen objects [16, 21, 29]. However, these approaches rely heavily on RGB supervision and struggle to accurately reconstruct hands and objects when parts of the object are occluded or only partially visible (see Figure 4).

In this paper, we introduce MagicHOI, a method designed to address restricted viewpoints and significant occlusions in realistic hand-object interaction scenarios. MagicHOI takes as input a monocular video and reconstructs the 3D surfaces of both the hand and the object, even in short video sequences with partially visible object regions. Our key insight is that, although 3D hand-object annotations are limited, foundational models such as novel view synthesis (NVS) diffusion models provide abundant object supervision [39]. This external supervision can be utilized as a prior to help regularize unseen or occluded object regions during hand-object interactions.

In detail, MagicHOI first employs structure-from-motion (SfM) to obtain the initial camera poses of the object from the input video. Once these observed views are calibrated through SfM, we utilize volumetric rendering to reconstruct the object geometry using an implicit signed distance field (SDF) [49, 80]. To regularize the reconstruction in unobserved regions, we sample novel views and apply a score-distillation sampling loss using an NVS diffusion model [39]. However, randomly sampling novel views to enforce a prior may negatively impact the reconstruction quality of object regions that are visible. To address this, we introduce a visibility-aware weighting strategy. Finally, MagicHOI aligns the hand and the object to the same space while incorporating hand-object interaction constraints.

We qualitatively and quantitatively demonstrate that our visibility-aware weighting strategy, combined with the integration of an NVS diffusion prior, significantly enhance the hand-object reconstruction quality in the partially-visible hand-object reconstruction setting, surpassing the performance of the state-of-the-art methods. To highlight the robustness of MagicHOI in challenging real-world scenarios, we also illustrate its ability to faithfully reconstruct short-form videos sourced from online content.

In summary, our contributions are as follows: 1) We present MagicHOI, a holistic framework that effectively leverages an NVS diffusion prior for accurate 3D hand-object reconstruction from short monocular video sequences, even in the presence of unobserved parts of the hand and object; 2) We demonstrate that the NVS model can be utilized to regularize template-free hand-object reconstruction; 3) We introduce a novel visibility-aware weighted sampling strategy to balance the regularization between the observed and unobserved object regions using this prior; 4) We evaluate our method both qualitatively and quantitatively, demonstrating its superior performance compared to the state-of-the-art methods and achieving realistic reconstruction results. Our project website can be found here: `byran-wang.github.io/MagicHOI`.

2. Related Work

Hand pose and shape recovery: The task of reconstructing hands in 3D is a long-standing problem in computer vision, where most methods focus on single-hand reconstruction [1, 5, 13, 17, 30, 35, 46, 48, 62–65, 75, 87, 90, 93, 94]. In particular, Zimmermann *et al*. [94] introduce the first convolutional networks for hand pose estimation. Spurr *et al*. [64] present bio-mechanical constraints allowing for semi-supervised learning on in-the-wild images. Contrastive learning objectives are also explored for learning hands in action on images without labels [65, 93]. With the introduction of the InterHand2.6M dataset by Moon *et al*. [46], the interacting hand pose estimation research has exploded [13, 19, 32, 34, 35, 44–47, 50, 74]. For example, Tse *et al*. [74] introduced a spectral graph transformer network for estimating the surfaces of two strongly interacting hands. Pavlakos *et al*. [52] train a vision transformer to achieve robust monocular hand pose estimation under unconstrained conditions. Yu *et al*. [86] propose Dyn-HaMR, which reconstructs interacting hand poses in a global frame by combining SLAM with a learned hand motion prior [12]

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

for spatial alignment. In contrast to these approaches, we focus on joint hand-object 3D reconstruction.

Template-based hand-object reconstruction: Reconstructing hand-object interaction is challenging and has gained increasing attention in recent years [9, 18, 23–25, 38, 40, 71, 73, 79, 79, 92]. Most existing methods assume a known object template and estimate only hand and object poses [3, 9, 15, 40, 71, 79]. For example, Yang *et al*. [79] introduce a contact potential field to improve hand-object contact. Liu *et al*. [40] propose a transformer-based contextual reasoning module that captures the synergy between hand and object features, yielding stronger responses at contact regions. Zhou *et al*. [91] learn an interaction motion prior to refine noisy motion estimates from a single-frame hand-object reconstruction approach. Fan *et al*. [14, 15] introduce the first approach to jointly reconstruct two dexterous hands and articulated objects from RGB inputs. Despite accurate pose estimation, these methods struggle with novel objects and in-the-wild videos due to dependence on known templates.

Template-free hand-object reconstruction: Recent advancements have introduced more generalizable techniques [4, 6, 16, 23, 29, 55, 66, 78, 81–84], incorporating differentiable rendering, data-driven priors, and compositional implicit representations. However, these methods often impose constraints such as requiring rigid hand-object interactions [29, 54], multi-view observations [55], category-level hand-object supervision [6, 23, 78, 81–84], or complete object observations [16]. Consequently, they struggle with occlusions, leading to incomplete object surfaces in cases of hand-induced or self-occlusions. In contrast, our approach integrates 3D priors with geometry information using a visibility-aware weighting strategy, enhancing object reconstruction quality in both occluded and observed regions, even in extremely short video clips.

3D object reconstruction: Reconstructing objects from images has been a long-standing challenge in computer vision. Structure from Motion (SfM) [22, 59, 67, 72] has remained a robust solution, particularly for multiview reconstruction and camera parameter estimation. However, traditional SfM methods are limited to scenarios where the object is either fully visible or nearly unoccluded in the input views. To address the reconstruction of occluded object shapes, many research efforts have turned to learning-based approaches. These approaches leverage large-scale 3D object datasets, such as Objaverse [11] and Objaverse-XL [10], and include methods like regression [36], retrieval [69], diffusion models [37, 39, 41, 43, 53, 61], and large-scale reconstruction models [28, 68, 76, 77]. Despite progress, standard object reconstruction methods underperform with hand-object interaction images due to hand occlusions. Unlike these, MagicHOI jointly reconstructs hands and objects from videos.

3. Method: MagicHOI

Figure 2 provides an overview of our approach, MagicHOI, which reconstructs hand–object interactions from a RGB video captured with limited viewpoints. We first initialize the hand and object poses for every frame (Sec. 3.1). Next, we integrate a novel view synthesis (NVS) model into the pipeline and align its coordinate frame with that of the object (Sec. 3.2). With the NVS model in place, we optimize an implicit signed distance field (SDF) that regularizes the reconstruction of object regions that are not directly observed (Sec. 3.3). Finally, we align the hand to the object using visible hand-object contact constraints (Sec. 3.4).

3.1. Initialization

Hand initialization: For each image $I^{\text{obs}} \in \mathbb{R}^{3\times H\times W}$ from a short video clip, we leverage an off-the-shelf hand pose estimator [51] to obtain MANO [56] parameters, which include the hand pose $\theta \in \mathbb{R}^{45}$, shape $\beta \in \mathbb{R}^{10}$, global rotation $\mathbf{R}_h \in \mathrm{SO}(3)$, and translation $\mathbf{t}_h \in \mathbb{R}^3$.

Object initialization: Following [16], we apply HLoc [57, 58] to perform structure-from-motion (SfM), obtaining the camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$, rotation $\mathbf{R}^{\text{SfM}} \in \mathrm{SO}(3)$, and translation $\mathbf{t}^{\text{SfM}} \in \mathbb{R}^3$. For each frame, we additionally generate a depth map $\mathbf{D} \in \mathbb{R}^{H\times W}$ using multi-view stereo (MVS) [26, 60]. These steps are applied to images segmented by the off-the-shelf model Cutie [7]. In fact, advances in learning-based pose estimation, such as MASt3R [33], can provide more robust object pose estimates without relying on SfM convergence.

3.2. Novel view synthesis model

A key contribution of MagicHOI is leveraging an NVS model Zero-1-to-3 to reconstruct objects with hands even in occluded object regions.

Inpainting: Before using the NVS model to supervise our reconstruction, one needs to condition it on a reference image I^{ref} where only the object is visible. To achieve this, we extract an object-only image based on the object segmentation mask and perform inpainting on the hand region to create the object-only reference image I^{ref}. We use InpaintAnything [85] to inpaint the hand region. The inpainted image $\tilde{I}^{\text{ref}}$ is then centered, cropped, and resized to suit the input format of the NVS model I^{ref}. We select the reference frame from the input video based on a simple heuristic. Specifically, we use off-the-shelf segmentation to extract hand and object masks for each frame, compute the object-to-hand pixel ratio, and choose the frame with the highest ratio. This automated selection method prioritizes views where the object is more clearly visible.

Novel view synthesis model: We use Zero-1-to-3 [39] as our view synthesis backbone. In a nutshell, conditioned on an input image I^{ref}, its camera pose $\mathbf{y}^{\text{ref}}$ and a target camera

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

view $\mathbf{y}$, the NVS model $f_{\text{NVS}}(I^{\text{ref}}, \mathbf{y}^{\text{ref}}, \mathbf{y})$ generates an image I^{NVS} for the novel view $\mathbf{y}$ where the view $\mathbf{y} = [\varphi, \phi, \rho]$ represents the camera pose in spherical coordinates, φ denotes elevation, ϕ denotes azimuth, and ρ denotes distance from the object center. The camera intrinsic of the reference image I^{ref} is represented by $\mathbf{K}^{\text{ref}} \in \mathbb{R}^{3\times 3}$, centered on the image plane and characterized by a predefined focal length. See more details in SupMat.

We use the pre-trained Zero-1-to-3 model with frozen weights in our framework. It is trained using a forward diffusion process that progressively corrupts a novel-view image with Gaussian noise, and a denoiser ϵ_ω is then learned to reconstruct the image by reversing this process [27]. Training is performed on large-scale synthetic data rendered from 3D object models. For full architectural and training details, we refer to the original Zero-1-to-3 paper [39].

Space alignment: To leverage both observations from RGB video input and object geometry prior in Zero-1-to-3, we must align our SfM object space to that of the Zero-1-to-3 object space. We use the object space from Zero-1-to-3 as our canonical object space and align the SfM object space to it. To achieve this, we first extract the 2D correspondences $\{\mathbf{p}^{\text{ref}}, \tilde{\mathbf{p}}^{\text{ref}}\}$ between the reference image I^{ref} and the same image but in the original resolution $\tilde{I}^{\text{ref}}$. Since the Zero-1-to-3 model can render the depth map $\mathbf{D}^{\text{ref}}$ of the reference view and the MVS process provides a depth map $\tilde{\mathbf{D}}^{\text{ref}}$, we apply inverse perspective projection to obtain 3D correspondences $\{\mathbf{P}^{\text{ref}}, \tilde{\mathbf{P}}^{\text{ref}}\}$ using the intrinsics from the Zero-1-to-3 model and the SfM estimates respectively.

Finally, we estimate the rigid transformation that brings the SfM object frame into the Zero-1-to-3 frame by solving the rotation $\mathbf{R}_{\text{a}} \in \text{SO}(3)$ and translation $\mathbf{t}_{\text{a}} \in \mathbb{R}^3$. The optimization objective combines a 3-D correspondence term $\mathcal{L}_{3D}$ and a Perspective-n-Point (PnP) term $\mathcal{L}_{2D}$:

$$\mathcal{L}_{\text{a}} = \mathcal{L}_{\text{3D}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}}, \tag{1}$$

where $\lambda_{\text{2D}} = 1.0$. See more details in SupMat.

Using the estimated rigid transformation $(\mathbf{R}_{\text{a}}, \mathbf{t}_{\text{a}})$, we convert each SfM camera pose $(\mathbf{R}^{\text{SfM}}, \mathbf{t}^{\text{SfM}})$ from the SfM frame to the Zero-1-to-3 frame:

$$\mathbf{R} = \mathbf{R}_{\text{a}}\mathbf{R}^{\text{SfM}}, \quad \mathbf{t} = \mathbf{R}_{\text{a}}\mathbf{t}^{\text{SfM}} + \mathbf{t}_{\text{a}}. \tag{2}$$

where $\mathbf{R} \in \text{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$.

3.3. Occlusion-robust object reconstruction

With the help of the NVS backbone, we can synthesize novel views to reveal parts of the object that are occluded in the original observations, thereby enabling occlusion-robust object reconstruction.

We represent the object using an implicit neural signed distance and texture field f_{ψ_o}, which can be volumetrically rendered into RGB images and defined as:

$$f_{\psi_o}(\mathbf{x}) : \mathbb{R}^3 \to \mathbb{R} \times \mathbb{R}^3 \tag{3}$$

where f_{ψ_o} maps a spatial point $\mathbf{x}$ to its signed distance to the object surface and its corresponding texture color. The function is parameterized by learnable parameters ψ_o.

Supervision on the observed object regions: For observed RGB images $\{I^{\text{obs}}\}$, we supervise our object model ψ_o with image evidence, namely RGB images and segmentation masks $\mathcal{L}_{\text{RGB}} + \lambda_{\text{segm}}\mathcal{L}_{\text{segm}}$ where $\mathcal{L}_{\text{RGB}}$ enforces photometric consistency between the rendered and observed object images, $\mathcal{L}_{\text{segm}}$ encourages alignment with the object segmentation mask (see more details in SupMat).

Supervision on the unobserved object regions: To supervise unobserved object regions using the object geometry prior, we adopt the Score Distillation Sampling (SDS) loss [53]. A novel view image I^{NVS} is rendered from the object model f_{ψ_o} using a randomly sampled camera pose $\mathbf{y}$. We sample a diffusion timestep $t \sim \text{Uniform}(\{0, \ldots, T\})$ and obtain the corresponding Gaussian noise ϵ_t. A noisy image $\mathbf{z}_t$ is computed by adding the noise ϵ_t to the rendered image I^{NVS}. Given the known noise ϵ_t from the step t of the forward diffusion process, we supervise the model by minimizing the discrepancy between it and the Zero-1-to-3 denoiser's prediction:

$$\mathcal{L}_{\text{NVS}} = w(t)\left\|\epsilon_t - \epsilon_\omega(\mathbf{z}_t; t, \mathcal{C})\right\|^2 \tag{4}$$

where ϵ_ω is the denoising network of the Zero-1-to-3 model, $\mathcal{C} = \Psi(I^{\text{ref}}, \mathbf{y}^{\text{ref}}, \mathbf{y})$ is the embedding of the condition. The scalar $w(t) \in \mathbb{R}$ is a weighting function predefined by the diffusion model and the step t.

Visibility-aware weighting strategy: To balance image evidence and object geometry prior from Zero-1-to-3, we devise a visibility-aware weighting strategy. In particular, after the object model has been optimized using observed views for some iterations, we derive a coarse 3D visibility grid $O(x, y, z)$ from the underlying 3D object representation, which is a coarse voxel representation around the object. For each pixel of the observed views, we cast a ray from an object pixel into the visibility grid to determine the occluded voxels for novel views. This process then labels a voxel as visible (valued 1) if the voxel is the first intersecting point of the ray. If a voxel is never the first intersected point of all rays, it is labeled as non-visible (valued 0).

For each novel view $\mathbf{y}$, we rasterize the visibility grid O to produce a visibility image $V(u, v)$, where each pixel takes a value from $\{1, 0, -1\}$: 1 indicates the voxel is observed by at least one ray in the observed input images, 0 indicates it has never been observed, and -1 denotes background. From $V(u, v)$, we compute the visibility ratio β for the given view $\mathbf{y}$. In detail, we compute the ratio of pixels in observed regions to those in unobserved regions

$$\beta = \frac{\sum_{(u,v)\in V} \mathbb{I}[V(u, v) = 1]}{\sum_{(u,v)\in V} \mathbb{I}[V(u, v) = 0]} \tag{5}$$

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

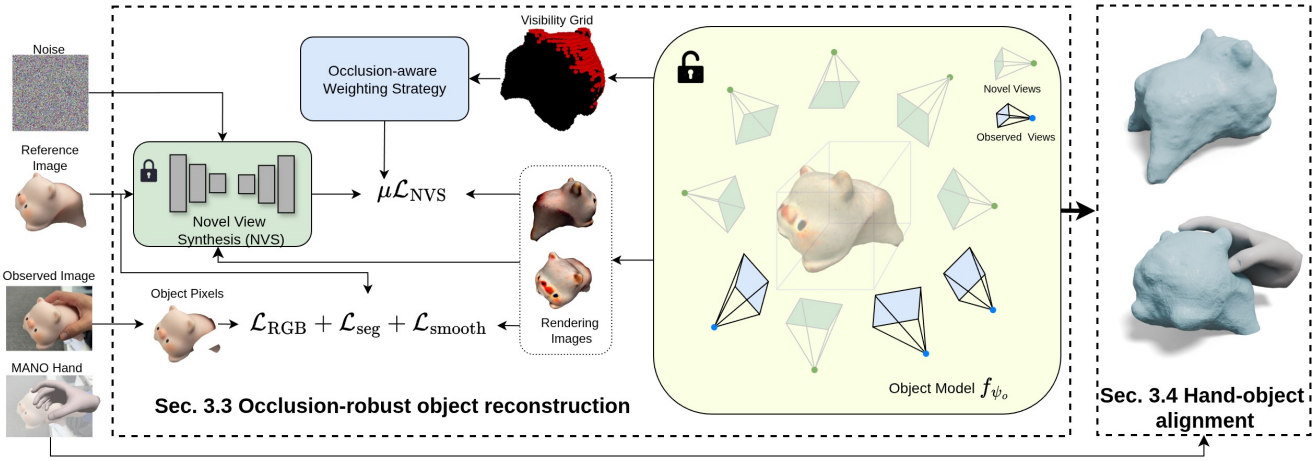

Figure 2. **Method overview**. MagicHOI learns an implicit object field f_{ψ_o} that reconstructs complete object geometry from a short video with partial visibility. We introduce a novel view synthesis (NVS) loss $\mathcal{L}_{\text{NVS}}$ that leverages an NVS diffusion prior to reconstruct unobserved object parts. Specifically, we render a novel view and take one denoising step of the NVS model; $\mathcal{L}_{\text{NVS}}$ guides f_{ψ_o} toward high-density regions to complete occluded areas. Updates from novel views are combined with RGB and segmentation supervision $\mathcal{L}_{\text{RGB}}$, $\mathcal{L}_{\text{seg}}$ and a smoothness regularizer $\mathcal{L}_{\text{smooth}}$ for both observed and reference images. A visibility-aware weighting factor μ emphasizes $\mathcal{L}_{\text{NVS}}$ for unobserved regions. Finally, we align the hand to the completed object geometry to obtain an accurate hand–object spatial relationships.

Given the ratio β, we compute a weighting factor as $\mu = e^{-\beta^2/0.6}$ for a randomly sampled novel view $\mathbf{y}$.

Total loss: To summarize, at each iteration, we sample one RGB image from the video input images $\{I^{\text{obs}}\}$, the reference image I^{ref}, and multiple novel views $\{\mathbf{y}\}$, and then compute the total loss $\mathcal{L}_o$ for learning the object network parameters ψ_o where

$$\mathcal{L}_o = \mathcal{L}_{\text{RGB}} + \lambda_{\text{segm}}\mathcal{L}_{\text{segm}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \mu\mathcal{L}_{\text{NVS}} \quad (6)$$

with the weights $\lambda_{\text{segm}} = 10.0$, $\lambda_{\text{smooth}} = 50.0$ and μ being view-dependent. $\mathcal{L}_{\text{smooth}}$ enforces smoothness in the rendered normal image. Both $\mathcal{L}_{\text{RGB}}$ and $\mathcal{L}_{\text{segm}}$ are supervised by $\{I^{\text{obs}}\}$ and I^{ref}, while $\mathcal{L}_{\text{NVS}}$ is regularized by $\{\mathbf{y}\}$. See more details in SupMat.

3.4. Hand-object alignment

Since the object lives in the Zero-1-to-3 space and it does not have a metric scale, we need to estimate the scale of the object and align the hand and object based on reliable contact from image evidence.

Visible contact: Prior work [16] identifies contact points as the closest hand–object vertex pairs. However, this strategy is unreliable when the object surface is heavily occluded or poorly reconstructed. Consequently, we enforce hand–object contact constraints only on *visible contact pairs*.

Specifically, we take the vertices that frequently make contact on the fingertips (cyan mesh in Figure 3a) [23] as potential contacts, denoted by $\mathcal{V}_{\text{h}}^{\text{p}}$ (green dots in Figure 3a), while excluding those on the palm. We project each $\mathbf{v}_h \in \mathcal{V}_{\text{h}}^{\text{p}}$ onto the image and label it as a *visible contact candidate*, $\mathcal{V}_{\text{h}}^{\text{v}}$ (blue dots in Figure 3b), if its projection falls inside the hand mask; otherwise, it is marked *object-occluded* (red dots in Figure 3b) and discarded. Some candidates in $\mathcal{V}_{\text{h}}^{\text{v}}$ are occluded by the front fingertips (yellow dots in Figure 3c) even though they lie inside the hand mask. Fingertips whose mask area exceeds a predefined threshold are deemed visible; we retain only the contacts on these fingertips, yielding the final reliable visible contacts $\mathcal{V}_{\text{h}}$.

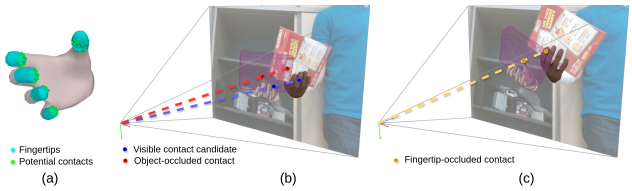

Figure 3. **Visible contact**. Our method first separate potential contacts on fingertip into visible contact candidates and object-occluded contacts using hand mask information. We then keep only those visible candidates whose projections are not occluded by other fingertips, yielding the final set of visible contact points.

For each contact in $\mathcal{V}_{\text{h}}$ we perform ray tracing to locate the corresponding object contact point $\mathcal{V}_{\text{o}}$, eliminating any non-surface contacts. This procedure yields a set of visible contact pairs $\{\mathcal{V}_{\text{h}}, \mathcal{V}_{\text{o}}\}$, which are then used to enforce reliable hand–object alignment.

Alignment process: For simplicity, we optimize only the hand while keeping the object fixed, using a contact-aware optimization procedure. Our experiments show that allowing hand pose θ to optimize slightly increases the hand pose error (MPJPE: 7.83 mm), compared to the original esti-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

mates (MPJPE: 4.62 mm). Therefore, we optimize only the hand translation $\mathbf{t}_h \in \mathbb{R}^3$ and scale $s \in \mathbb{R}$ using the following objective function:

$$\mathcal{L}_{\text{h}} = \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{kpoints}}\mathcal{L}_{\text{kpoints}} + \lambda_{\text{vsmooth}}\mathcal{L}_{\text{vsmooth}} + \lambda_{\text{penetr}}\mathcal{L}_{\text{penetr}} \quad (7)$$

where $\mathcal{L}_{\text{contact}}$ encourages a realistic hand-object contact relationship, $\mathcal{L}_{\text{kpoints}}$ enforces consistency in the hand joints 2D projections, $\mathcal{L}_{\text{vsmooth}}$ ensures temporal smoothness of hand vertices, and $\mathcal{L}_{\text{penetr}}$ penalizes hand-object penetration. The corresponding weights are: $\lambda_{\text{contact}} = 200.0$, $\lambda_{\text{kpoints}} = 20.0$, $\lambda_{\text{vsmooth}} = 20.0$, $\lambda_{\text{penetr}} = 0.003$. See more details in SupMat.

4. Experiments

We evaluate our approach against state-of-the-art baselines on the challenging task of *partially visible hand–object reconstruction*. The objective is to accurately recover both 3D object geometry and hand pose from short monocular video sequences that provide only limited observation.

4.1. Datasets

HO3D [21]: We evaluate our method using the HO3D-v3 dataset [20], which contains RGB videos of hands interacting with YCB objects [2], along with annotations of hand poses and object poses. We use the same sequences as HOLD [16], resulting in a total of 14 sequences. Since our focus is on reconstruction from short videos, we divide long sequences into 30-frame clips (see SupMat), each providing limited observations. The novel view synthesis (NVS) reference frame is also used as input to iHOI [82] and EasyHOI [42], which operate on a single image.

In-the-wild sequences: To evaluate real-world performance, we record short video clips of household objects in both indoor and outdoor environments, that do not reveal the complete objects. We also include short videos from the internet to further assess our method's generalization.

4.2. Metrics

Following [16, 70, 83], we use the root-relative mean per-joint error (MPJPE) in millimeters to evaluate hand pose accuracy and the Chamfer distance (CD) in centimeters to assess object reconstruction quality. We also compute the F-score in percentage at 5mm (F5) and 10mm (F10) to compare object local shape details. To measure the object's 3D pose and shape relative to the hand, we first subtract each object mesh using the predicted hand root position. We then compute the hand-relative Chamfer distance (CD_h) to measure the alignment accuracy. To evaluate the size of the reconstructed object relative to its real size after hand alignment, we introduce the Relative Scale (RS) metric. In particular, RS equals $\frac{1}{s} - 1$ if $s < 1$ and equals $1 - \frac{1}{s}$, if $s \geq 1$, where s is the scale by performing ICP alignment from the reconstructed object mesh to the ground-truth. Intuitively, when $RS = 0$, the reconstructed object size matches the actual physical size. As RS increases, the deviation from the real size becomes larger.

	CD [cm] ↓	F5 [%] ↑	F10 [%] ↑	MPJPE [mm] ↓	CD_h [cm] ↓	RS ↓
iHOI [81]	2.37	35.78	62.11	27.75	25.45	0.17
DiffHOI [83]	2.30	39.59	64.49	16.02	33.33	0.13
EasyHOI [42]	1.86	46.10	70.92	16.69	19.55	0.28
HOLD [16]	1.31	57.20	80.23	30.79	21.28	0.62
Ours	**0.87**	**69.72**	**92.15**	**4.62**	**2.39**	**0.11**

Table 1. **SOTA comparison in hand-object reconstruction**. We evaluate our method against four baselines on the HO3D dataset.

4.3. State-of-the-art comparison

Table 1 presents a comparison between our method and state-of-the-art (SOTA) hand-object reconstruction methods, including HOLD [16], EasyHOI [42], DiffHOI [83], and iHOI [81]. All methods are template-free. In particular, HOLD is a geometry-driven method that optimizes from long video inputs without using priors, while iHOI, EasyHOI, and DiffHOI are prior-driven methods. Our results indicate that MagicHOI achieves superior object reconstruction accuracy, as evidenced by the CD, F5, and F10 metrics. Among all baselines, HOLD overall struggles the most in this short video clip setting. Our method has the best object scale estimate indicated by the RS metric.

Figure 4 shows a qualitative comparison of all methods to provide an intuition of the reconstruction results. From these results, we observe that prior-driven methods, *e.g.* EasyHOI, DiffHOI, and iHOI, are capable of reconstructing complete object shapes even in regions that are occluded by the hand or not directly observed. This is in contrast to the non-prior method, such as HOLD. However, their reconstructed areas on the non-observed regions often exhibit distortions, particularly apparent in category-specific priors like those in iHOI and DiffHOI. EasyHOI generally does not reconstruct object surfaces that resemble those in the input images. In contrast, MagicHOI combines the observations from the RGB input with a category-agnostic NVS prior, allowing for the most accurate and complete object reconstruction from short videos. For short video sequences, artifacts in the reconstructed occluded regions can lead to significant pose alignment errors, as evidenced by the CD_h values for the HOLD method in Table 1. In contrast, our method reduces reconstruction noise in the non-observed object region, resulting in more stable and accurate hand alignment. Additional details are provided in SupMat. In addition, Figure 5 shows the video results across different frames, which indicate our method can generalize to bimanual setting in the rigid object sequences on ARCTIC [14] where there are more challenging hand-object occlusion, depth ambiguities, and more diverse hand pose variations.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

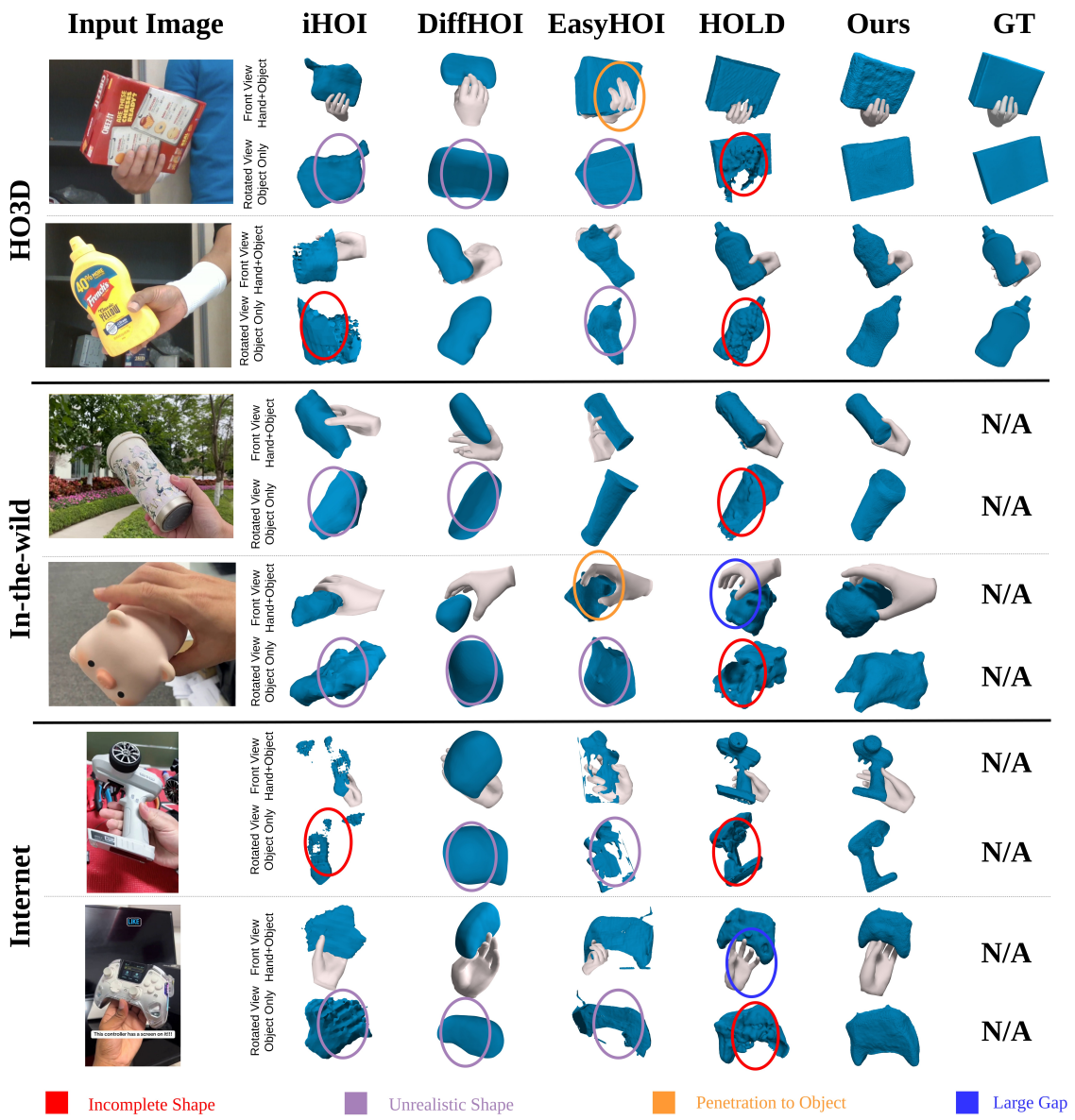

Figure 4. **Qualitative comparison with SOTA**. Reconstruction results from HO3D, in-the-wild, and internet short video sequences, comparing our method with SOTA baselines in both hand-object front view and object only rotated view. Our approach excels in completing unobserved object regions and reconstructing complex shapes, even under severe hand-induced and object-self occlusion, yielding detailed geometry and realistic hand-object spatial relationships. From the HOLD method column, it is evident that without an NVS prior, it fails to reconstruct the complete object shape in unobserved regions and there is a large gap between hand and object. Even with an NVS prior, the category-specific iHOI method struggles to recover full object geometry. EasyHOI, despite using an advanced NVS model, often hallucinates unrealistic shapes and results in hand-object penetrations. DiffHOI, which integrates both prior-driven and geometry-driven approaches, tends to reconstruct only simple shapes.

4.4. Ablation

To analyze the complementary roles of our novel view synthesis (NVS) backbone, and RGB image observations in the reconstruction quality in our short video clip setting, we conduct an ablation study in Table 2 based on different baselines set by changing the loss function in Equation 6:

- *RGB* (row 2): Object reconstruction using the loss terms $\mathcal{L}_{\text{RGB}}, \mathcal{L}_{\text{segm}}, \mathcal{L}_{\text{smooth}}$. Both $\mathcal{L}_{\text{RGB}}$ and $\mathcal{L}_{\text{segm}}$ are supervised only by the observed images $\{I^{\text{obs}}\}$.
- *NVS* (row 3): Reconstruction guided solely by NVS-derived SDS loss condition on the reference image. In other words, we use all loss terms, but supervise $\mathcal{L}_{\text{RGB}}$

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

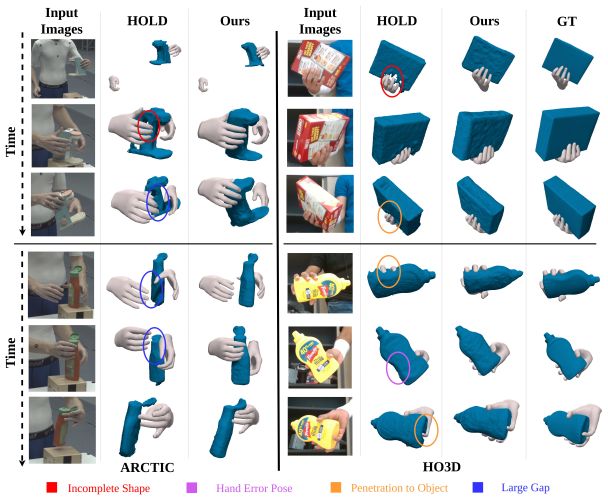

Figure 5. **Video qualitative comparison with SOTA**.

RGB	NVS	WS	CD [cm] ↓	F5 [%] ↑	F10 [%] ↑	CD_h [cm] ↓	RS ↓
✓	×	×	1.37	55.02	78.24	3.21	0.20
×	✓	×	1.54	55.59	78.97	11.97	0.26
✓	✓	×	0.91	68.62	90.81	3.77	0.17
✓	✓	✓	**0.89**	**69.72**	**92.15**	**2.39**	**0.11**

Table 2. **Ablation study**. Integrating all observed RGB images into the NVS with the occlusion-aware weighting strategy (WS) significantly improves object reconstruction, as indicated by CD, F5, and F10. The refined object shape further enhances the hand-object relationships, as reflected in CD_h and RS.

and $\mathcal{L}_{\text{segm}}$ only with the reference image I^{ref} (no observed view supervision) and fix $\mu = 1$.

- *RGB+NVS* (row 4): Integration of RGB inputs with NVS via alignment without weighting strategy by setting μ to one. In other words, we use all loss terms, but fix $\mu = 1$.
- *Ours* (row 5): Object optimization with all loss terms.

Mutual benefits of NVS model and RGB observations: As shown in Table 2, combining RGB image observations from the input video and the NVS prior significantly improves reconstruction quality (see *RGB* or *NVS vs. RGB+NVS*), evidenced by lower CD, hand-object alignment score CD_h, RS, and higher F5/F10 scores. Figure 6 provides qualitative results to highlight the improvements. We see that only using RGB, the model fails to reconstruct regions occluded by hands or the object itself. Furthermore, only using NVS model without incorporating other image evidence results in distorted and unrealistic object shapes (*e.g.*, overly thick box/driller base, and unrealistic bottle geometry near occlusions). By combining with image observations and the NVS prior, our model reconstructs more complete and accurate object shapes with minimal artifacts, closely matching the ground-truth.

Visibility-aware weighting strategy: To assess the effect of our visibility-aware weighting strategy, we compare

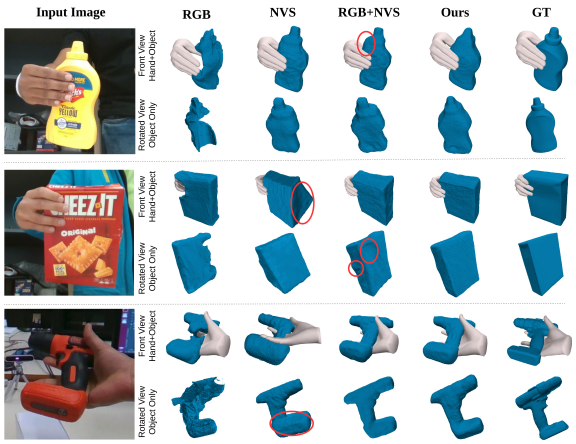

Figure 6. **Ablation Study**. Comparing *RGB+NVS* and *Ours* columns, the yellow bottle (row 1) exhibits significant distortion near the hand-occluded region, while the red box (row 2) shows numerous artifacts on observed surfaces in the *RGB+NVS* column due to lack of a weighting strategy. Comparing *NVS* and *Ours* columns, the red box reconstructs an unrealistic shape on unobserved surfaces, and the driller (row 3) has an excessively thick base in the *NVS* column. Comparing *RGB* and *Ours* columns, the object fails to reconstruct a complete shape in occluded regions.

our full model *Ours* to the baseline *RGB+NVS* that omits this strategy while retaining both RGB observations and NVS priors (Table 2). Quantitatively, the weighting strategy leads to improved reconstruction accuracy, as indicated by lower CD and higher F5/F10 scores. Qualitatively, Figure 6 illustrates that removing this strategy results in surface distortions across both visible and occluded regions, and produces less smooth surfaces, especially in occluded areas.

5. Conclusion

In this paper, we present MagicHOI, a hand-object reconstruction framework that incorporates novel view synthesis (NVS) priors for reconstructing from short video sequences with significant occlusions. By integrating NVS into the reconstruction pipeline, MagicHOI enables accurate joint reconstruction of hands and objects even in such challenging interaction scenarios. Qualitative and quantitative results show that MagicHOI outperforms SOTA methods without requiring full object or hand visibility in both controlled and in-the-wild settings.

Acknowledgment: We thank Muhammed Kocabas, Xu Chen, Bonan Liu for detailed discussions and insightful feedback, Handi Yin for support and International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Maria Parelli.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3D hand shape and pose from images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10843–10852, 2019. 2

[2] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 6

[3] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, pages 12417–12426, 2021. 2, 3

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[5] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[6] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12890–12900, 2023. 3

[7] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *arXiv*, 2023. 3

[8] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20545–20554, 2022. 2

[9] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020. 2, 3

[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. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 3

[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*, pages 13142–13153, 2023. 3

[12] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. HMP: Hand motion priors for pose and shape estimation from video. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2

[13] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J. Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *International Conference on 3D Vision (3DV)*, pages 1–10, 2021. 2

[14] 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

[15] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, Fei Li, Liu Zheng, Feng Lu, Karim Abou Zeid, Bastian Leibe, Jeongwan On, Seungryul Baek, Aditya Prakash, Saurabh Gupta, Kun He, Yoichi Sato, Otmar Hilliges, Hyung Jin Chang, and Angela Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. *arXiv preprint arXiv: 2403.16428*, 2024. 2, 3

[16] 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. 2, 3, 5, 6

[17] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M. Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 23600–23611, 2023. 2

[18] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021. 3

[19] Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, and Houqiang Li. HandNeRF: Neural radiance fields for animatable interacting hands. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21078–21087, 2023. 2

[20] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020. 6

[21] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[22] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[23] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 2, 3, 5

[24] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

[25] 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. 3

[26] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *arxiv*, 2023. 3

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[28] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[29] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3

[30] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2

[31] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 2

[32] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21169–21178, 2023. 2

[33] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 3

[34] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 20395–20405, 2023. 2

[35] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2761–2770, 2022. 2

[36] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 677–693. Springer, 2020. 3

[37] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6517–6526, 2024. 3

[38] Bonan Liu, Handi Yin, Manuel Kaufmann, Jinhao He, Sammy Christen, Jie Song, and Pan Hui. Egohdm: An online egocentric-inertial human motion capture, localization, and dense mapping system. *arXiv preprint arXiv:2409.00343*, 2024. 3

[39] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2, 3, 4

[40] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2021. 2, 3

[41] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[42] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild, 2024. 6

[43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 3

[44] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3D interacting hand pose estimation by hand de-occlusion and removal. In *European Conference on Computer Vision (ECCV)*, pages 380–397. Springer, 2022. 2

[45] Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 17028–17037, 2023.

[46] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, pages 548–564, 2020. 2

[47] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *arXiv preprint arXiv:2310.17768*, 2023. 2

[48] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *Computer Vision and Pattern Recognition (CVPR)*, pages 49–59, 2018. 2

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

[50] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 2

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

[51] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 3

[52] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 2

[53] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3, 4

[54] Aditya Prakash, Matthew Chang, Matthew Jin, and Saurabh Gupta. Learning hand-held object reconstruction from in-the-wild videos. *arXiv*, 2305.03036, 2023. 3

[55] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *International Conference on Computer Vision (ICCV)*, pages 15100–15111, 2023. 3

[56] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 3

[57] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[58] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[59] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 3

[60] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[61] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[62] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017. 2

[63] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 89–98, 2018.

[64] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision (ECCV)*, pages 211–228, 2020. 2

[65] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3D hand pose estimation from monocular RGB via contrastive learning. In *International Conference on Computer Vision (ICCV)*, pages 11210–11219, 2021. 2

[66] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. SHOWMe: Benchmarking object-agnostic hand-object 3d reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 1935–1944, 2023. 3

[67] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 3

[68] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3

[69] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414, 2019. 3

[70] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Computer Vision and Pattern Recognition (CVPR)*, pages 3405–3414, 2019. 6

[71] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2019. 2, 3

[72] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9:137–154, 1992. 3

[73] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1674, 2022. 3

[74] Tze Ho Elden Tse, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, et al. Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images. In *International Conference on Computer Vision (ICCV)*, pages 14666–14677, 2023. 2

[75] Dimitrios Tzionas and Juergen Gall. A comparison of directional distances for hand pose estimation. In *German Conference on Pattern Recognition (GCPR)*, pages 131–141, 2013. 2

[76] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3

[77] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pages 57–74. Springer, 2024. 3

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.

[78] Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d from images and videos, 2024. 3

[79] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[80] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 2

[81] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

[82] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3895–3905, 2022. 2, 6

[83] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *International Conference on Computer Vision (ICCV)*, 2023. 6

[84] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis, 2024. 3

[85] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 3

[86] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dynhamr: Recovering 4d interacting hand motion from a dynamic camera. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[87] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3D pose and shape reconstruction from single color image. In *International Conference on Computer Vision (ICCV)*, pages 11354–11363, 2021. 2

[88] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision (ECCV)*, pages 386–403. Springer, 2024. 2

[89] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024. 2

[90] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *International Conference on Computer Vision (ICCV)*, pages 2354–2364, 2019. 2

[91] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*, pages 1–19, 2022. 3

[92] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5345–5354, 2020. 3

[93] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. TempCLR: Reconstructing hands via time-coherent contrastive learning. In *International Conference on 3D Vision (3DV)*, pages 627–636, 2022. 2

[94] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017. 2

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:59:52 UTC from IEEE Xplore. Restrictions apply.