In-Hand 3D Object Reconstruction from a Monocular RGB Video

Shijian Jiang[1], Qi Ye[1,2*], Rengan Xie[3], Yuchi Huo[3,4], Xiang Li[5], Yang Zhou[5], Jiming Chen[1]

[1]College of Control Science and Engineering, Zhejiang University
[2]Key Lab of CS&AUS of Zhejiang Province
[3]State Key Lab of CAD&CG, Zhejiang University
[4]Zhejiang Lab
[5]OPPO US Research Center
{jsj630, qi.ye, rgxie}@zju.edu.cn, huo.yuchi.sc@gmail.com,
{xiang.li, yang.zhou}@oppo.com, cjm@zju.edu.cn

Abstract

Our work aims to reconstruct a 3D object that is held and rotated by a hand in front of a static RGB camera. Previous methods that use implicit neural representations to recover the geometry of a generic hand-held object from multi-view images achieved compelling results in the visible part of the object. However, these methods falter in accurately capturing the shape within the hand-object contact region due to occlusion. In this paper, we propose a novel method that deals with surface reconstruction under occlusion by incorporating priors of 2D occlusion elucidation and physical contact constraints. For the former, we introduce an object amodal completion network to infer the 2D complete mask of objects under occlusion. To ensure the accuracy and view consistency of the predicted 2D amodal masks, we devise a joint optimization method for both amodal mask refinement and 3D reconstruction. For the latter, we impose penetration and attraction constraints on the local geometry in contact regions. We evaluate our approach on HO3D and HOD datasets and demonstrate that it outperforms the state-of-the-art methods in terms of reconstruction surface quality, with an improvement of 52% on HO3D and 20% on HOD. Project webpage: https://east-j.github.io/ihor.

Introduction

3D object reconstruction from images has many applications in fields such as robotic manipulation and AR/VR. A handy and low-cost way to obtain 3D models is to rotate an object in hand in front of a camera and reconstruct the 3D objects from a captured video, which is the focus of this work. However, in-hand 3D object reconstruction in this setting poses several challenges, such as the lack of prior knowledge of the object shape, the estimation of the relative poses between the camera and the object, and particularly, the occlusion caused by the hand-object interaction.

Implicit neural representations, combined with volume rendering techniques (Wang et al. 2021; Yariv et al. 2021), have proven to be remarkably effective in reconstructing 3D geometry from multi-view images without requiring any prior knowledge of the object. Several in-hand object reconstruction works based on these representations (Hampali et al. 2022; Huang et al. 2022; Wen et al. 2023) have achieved compelling results in the visible part of the object. However, their performance degrades significantly when objects are heavily occluded by the hand as these methods optimize 3D object models to fit the observed images only.

In this paper, we argue that dealing with object surface reconstruction under occlusion demands the incorporation of additional priors beyond direct observation. Humans are capable of intuitively elucidating objects under occlusion. Some works (Back et al. 2022; Zhan et al. 2020; Zhou et al. 2021) therefore explore large-scale data to learn the capability of occlusion elucidation for images with amodal mask completion. However, leveraging this 2D elucidation capability for multi-view 3D reconstruction is challenging as the amodal masks may be inaccurate and inconsistent across multiple views, especially in the heavily occluded areas. To address this issue, we add a semantic amodal mask head to the implicit 3D reconstruction neural network and refine the masks by jointly optimizing the parameters of both networks.

Though the completed amodal mask can help to constrain a rough global shape of an object, it may not reconstruct local surfaces well, as small changes in the 3D local surfaces might not render apparent changes in the 2D masks. On the other hand, we humans can feel the object shapes and manipulate objects by hands without seeing them and an attempt has been made in (Yin et al. 2023) that robotic hands can accomplish similar tasks with only simple tactile information (touch object surfaces or not) collected by tactile sensors attached on robotic hands. With this inspiration, we propose to infer the occluded local object surfaces by reasoning about the physical contact between objects and hands: the reconstructed hands and objects must not intersect with each other and must be in contact to enforce friction so that objects will not fall due to gravity. To this end, we introduce penetration and attraction penalties to guide the inference of the occluded surface in contact areas with hands.

By incorporating the 2D occlusion elucidation and the physical contact priors, we propose a novel in-hand 3D object reconstruction method based on implicit representations from a monocular RGB video sequence. We evaluate our

*Corresponding author.

method on two datasets HO3D (Hampali et al. 2020) and HOD (Huang et al. 2022). The experiments show that our method can accurately reconstruct objects in both visible and invisible parts and significantly outperforms the state-of-the-art methods in terms of reconstruction quality. Our contributions can be summarized as follows:

- We propose a novel method for implicit hand-held object reconstruction that first leverages priors of 2D occlusion elucidation and physical contact constraints.
- For the 2D occlusion elucidation prior, we introduce an amodal mask head and a joint optimization method for both amodal mask refinement and 3D object reconstruction to ensure the accuracy and view consistency of the predicted amodal masks.
- For the physical contact prior, we devise penetration loss and attraction loss to regularize the occluded object surface.
- We conduct extensive experiments on HO3D and HOD datasets and demonstrate that our approach outperforms state-of-the-art methods in terms of surface quality, with an improvement of 52% on HO3D and 20% on HOD.

Related Works

Multi-view 3D Reconstruction. Recovering 3D geometry from multi-view images has a history in computer graphics and computer vision. Traditional methods (Schönberger et al. 2016; Barnes et al. 2009) involve SFM (Schonberger and Frahm 2016) for camera estimation, dense point clouds via MVS, and Poisson reconstruction (Kazhdan, Bolitho, and Hoppe 2006) for meshing. Recently, there is a growing trend to use MLPs to represent 3D appearance and geometry. For instance, NeRF (Mildenhall et al. 2021) combines the volume rendering with implicit functions by minimizing observed-rendered differences. Inspired by NeRF (Mildenhall et al. 2021) and SDF (Park et al. 2019), NeuS (Wang et al. 2021) and VolSDF (Yariv et al. 2021) advance surface quality by replacing the density field with the signed distance fields. We experimentally find these methods tend to produce poor results in the invisible part due to insufficient observation information. Our method is specially designed to handle this occlusion in hand-object interaction scenes.

3D Hand-held Objects Reconstruction. The 3D reconstruction of manually manipulated objects is a very challenging task due to the heavy occlusion and the variety of objects. To simplify the reconstruction task, several methods (Cao et al. 2021; Yang et al. 2022) reduce the reconstruction to a 6DoF pose estimation. Some other works rely on additional depth information (Zhang et al. 2019) or point cloud (Chen et al. 2022a) to address this challenge. Recent learning-based approaches attempt to directly infer the representations of hands and objects from a monocular RGB image. Hasson et al. (2019) utilizes AtlasNet (Groueix et al. 2018) to recover the object meshes, limited to reconstructing simple objects. (Ye, Gupta, and Tulsiani 2022; Karunratanakul et al. 2020; Chen et al. 2022b) use the implicit function to predict the object shape. However, these learning-based methods rely heavily on the dataset, and the reconstructed meshes lack details. In contrast, our method only needs RGB images as supervision and does not need any prior knowledge of the objects. Furthermore, our method excels in recovering object meshes with more details. Most related to our work, (Huang et al. 2022; Hampali et al. 2022) reconstructs a hand-held object from RGB monocular video, leveraging the differentiable SDF rendering technique. Huang et al. (2022) treats the interacting hand and object as a whole and separates them using an estimated semantic class of each vertex. Hampali et al. (2022) focus only on the object part. Therefore, these methods do not take occlusion into account, resulting in incomplete surfaces in the hand-occluded part of the object. In contrast, we incorporate contact physical constraints and 2D amodal priors, leading to substantial improvements in the quality of object reconstruction.

Occlusion Handling. As hands/humans are often severely occluded by objects, several approaches aim to recover the content of the occluded parts. The first approach involves utilizing temporal information. Cheng et al. (2019) feed filtered reliable 2D keypoints to 2D and 3D temporal convolutional networks that enforce temporal smoothness to produce a complete 3D pose. The second approach utilizes the spatial attention mechanism. Park et al. (2022) propose a feature injection mechanism for occlusion-robust 3D hand mesh reconstruction. The third applies the amodal mask to perceive the invisible part. Amodal mask refers to the ability to perceive entire objects despite partial occlusion, which has the potential to make computers more human-like in handling occlusion. Ours is related to the amodal mask. Prior studies (Zhan et al. 2020; Zhou et al. 2021) have employed amodal masks to aid in recovering occluded 2D content from images. In our method, we leverage amodal masks to significantly enhance the optimization of neural implicit fields, thereby introducing a novel means of improving reconstruction in occluded regions. However, simply applying initial masks suffers from two issues: (1) some of them may be incorrect, and (2) they are not multi-view consistent. To address these issues, we use a semantic head to refine the masks, resulting in improved reconstruction quality.

Methods

Our objective is to reconstruct the 3D object from a video sequence $\{I_k\}_{k=0,\dots,N}$, where a hand holds a rigid object and rotates it in front of a static RGB camera. In our problem, hand poses are assumed to be fully constrained by objects. Therefore, hand poses are the same across different frames; only global translation and rotation of the hand may differ.

We adopt the widely used 3D parametric model MANO (Romero, Tzionas, and Black 2017) to represent the hand. MANO can generate hand mesh by inputting two sets of parameters. Shape parameters $\beta \in \mathbb{R}^{10}$ control the hand shape and pose parameters $\theta \in \mathbb{R}^{16\times 3}$ represent the rotation of 16 joints. We estimate the parameter of MANO along with the relative rotation $R \in SO(3)$ and translation $T \in \mathbb{R}^3$ between the hand and camera from RGB sequence images. Thus, the hand mesh can be defined as $H_k = \{MANO(\beta, \theta), R_k, T_k\}$, where k indicates the k_{th}

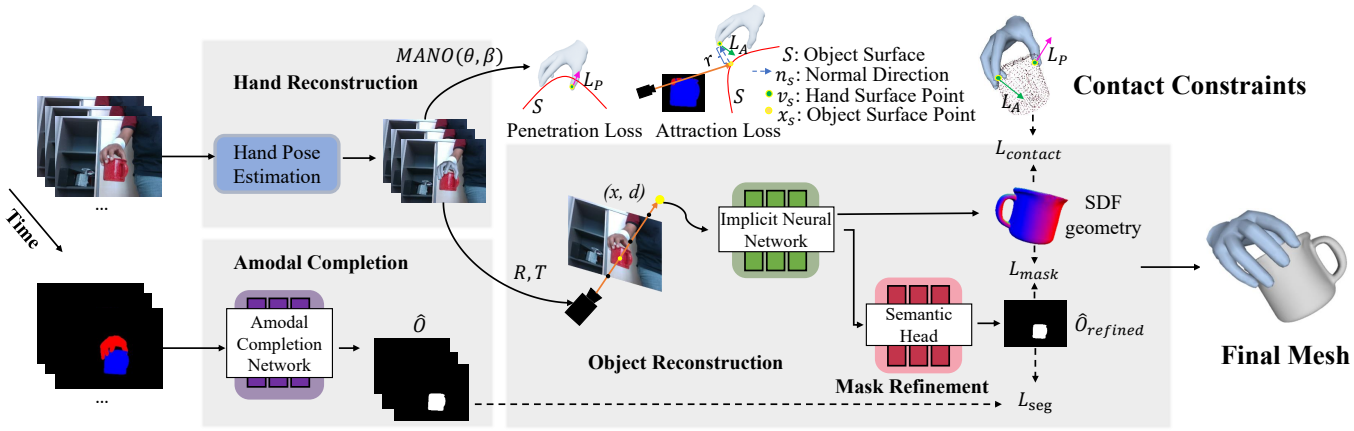

Figure 1: Overview of our framework, consisting of three parts. Hand reconstruction: we optimize the MANO parameters to reconstruct both the hand mesh and camera-relative motion. Amodal completion: by utilizing the amodal completion network, we can recover amodal masks from the input hand and object segmentation maps. Object reconstruction: we learn the 3D objects with the neural implicit field, which is supervised by input images, hand meshes, and amodal masks. To improve the consistency and quality of the predicted amodal masks, we refine them with a semantic head.

frame, and β, θ are shared for all the frames.

The 3D object shape is represented by an SDF-based implicit function F_g. By mapping a query 3D point to a signed distance from the object surface, we can extract the zero-level set as the object surface. As the object shape reconstruction is supervised by image sequences, we represent the object appearance by an extra implicit function F_c. Both F_g and F_c are optimized through volume rendering to minimize differences between input images I_k and rendered images $\hat{I}_k$. However, employing this technique alone leads to incomplete reconstruction results due to the absence of observations in the occluded region (hand-object contact area).

To address these challenges, we incorporate amodal masks and physical contact guidance into the neural rendering framework for constraining the reconstruction of the invisible parts. Specifically, an overview of our method is shown in Fig. 1. Given a monocular RGB input video of a moving hand-held object, our method reconstructs the hand-held object without any prior of the object category. Same with (Huang et al. 2022), we assume that the object is firmly grasped, enabling us to jointly predict the object motion by hand estimator. We apply the SDF-based implicit function to represent the object. To improve the reconstruction quality in contact areas, the 2D occlusion elucidation and the physical contact priors are leveraged. First, we utilize amodal masks to detect and supervise these regions. We ensure the consistency and quality of the amodal masks by refining them using an additional semantic head after the implicit neural network. Moreover, we apply contact constraints, which require that the object does not intersect with the hand and is in close proximity when they make contact.

3D Hand Reconstruction

The first step of our framework is to perform hand pose estimation. We employ a learning-based approach to achieve a robust initialization and further optimize the hand model by fitting it to 2D keypoints.

Previous research (Lv et al. 2021) has demonstrated that directly fitting MANO to 2D keypoints is highly non-linear and very sensitive to initial parameters. Therefore, we first utilize the pre-trained monocular hand reconstruction model HandOccNet (Park et al. 2022) to estimate the hand model parameters of each frame and then average them to obtain a more robust initialization. The hand model can be optimized by minimizing the difference between 2D keypoints detections and reprojection of 3D joints:

$$H_k^* = \arg\min_{H_k} \left(\sum_{i=1}^{16} ||\pi(J_{3d}^i(H_k)) - J_{2d}^i|| + \lambda_{reg}\mathcal{L}_{reg}\right), \quad (1)$$

where $\pi(.)$ denotes the projection operation, J_{3d} and J_{2d} represent the 3D and 2D joint location respectively. The last term $\mathcal{L}_{reg} = ||\beta||_2^2 + ||\theta||_2^2$ is for regularization. We use Mediapipe (Lugaresi et al. 2019) to obtain 2D keypoints.

Inspired by (Hasson et al. 2021), we add an additional term into the optimization over k frames to force temporal smoothness:

$$\sum_k ||H_k - H_{k-1}||. \quad (2)$$

Jointly optimizing the energy function may be unstable. Therefore, we first optimize the relative rotation R and translation T, followed by the optimization of the MANO parameters. After optimization, we can transfer RGB video sequences into multi-view images in the hand-centric coordinates, with the hand wrist serving as the origin.

Object Reconstruction

A key to our framework is learning a hand-centric Signed Distance Function (SDF) representation, enabling the learning of a consistent 3D shape and appearance of the object. It is learned per sequence and does not require pre-training. To optimize the SDF, We adopt the NeuS method (Wang et al.

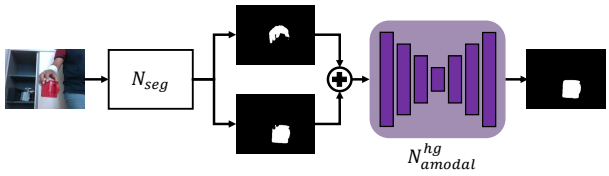

Figure 2: Overview of the amodal completion network architecture. Given initial segmentation maps of hand and object as input, an hourglass module is applied to produce the object amodal masks.

2021) to utilize the volume rendering technique, while also integrating amodal mask and contact constraints.

SDF-based Implicit Representation. We represent the geometry and appearance by two MLP networks, a geometry network $F_g : \mathbb{R}^3 \rightarrow \mathbb{R}$ and a color network $F_c : \mathbb{R}^3 \times \mathbb{S}^2 \rightarrow \mathbb{R}^3$. Given a 3D point x, the geometry network maps it to the SDF value $F_g(x)$, and the color network takes x along with view direction d as inputs and outputs color $F_c(x, d)$. The object surface is then extracted as the zero-level set of the SDF:

$$\mathcal{S} = \{x | F_g(x) = 0\}. \tag{3}$$

For each pixel, we sample a set of points along the corresponding camera ray, denoted as $\{p_i = o + t_i d | t_i \in [t_n, t_f]\}$, where p_i are the sampled points, o is the camera position, d is the viewing direction, and t_n, t_f denote the bound of the sample ray. Then we can get the rendered color as:

$$\hat{c} = \sum_i T_i \alpha_i F_c(p_i, d), \tag{4}$$

where $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ is the discrete accumulated transmittance, and $\alpha_i = 1 - exp(-\int_{t_i}^{t_{i+1}} \rho(t)dt)$ denotes the discrete opacity values. $\rho(t)$ is the opaque density transferred from SDF as defined in (Wang et al. 2021).

Amodal for Shape Completion. The amodal completion network targets at segmenting the invisible part of the object to offer the understanding of its complete shape, which can then be utilized to supervise the object geometry. Modern 2D amodal segmentation models (Tran et al. 2022) trained on large labeled datasets can provide reasonable predictions. However, in our case, we require amodal masks for a range of categories that may not be present in the training dataset. Thus, we complete the hand and object segmentation maps into object amodal masks. This simultaneous input of hand and object segmentation maps is not restricted by object categories and effectively captures the patterns of hand-object interaction.

With this observation, we utilize a simple hourglass network (Newell, Yang, and Deng 2016) N^{hg}_{amodal} to estimate amodal masks $\hat{O}$ ignoring the category information. Specifically, as Fig. 2 shows, we first obtain the segmentation maps of the hand and object from N_{seg}, an off-the-shelf method (Boerdijk et al. 2021). Then, the segmentation maps of hand and object M are concatenated and fed into

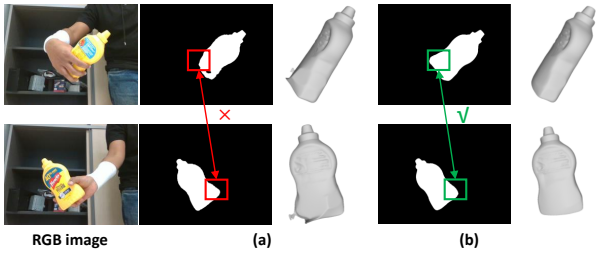

Figure 3: Qualitative results of mask refinement and reconstructed results. (a) Predicted masks from the completion network and corresponding mesh. (b) Refined results and corresponding meshes.

N^{hg}_{amodal} to generate the amodal results. The network is trained on ObMan (Hasson et al. 2019). ObMan is a large-scale hand-object interaction dataset, wherein ground-truth amodal masks O can be obtained by rendering 3D models. The cross-entropy loss $\mathcal{L}_{CE}(.)$ is applied to supervise the predictions:

$$\mathcal{L}_{amodal} = \mathcal{L}_{CE}(O, \hat{O}). \tag{5}$$

Mask Refinement with View Consistency. Since the amodal mask of each frame is predicted independently, they lack multi-view consistency and are often inaccurate, especially when the object is heavily occluded, as shown in Fig. 3(a). To resolve these inconsistencies and refine the masks, we use an additional semantic head. As demonstrated in (Zhi et al. 2021), the semantic neural field can naturally leverage the multi-view consensus to improve the accuracy of segmentation. Given a 3D point x, the semantic head predicts a logit $l(x)$, which is defined as:

$$l(x) = F_s(x), \tag{6}$$

where F_s is also an MLP network. Similar to color, we adopt volume rendering to convert the semantic logits into 2D semantic maps denoted as L, as presented in Eq. 4:

$$\hat{L} = \sum_i T_i \alpha_i F_s(p_i). \tag{7}$$

We then use a softmax to compute the probabilities and supervise by predicted amodal masks $\hat{O}$ using the classification loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{CE}(\hat{O}, \hat{L}). \tag{8}$$

The semantic head is trained together with the implicit neural network. After several iterations, we obtain refined amodal masks $\hat{O}_{refined}$ by thresholding the probabilities of $\hat{L}$, which are then used to supervise the geometry again. Fig. 3(b) presents examples of refined masks, which demonstrate the effectiveness in improving the accuracy and consistency of masks.

Hand-Object Contact Constraints. We leverage the constraints that guide objects interacting in physical contact. In particular, when grasping the objects, there is no interpenetration between the hand and the object, and that, contacts occur at the surface of both. We express these contact constraints as a differentiable loss $\mathcal{L}_{contact}$, that can be easily applied in the neural rendering framework.

Penetration. To prevent penetration between the hand and object, we define a penetration loss $\mathcal{L}_P$. Following (Ye, Gupta, and Tulsiani 2022), we penalize if any hand mesh vert v is predicted to have negative SDF value by geometry network F_g, which can be formulated as:

$$\mathcal{L}_P = \sum_{v \in H} ||max(-F_g(v), 0)||. \tag{9}$$

Attraction. We further define an attraction loss $\mathcal{L}_A$ to encourage the contact. We sample the rays only from pixels with the amodal mask value of 1. By calculating the surface intersection of these rays, we can obtain the object surface point x_s, along with its corresponding normal n_s in the contact area. Then we cast a ray along n_s direction and find the nearest point it intersects the hand mesh. We determine whether the object is in contact with the hand based on the distance r between the surface point x_s and the intersection point v_s. The process is illustrated in Fig. 1. For object surface points in contact with hands (the distance r smaller than a threshold τ), we first encourage the object surface to be close to the hand surface. This involves ensuring that $F_g(v_s)$ approaches 0, denoted as $\mathcal{L}_A = ||F_g(v_s)||$. However, we find this constraint for points in contact only usually does not reconstruct the surface we desire, as shown in Fig. 4. This can be attributed to: 1) objects surface points near the hand but not in contact are usually occluded in all images, lacking constraints for the reconstruction; 2) the hard threshold results in abrupt surface constraint changes near contact and non-contact regions; 3) the hard threshold is sensitive to the hand reconstruction quality. Consequently, we also introduce constraints for object surface points near the contact regions but not in contact by encouraging the SDF values to be a function of the distance between a surface point and the hand surface.

Therefore, our overall attraction loss $\mathcal{L}_A$ is defined as :

$$\mathcal{L}_A = \begin{cases} ||F_g(v_s)|| & r < \tau, \\ ||F_g(v_s) - \beta tanh(\frac{r-\tau}{\beta})|| & r \geq \tau, \end{cases} \tag{10}$$

where τ, β are the hyper-parameters. In our study, we empirically set $\tau = 0.001, \beta = 0.5$.

Finally, we employ the surface smooth regularization (Oechsle, Peng, and Geiger 2021) in the contact region, which encourages n_s to be similar with its neighborhood:

$$\mathcal{L}_S = \sum ||n_s - n_{s+\epsilon}||_2. \tag{11}$$

Our final contact loss can be formulated as:

$$\mathcal{L}_{contact} = \lambda_P \mathcal{L}_P + \lambda_A \mathcal{L}_A + \lambda_S \mathcal{L}_S. \tag{12}$$

Training

In the training stage, we employ multiple loss functions to optimize the neural implicit field. Specifically, during training, we sample N_r rays and their corresponding reference colors C_i, and amodal mask values $\hat{O}_i$. We use the amodal masks obtained by mask refinement as ground truth for the corresponding calculation. For each ray, we sample N_p points. The color loss $\mathcal{L}_{color}$ is defined as:

$$\mathcal{L}_{color} = \frac{1}{N_r} \sum_i ||\hat{C}_i - C_i||. \tag{13}$$

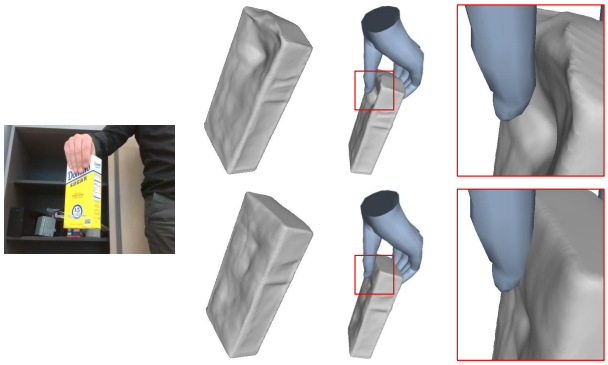

Figure 4: Comparison of different contact loss designs. Top: reconstructed mesh using the contact loss only constraining the object surface points in contact. Bottom: reconstructed mesh using our loss.

We apply Eikonal loss (Gropp et al. 2020) and mask loss to regularize the SDF:

$$\mathcal{L}_{eik} = \frac{1}{N_r N_p} \sum_{k,i} (||\nabla F_g(p_{k,i})||_2 - 1)^2, \tag{14}$$

$$\mathcal{L}_{mask} = \mathcal{L}_{CE}(\hat{W}_i, \hat{O}_i), \tag{15}$$

where $\hat{W}_i = \sum_k T_k \alpha_{i,k}$ is the sum of weight along the ray. The overall training loss is:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{color} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{eik}\mathcal{L}_{eik} + \\ \lambda_{seg}\mathcal{L}_{seg} + \lambda_{contact}\mathcal{L}_{contact}, \end{aligned} \tag{16}$$

where $\lambda_{mask} = 10, \lambda_{eik} = 0.1, \lambda_{seg} = 0.1, \lambda_{contact} = 5$ are set empirically.

Optimize Camera Poses. The estimated camera pose from the hand pose may not be accurate due to occlusion between the hand and object, leading to a significant degradation in the quality of the reconstruction. Pose refinement has been explored in previous NeRF-based models (Lin et al. 2021). We incorporate this to effectively optimize for the poses jointly with the object representation.

Experiments

In this section, we first present the hand-object interaction datasets and evaluation metrics. Subsequently, we compare our method to state-of-the-art approaches and provide ablation results.

Experimental Setups

Implementation Details. We use the same network architecture as NeuS (Wang et al. 2021), following them to normalize all cameras within a unit sphere and initialize network parameters to approximate the SDF to a unit sphere. For training the model, we use Adam optimizer (Kingma and Ba 2014) with a learning rate of $5e{-}4$ and sampled 1024 rays per batch for a total of $100k$ iterations. The training takes about 14 hours in total on a single NVIDIA RTX3090 GPU. Our implementation is based on PyTorch.

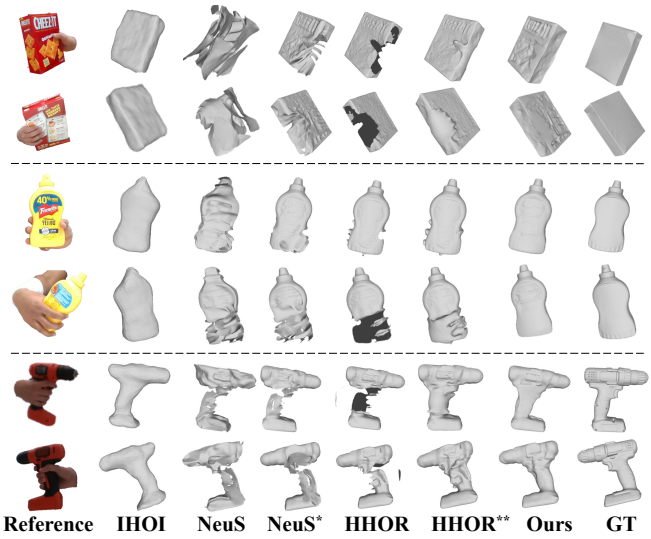

Figure 5: Qualitative comparison with baseline methods on the HO3D dataset. * indicates that the method uses the gound-truth camera pose. ** indicates that the method with post-processing. Compared with other methods, we can produce more complete and detailed reconstruction results. Zoom in for details.

Figure 6: Qualitative comparison with baselines methods on HOD dataset.

Datasets. To evaluate our method, we perform the experiments on HO3D (Hampali et al. 2020) and HOD (Huang et al. 2022).

- HO3D is a dataset that contains the RGBD video of a hand interacting with YCB objects (Calli et al. 2015) with 3D annotations of both hand and object. We select the 5 sequences in which the objects are firmly grasped by users for our experiments.
- HOD aims to reconstruct hand-held objects from RGB sequences, containing 35 objects. However, only 14 ground truth scanned meshes are available for evaluation.

In the experiment, we use 500 frames from each sequence of HO3D and all provided frames of HOD.

Evaluation Metrics. We evaluate both the quality of object reconstruction and the relationship between the hand and object. Firstly, we use Marching Cubes (Lorensen and Cline 1987) to extract the object mesh from SDF. Following prior research, we then evaluate the object reconstruction quality using **Chamfer Distance (CD)**. As the reconstructed result and ground truth mesh are in different coordinates, we follow Huang et al. (2022) to normalize each mesh to unit size and apply the ICP to register the reconstructed mesh with the ground truth mesh. For evaluating the relationship between the hand and object, we report the **Intersection Volume (Vol)** in cm^3 between the hand mesh and object mesh, similar to (Ye, Gupta, and Tulsiani 2022).

Methods	HO3D		HOD	
	CD↓	Vol↓	CD↓	Vol↓
IHOI	2.206	2.192	6.607	**0.505**
NeuS	3.310	-	3.093	-
NeuS*	0.872	-	-	-
HHOR	1.256	-	0.589	-
HHOR**	0.591	7.771	0.347	1.738
Ours	**0.282**	**0.327**	**0.277**	0.757

Table 1: Quantitative results of object reconstruction on the HO3D and HOD datasets using Chamfer distance (in unit size) and intersection volume (cm^3). * indicates that the method uses the ground-truth camera pose. ** indicates that the method with post-processing.

Comparison Baselines. In our evaluation, we compare our method with several existing approaches, including (1) **HHOR** (Huang et al. 2022), which addresses the same problem as ours. Additionally, we present the results of HHOR with post-processing (denoted as **HHOR****) using MeshLab (Cignoni et al. 2008) to remove unnecessary parts and fill holes; (2) **NeuS** (Wang et al. 2021) serves as the foundation for our method. Since the reconstruction quality of NeuS is greatly influenced by the accuracy of the camera pose, we also report its results on HO3D using ground truth camera poses (denoted as **NeuS***) for a fair comparison, as HOD does not offer ground truth data; (3) **IHOI** (Ye, Gupta, and Tulsiani 2022), a learning-based single image hand-held object reconstruction method that is pre-trained on sequences of the HO3D and other datasets. We evaluate IHOI on each frame of the sequence and report the average results. As the results of NeuS and HHOR are not watertight, we do not report the intersection volume metric.

Comparisons with the State-of-the-Art Methods

We evaluate reconstructed 3D meshes on HO3D and HOD. Averaged quantitative results are presented in Table 1. Please refer to the Supp. for more detailed results.

Comparison Results on HO3D. We visualize the reconstructed objects in Fig. 5. The learning-based method IHOI can predict the coarse shape, but it typically loses the finer details of the object surface when compared to neural rendering methods. Inaccurate camera poses significantly decrease the reconstruction quality of NeuS, but when ground-truth poses are used (NeuS*), it achieves similar quality to HHOR in the visible part of the object. However, both NeuS and HHOR struggle to handle occlusion, which leads to incomplete surface reconstructions. While HHOR** (HHOR with post-processing) can use Poisson Reconstruction to aid in filling the holes, the resulting object reconstruction may contain obvious artifacts. This is because the Poisson Reconstruction can not correctly fill the surface for the missing part when a large part of the object is occluded. In contrast,

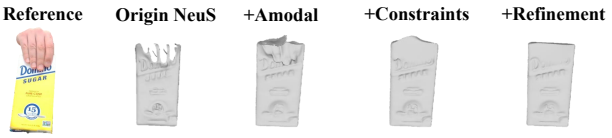

Figure 7: Qualitative results of ablation study. +Amodal indicates the utilization of amodal mask supervision, +Constraints indicates the incorporation of $\mathcal{L}_{contact}$, and +Refinement indicates the application of mask refinement.

NeuS	Amodal	Constraints	Refinement	CD↓
✔				1.094
✔	✔			0.448
✔	✔	✔		0.333
✔	✔	✔	✔	0.282

Table 2: Ablation studies of each component of our method over 5 objects on the HO3D dataset. We report the averaged 3D object reconstruction metric.

our method can recover detailed object meshes in both the visible and invisible parts without any post-processing.

By analyzing quantitative results on HO3D, our method significantly outperforms the comparison methods in both 3D reconstruction and hand-object relationships. Other volume rendering-based methods, NeuS with ground-truth poses and HHOR achieve better performance of reconstruction than the learning-based method IHOI. However, they still struggle in reconstructing complete geometry. When HHOR** can obtain complete surfaces with 0.591 CD, our approach achieves even lower 0.282 CD values, demonstrating an improvement of 52%. In terms of intersection volume, our approach outperforms IHOI and significantly surpasses HHOR**. This can be attributed to the evident hand artifacts in HHOR** that lead to increased volume, highlighting the effectiveness of our integrated contact constraints.

Comparison Results on HOD. HOD contains objects with more complex shapes and textures, but less hand occlusion. For reconstruction quality, our method continues to surpass the state-of-the-art methods. Compared to HHOR**, our method improves by 20%. Furthermore, visualizations in Fig. 6 highlight that other results still contain obvious artifacts, whereas our outcomes appear more reasonable. Our method can reconstruct a complete and detailed object mesh regardless of whether the hand-grasping type involves weak or heavy occlusion. Note that the learning-based method heavily relies on the learned prior, and therefore does not work well for objects beyond the training dataset. They cannot recover the shape accurately. Regarding the hand-object relationship, our method outperforms HHOR**, emphasizing contact constraints' importance in less occluded scenarios. Though IHOI results in lower intersection volume, their predicted object shapes are absolutely inaccurate. Conversely, our method can reconstruct detailed object meshes with a reasonable hand-object relationship.

Methods	CD↓	Vol↓
Ours w/o $\mathcal{L}_A$	0.376	**0.108**
Ours w/o $\mathcal{L}_P$	0.319	0.614
Ours w/ ${\mathcal{L}_A}^-$	0.341	0.439
Ours	**0.282**	0.327

Table 3: Ablation on $\mathcal{L}_{contact}$ variants.

Ablation Studies

To evaluate the effectiveness of our proposed components, we perform experiments on HO3D across four distinct settings: (1) NeuS; (2) NeuS with amodal masks; (3) NeuS with amodal masks and contact constraints; (4) Ours: NeuS with amodal masks, contact constraints, and mask refinement.

According to Table 2, incorporating amodal masks significantly improves reconstruction quality, reducing the CD value by 0.646. Visualizations in Fig. 7 demonstrate successful recovery of overall shape, indicating that amodal masks effectively fuse observations for complete reconstruction.

With contact constraints, we can further improve the geometry quality with 0.115 CD value decrement. As demonstrated in Fig. 7, the utilization of $\mathcal{L}_{contact}$ effectively reduces the ambiguities caused by occlusion in the contact region, resulting in a smoother surface.

Moreover adding mask refinement can remove the wrongly estimated part caused by inaccurate amodal masks, leading to a 0.051 CD value reduction. The visualization results illustrate that the mask refinement effectively removes wrong results at the object boundaries. These results demonstrate the effectiveness of our proposed components.

Different $\mathcal{L}_{contact}$ Design. We analyze the results with various $\mathcal{L}_{contact}$ designs in Table 3. We conduct experiments without attraction or penetration loss. Incorporating only penetration loss minimizes intersection volume, yet its reconstruction quality lags behind other methods. Conversely, solely applying attraction loss increases intersection volume while enhancing reconstruction. To balance reconstruction quality and intersection volume, we simultaneously apply these two losses in our method. Furthermore, we conducted a comparison by substituting our formulated $\mathcal{L}_A$ with constraints on object surface points in contact only (denoted as ${\mathcal{L}_A}^-$). Our approach reaches lower CD values and intersection volume, demonstrating the efficacy of guidance in the vicinity of the hand but not in the contact area.

Conclusions and Future Work

In this paper, we present a framework for reconstructing the 3D generic objects in hand using a monocular RGB video. The key insights of our method are to incorporate the amodal masks and physical contact guidance for dealing with surface reconstruction under occlusion. On several datasets, we have demonstrated state-of-the-art results compared with existing methods. In the future, we aim to speed up the training process by integrating hybrid neural representations such as (Müller et al. 2022), and relax the assumption of fixed grasping by inferring the object pose (Wen et al. 2023).

Acknowledgments

This work was supported in part by NSFC under Grants (62103372, 62088101, 62233013), the Fundamental Research Funds for the Central Universities (226-2023-00111), and the OPPO Research Fund.

References

Back, S.; Lee, J.; Kim, T.; Noh, S.; Kang, R.; Bak, S.; and Lee, K. 2022. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *2022 International Conference on Robotics and Automation (ICRA)*, 5085–5092. IEEE.

Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3): 24.

Boerdijk, W.; Sundermeyer, M.; Durner, M.; and Triebel, R. 2021. "What's This?"-Learning to Segment Unknown Objects from Manipulation Sequences. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 10160–10167. IEEE.

Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. M. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, 510–517. IEEE.

Cao, Z.; Radosavovic, I.; Kanazawa, A.; and Malik, J. 2021. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12417–12426.

Chen, J.; Yan, M.; Zhang, J.; Xu, Y.; Li, X.; Weng, Y.; Yi, L.; Song, S.; and Wang, H. 2022a. Tracking and Reconstructing Hand Object Interactions from Point Cloud Sequences in the Wild. *arXiv preprint arXiv:2209.12009*.

Chen, Z.; Hasson, Y.; Schmid, C.; and Laptev, I. 2022b. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, 231–248. Springer.

Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; and Tan, R. T. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, 723–732.

Cignoni, P.; Callieri, M.; Corsini, M.; Dellepiane, M.; Ganovelli, F.; Ranzuglia, G.; et al. 2008. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, 129–136. Salerno, Italy.

Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 216–224.

Hampali, S.; Hodan, T.; Tran, L.; Ma, L.; Keskin, C.; and Lepetit, V. 2022. In-Hand 3D Object Scanning from an RGB Sequence. *arXiv preprint arXiv:2211.16193*.

Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3196–3206.

Hasson, Y.; Varol, G.; Schmid, C.; and Laptev, I. 2021. Towards unconstrained joint hand-object reconstruction from RGB videos. In *2021 International Conference on 3D Vision (3DV)*, 659–668. IEEE.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11807–11816.

Huang, D.; Ji, X.; He, X.; Sun, J.; He, T.; Shuai, Q.; Ouyang, W.; and Zhou, X. 2022. Reconstructing Hand-Held Objects from Monocular Video. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Karunratanakul, K.; Yang, J.; Zhang, Y.; Black, M. J.; Muandet, K.; and Tang, S. 2020. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, 333–344. IEEE.

Kazhdan, M.; Bolitho, M.; and Hoppe, H. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 0.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5741–5751.

Lorensen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics*, 21(4): 163–169.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Lv, J.; Xu, W.; Yang, L.; Qian, S.; Mao, C.; and Lu, C. 2021. HandTailor: Towards high-precision monocular 3d hand recovery. *arXiv preprint arXiv:2102.09244*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 483–499. Springer.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5589–5599.

Park, J.; Oh, Y.; Moon, G.; Choi, H.; and Lee, K. M. 2022. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1496–1505.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6).

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 501–518. Springer.

Tran, M.; Vo, K.; Yamazaki, K.; Fernandes, A.; Kidd, M.; and Le, N. 2022. AISFormer: Amodal Instance Segmentation with Transformer. *arXiv preprint arXiv:2210.06323*.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *arXiv preprint arXiv:2106.10689*.

Wen, B.; Tremblay, J.; Blukis, V.; Tyree, S.; Muller, T.; Evans, A.; Fox, D.; Kautz, J.; and Birchfield, S. 2023. BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects. *arXiv preprint arXiv:2303.14158*.

Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2750–2760.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34: 4805–4815.

Ye, Y.; Gupta, A.; and Tulsiani, S. 2022. What's in your hands? 3D Reconstruction of Generic Objects in Hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3895–3905.

Yin, Z.-H.; Huang, B.; Qin, Y.; Chen, Q.; and Wang, X. 2023. Rotating without Seeing: Towards In-hand Dexterity through Touch. *arXiv preprint arXiv:2303.10880*.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3784–3792.

Zhang, H.; Bo, Z.-H.; Yong, J.-H.; and Xu, F. 2019. InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. *ACM Transactions on Graphics (TOG)*, 38(4): 1–11.

Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. J. 2021. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15838–15847.

Zhou, Q.; Wang, S.; Wang, Y.; Huang, Z.; and Wang, X. 2021. Human de-occlusion: Invisible perception and recovery for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3691–3701.