

MOHO: Learning Single-view Hand-held Object Reconstruction with Multi-view Occlusion-Aware Supervision

Chenyanguang Zhang^{1*}, Guanlong Jiao^{1*}, Yan Di², Gu Wang¹, Ziqin Huang¹,
Ruida Zhang¹, Fabian Manhardt³, Bowen Fu¹, Federico Tombari^{2,3} and Xiangyang Ji¹

¹Tsinghua University, ²Technical University of Munich, ³ Google

{zcyg22@mails., xyji@}tsinghua.edu.cn * †

Abstract

Previous works concerning single-view hand-held object reconstruction typically rely on supervision from 3D ground-truth models, which are hard to collect in real world. In contrast, readily accessible hand-object videos offer a promising training data source, but they only give heavily occluded object observations. In this paper, we present a novel synthetic-to-real framework to exploit **Multi-view Occlusion-aware supervision** from hand-object videos for **Hand-held Object reconstruction (MOHO)** from a single image, tackling two predominant challenges in such setting: hand-induced occlusion and object’s self-occlusion. First, in the synthetic pre-training stage, we render a large-scaled synthetic dataset SOMVideo with hand-object images and multi-view occlusion-free supervisions, adopted to address hand-induced occlusion in both 2D and 3D spaces. Second, in the real-world finetuning stage, MOHO leverages the amodal-mask-weighted geometric supervision to mitigate the unfaithful guidance caused by the hand-occluded supervising views in real world. Moreover, domain-consistent occlusion-aware features are amalgamated in MOHO to resist object’s self-occlusion for inferring the complete object shape. Extensive experiments on HO3D and DexYCB datasets demonstrate 2D-supervised MOHO gains superior results against 3D-supervised methods by a large margin.

1. Introduction

Understanding hand-object interaction is becoming increasingly important in many practical scenarios including robotics [31, 70], augmented and virtual reality [29], as well as embodied artificial intelligence systems [30, 60]. Although there exist previous works [8, 25, 62, 64] aiming at reconstructing fine hand-held object meshes from multi-view image sequences, single-view methods [12, 13, 23, 28, 63,

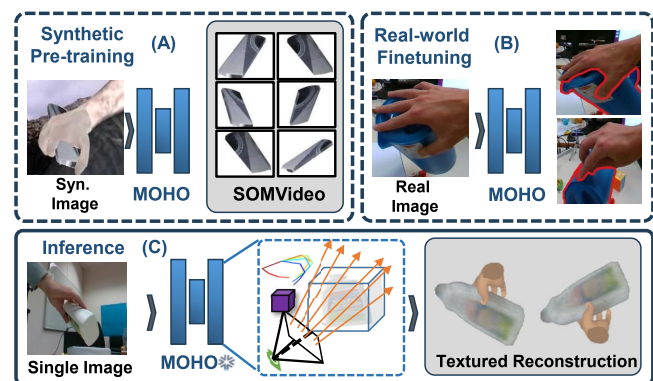


Figure 1. As a synthetic-to-real framework, MOHO is pre-trained by the rendered occlusion-free supervisions on SOMVideo, and then finetuned by the real-world hand-occluded supervising views. In the inference stage, MOHO generates the photorealistic reconstructed mesh given a single reference view, resisting both hand-induced occlusion and object’s self-occlusion.

[66] are drawing more attention recently since they can be applied more conveniently in real-world environment.

Given the ill-posed nature of single-view reconstruction, current top-performing methods [12, 13, 63] typically use Signed Distance Fields (SDFs) as the geometric representation and employ 3D ground-truth meshes as supervision for training. However, the applicability of such approaches in real-world scenarios is highly challenging, as obtaining clean and precise object meshes remains a formidable task. In contrast, readily accessible raw videos capturing hands interacting with objects offer a promising training data source. Nevertheless, leveraging these videos as multi-view supervision for single-view hand-held object reconstruction introduces two significant challenges: hand-induced occlusion and object’s self-occlusion. Firstly, hand-induced occlusion is an unavoidable issue in our easily obtained training data, leading to frequent instances of incomplete object views as objects are manipulated by hands. This incompleteness poses a significant hurdle for the network in effectively learning the reconstruction of the complete object shape. Thus, we

* * Authors with equal contributions.

† Codes and datasets: <https://github.com/ZhangCYG/MOHO>

adopt additional occlusion-free information from synthetic environments to mitigate the unfaithful guidance caused by the occluded supervising views in real world. Additionally, the single-view setting exacerbates the problem with object’s self-occlusion, as only one reference view is available, leaving the visible portion of the object incomplete and further complicating the task of enabling the network to recover the object’s full shape. Therefore, the occlusion-aware features need to be imposed for the network for full reconstruction.

To address the aforementioned problems, we present a novel synthetic-to-real framework to exploit **Multi-view Occlusion-aware** supervision from hand-object videos for single-view **Hand-held Object** reconstruction (MOHO). First, in the synthetic pre-training stage, we render a large-scale synthetic dataset SOMVideo with hand-object images and multi-view occlusion-free supervisions (Fig. 1 (A)). MOHO takes one hand-object image as input and the other occlusion-free image describing the complete object in a novel view as supervision. Thus, MOHO is empowered to remove hand-induced occlusion in 3D space. Simultaneously, an auxiliary 2D amodal mask recovery head is integrated into the pre-training process, which predicts the hand-occluded parts of the object in the reference view. Second, in the real-world finetuning stage, we freeze the 2D amodal mask recovery head to establish the amodal-mask-weighted geometric supervision, designed to combat the incomplete and defective supervisions presented by real-world hand-occluded videos (Fig. 1 (B)). Moreover, in order to overcome object’s self-occlusion in the whole synthetic-to-real process, we leverage domain-consistent occlusion-aware features including generic semantic cues and hand-articulated geometric embeddings. These features are obtained with small cross-domain discrepancy, indicating which portions are visible in the reference view as well as hallucinating the shape of the self-occluded object surfaces. Consequently, MOHO recovers complete 3D shape with photorealistic textures of the hand-held object via the geometric volume rendering technique during real-world inference (Fig. 1 (C)).

To summarize, our main contributions are threefold:

- We propose a synthetic-to-real framework MOHO to pursue photorealistic hand-held object reconstruction from a single-view image without relying on 3D ground-truth supervision. To mitigate hand-induced occlusion, the rendered SOMVideo is adopted in the synthetic pre-training stage for occlusion-free supervisions, while the amodal-mask-weighted geometric supervision is proposed during the real-world finetuning.
- The domain-consistent occlusion-aware features are exploited in order to overcome object’s self-occlusion in the whole synthetic-to-real process.
- Extensive experiments on real-world datasets HO3D [22] and DexYCB [8] demonstrate that 2D-supervised MOHO gains superior results against 3D-supervised methods.

2. Related Works

Hand and Object Pose Estimation. The separate regression of hand pose and object pose constitutes a methodological stream for reconstructing hand-held objects. Hand pose estimation from RGB(-D) input can be broadly categorized into two streams: model-free methods that lift detected 2D keypoints to 3D joint positions [26, 38–40, 43, 44, 72], as well as model-based approaches that estimate statistical models with low-dimensional parameters [3, 45, 46, 49, 69, 71]. On the other hand, many works focus on initially regressing object poses based on predefined object templates [15, 16, 52, 55, 56], and subsequently, they proceed to reconstruct object meshes. In contrast, MOHO stands apart by its capability to reconstruct agnostic objects without relying on any prior assumptions. Our approach adopts the MANO model [45], which shows more robustness to occlusion [63], to provide hand articulations.

Hand-held Object Reconstruction. Hand-held object reconstruction plays a crucial role in advanced understanding of human-object interaction. Previous works [17, 22, 50, 51] typically assume access to predefined object templates, employing joint regression techniques to estimate both hand poses and 6DoF object poses. Some studies [11, 18, 34, 47] explore implicit feature fusion, incorporating geometric constraints [4, 5, 14, 19, 67] or promoting physical realism [42, 53] for such joint reasoning. Recent studies have shifted their focus toward directly reconstructing hand-held object meshes from monocular RGB inputs without relying on any prior assumptions. For instance, [23] develops a joint network that predicts object mesh vertices and MANO parameters of the hand, while [28] predicts these parameters within a latent space. Additionally, [12, 13, 63, 66] leverage Signed or Directed Distance Field (S/DDF) representations for hand and object shapes. However, these methods require hard-to-collect 3D ground-truth data for training, limiting their applicability in real-world scenarios. Contrarily, MOHO alleviates the need for 3D ground-truth by exclusively utilizing 2D video supervision.

Volume Rendering Techniques. There has been a surge in the utilization of volume rendering techniques in the context of neural radiance fields (NeRFs) [2, 32, 36, 41, 57], which have proven to be influential for advancing novel view synthesis and photorealistic scene reconstruction. In the earlier stages, the focus is primarily on the development of scene-specific volume rendering [36, 57], where one network can only represent a single scene. Subsequent studies [9, 27, 33, 59, 65] extend the scope of the problem to scene-agnostic, focusing on reconstruction of various objects from one single reference view or sparse views, which closely resembles the setting of single-view hand-held object reconstruction. However, the performance of these previous methods is largely contingent on ideal conditions, where occlusion is not a significant factor. MOHO addresses this

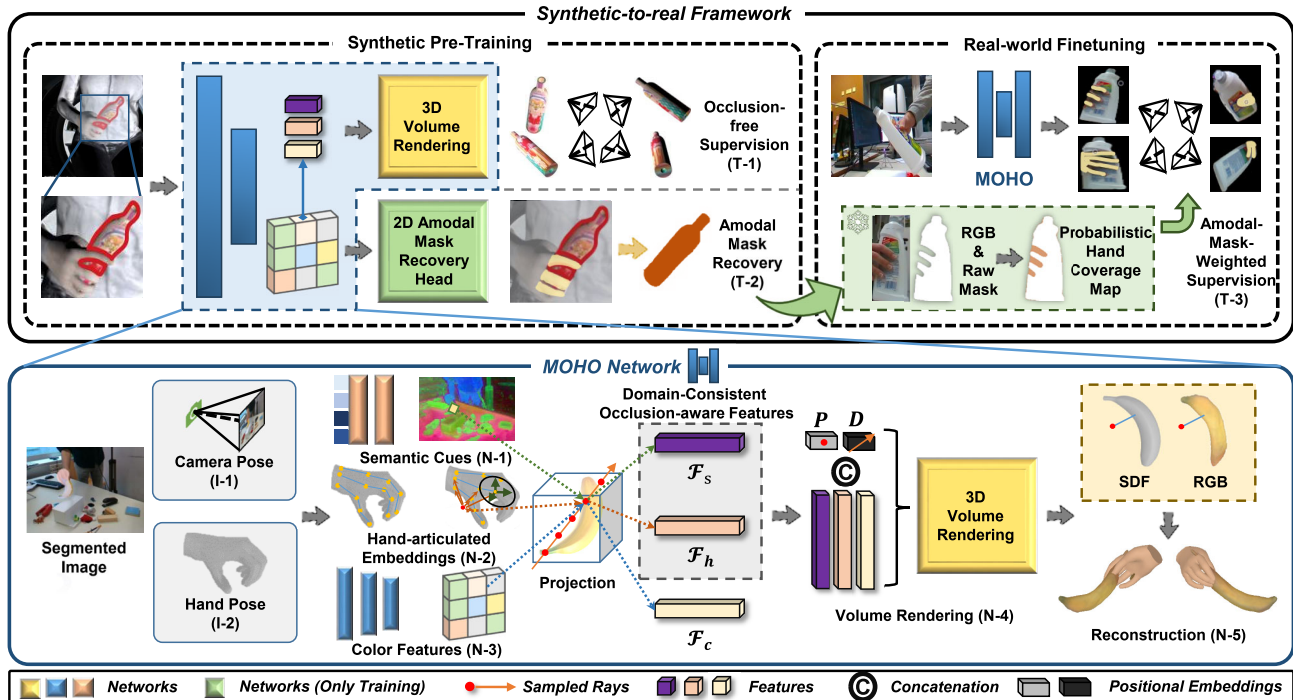


Figure 2. **Overview of MOHO. Synthetic-to-real Framework:** We pre-train MOHO on the SOMVideo to resist hand-induced occlusion in both 3D (T-1) and 2D (T-2) spaces. The 2D recovered amodal masks are transferred into the real-world finetuning for releasing the incomplete hand-occluded supervisions (T-3). **Network:** Given a segmented hand-object image as input, the estimated camera pose (I-1) and hand pose (I-2) are initialized by an off-line system [46]. Subsequently, MOHO extracts domain-consistent occlusion-aware features including generic semantic cues (N-1) and hand-articulated geometric embeddings (N-2), as well as color features (N-3) for the volume rendering heads (N-4) to yield the textured mesh reconstruction of the full hand-held object (N-5).

limitation by introducing a novel synthetic-to-real framework and leveraging domain-consistent occlusion-aware features. These additions aim to effectively mitigate the challenges posed by heavy occlusion in real world.

3. Method

3.1. Overview

As is shown in Fig. 2, MOHO, following the synthetic-to-real paradigm, is first pre-trained by the large-scaled rendered dataset SOMVideo for gaining the removal ability against hand-induced occlusion in both 3D and 2D spaces ((T-1) and (T-2)), and then finetuned on real-world videos with incomplete object observations. We adopt the proposed amodal-mask-weighted geometric supervision (T-3) to mitigate the misguidance caused by heavily occluded real-world supervisions in the finetuning stage. Meanwhile, the domain-consistent occlusion-aware features are leveraged in the whole synthetic-to-real process, which include generic semantic cues \mathcal{F}_s extracted by the pre-trained DINO [1] model \mathcal{D} (N-1), and hand-articulated geometric embeddings \mathcal{F}_h calculated by the predicted hand pose θ_A (N-2). These domain-consistent occlusion-aware features, as well as color features \mathcal{F}_c yielded by the CNN-based encoder ϕ (N-3), are con-

catenated as the condition for geometric volume rendering heads [57] ψ_S, ψ_C (N-4) to respectively predict the SDF value and the color density, enabling reconstruction of the agnostic occluded hand-held object without any instance priors. During inference, given a single input reference image \mathcal{I} depicting hand-object interaction, its corresponding camera pose \mathcal{P}_I (I-1), offline-estimated object segmentation \mathcal{S}_o , and hand pose prediction θ_A (I-2), MOHO synthesizes the novel views as well as reconstructs the textured mesh of the complete hand-held object (N-5).

3.2. Domain-consistent Occlusion-aware Features for Geometric Volume Rendering

Generating novel views and the full mesh for a hand-held object given by the only reference view is typically an ill-posed problem due to severe object’s self-occlusion. As [59] demonstrates, reconstructing the whole scene from a single reference view may cause the volume rendering technique to generate unsatisfactory results, whose surface toward the reference view is recovered decently, but the reconstruction of unseen parts is degraded. Hence, we need to feed sufficient information to the network for completing the unobserved area in the reference image. Such imposed information should have least domain discrepancy, to ensure

effective knowledge transfer in the whole synthetic-to-real framework. Specifically, we exploit the domain-consistent occlusion-aware features from two aspects: generic semantic cues \mathcal{F}_s (Fig. 2 (N-1)) and hand-articulated geometric embeddings \mathcal{F}_h (Fig. 2 (N-2)).

Generic Semantic Cues. The generic semantic cues \mathcal{F}_s are exploited to provide MOHO with high-level structural priors for amodal object perception. Concretely, we harness semantic cues from the pre-trained DINO [1] model, which provides local descriptors with consistent structural information to indicate the position of the observed object parts within the whole shape. Note that DINO is well demonstrated for its semantic stability across different domains [1]. With such domain-consistent semantic cues, MOHO learns to complement the full object better from the partial observation under the whole synthetic-to-real process. Specifically, the pre-trained DINO model \mathcal{D} extracts the patch-wise feature maps $\mathcal{F}_s^I = \mathcal{D}(\mathcal{I}) \odot \mathcal{S}_o$. We use the top three principal components of \mathcal{F}_s^I by principal component analysis (PCA) considering the trade-off between efficiency and performance. Since MOHO adopts the volume rendering technique, the feature map \mathcal{F}_s^I needs to be converted to the features corresponding to the 3D sampled points. Given the camera poses \mathcal{P}_I and camera parameters \mathcal{K} , the 3D points $\{P_i\}_{i=1}^n$ along the sampled rays are first projected onto the image plane to get the corresponding pixel positions. The patched color features of each sampled point \mathcal{F}_s^i are fetched on \mathcal{F}_s^I via bilinear interpolation.

Hand-articulated Geometric Embeddings. Considering that the holding hand shape implies the unobserved hand-held object shape, we add the hand-articulated geometric embeddings \mathcal{F}_h^i for each sampled point P_i . The adopted embeddings are explicitly yielded by calculating the geodesic distances from the sampled point P_i to the nearest hand joints. Such explicit embeddings remain stable and consistent during the whole synthetic-to-real transferring without any domain gaps. Specifically, we first use an offline hand pose estimator [46] to get θ_A from the reference image. Then, we run forward kinematics of MANO model [45] to derive the transformation $T(\theta_A)$ as well as the hand joint coordinates. Afterwards, the sampled point P_i is mapped to the nearest K hand joint coordinates by their transformation matrices. Finally, such K positions of the sampled point in the nearest K hand joint coordinates are concatenated as \mathcal{F}_h^i to provide the distance information. We select $K = 6$ during the implementation. Notably, we treat the hand articulation locally using nearest hand joints rather than globally using all hand joints as the previous methods [13, 63] do. We find that taking all the joints' coordinates is unnecessary and leads to more complexity empirically (Sec. 4.4).

Color Features. Since MOHO needs to recover the texture of the hand-held object, we follow [65] to use a ResNet34 [24] as the image encoder ϕ to extract the ob-

ject color feature map by $\mathcal{F}_c^I = \phi(\mathcal{I} \odot \mathcal{S}_o)$ of the reference view \mathcal{I} . The \mathcal{F}_c^i of the sampled 3D point P_i is obtained by the same projection and interpolation operations as the generation of \mathcal{F}_s^i .

Conditional Geometric Volume Rendering. After incorporating all the domain-consistent occlusion-aware features, we construct the conditional geometric volume rendering technique to render novel views as well as generate textured meshes (Fig. 2 (N-4)). Specifically, given 3D sampled points $\{P_i\}$, ray directions $\{D_i\}$ and corresponding point features $\{\mathcal{F}_{con}^i\} = \{\text{Cat}(\mathcal{F}_s^i, \mathcal{F}_h^i, \mathcal{F}_c^i)\}$ extracted from the single reference view \mathcal{I} , a geometric field ψ_S predicting the SDF value $s_i = \psi_S(P_i | \mathcal{F}_{con}^i)$ as well as a color field ψ_C predicting the RGB density $c_i = \psi_C(P_i, D_i | \mathcal{F}_{con}^i)$ are constructed. As for volume rendering, 3D points $\{P_i\}$ are sampled along camera rays by $\{P\} = \{P(z) | P(z) = O + zD, z \in [z_n, z_f]\}$, where O denotes the origin of the camera, D refers to the viewing direction of each pixel, z_n, z_f are the near and far bounds of the ray. O and D are calculated by the input camera pose \mathcal{P}_I and camera intrinsic \mathcal{K} . Then, the color of the pixel is rendered by

$$\hat{c} = \int_{z_n}^{z_f} \omega(z) \psi_C(P(z), D | \mathcal{F}_c) dz, \quad (1)$$

where $\omega(z) = T(z)\rho(z)$ is an unbiased and occlusion-aware function proposed by [57], converting the SDF value $\psi_S(P | \mathcal{F}_{con})$ to $T(z), \rho(z)$ by $T(z) = \exp\left(-\int_{z_n}^{z_f} \rho(z) dz\right)$, $\rho(z) = \max\left(\frac{-\frac{d\sigma^h}{dz}(\psi_S(P(z) | \mathcal{F}_{con}))}{\sigma^h(\psi_S(P(z) | \mathcal{F}_{con}))}, 0\right)$. σ^h here denotes the Sigmoid function with a trainable parameter h . During implementation, the formulations above are numerically discretized as referred to [57].

3.3. Synthetic-to-real Training Framework

We propose a synthetic-to-real training framework for MOHO to overcome the omnipresent hand-induced occlusion met in real-world single-view hand-held object reconstruction. In the synthetic pre-training stage, we foster MOHO for the capability to be aware of the hand-occluded regions of the object in both 3D and 2D spaces, with the utilization of our large-scaled rendered dataset SOMVideo. For removing hand-induced occlusion in 3D space, MOHO inputs the hand-occluded reference view $\mathcal{I} \odot \mathcal{S}_o$ (\odot means bitwise multiplication), and is supervised by the synthetic complete object in novel views (Fig. 2 (T-1)). Further, an auxiliary 2D amodal mask recovery head Γ (Fig. 2 (T-2)) is utilized to predict the probabilistic hand coverage map in 2D space. After pre-training, MOHO is finetuned with real-world hand-object videos, so as to be better applied for real-world inference. However, real-world hand-object multi-view images oftentimes contain truncated regions and incomplete views, resulting in detrimental effects when directly used for training. Naively utilizing the defective masks

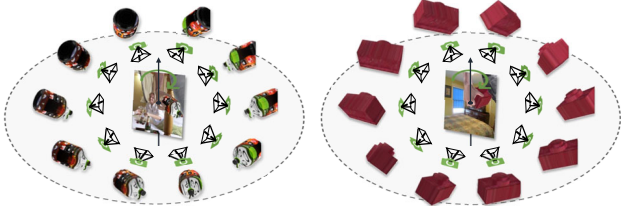


Figure 3. Visual illustration of SOMVideo rendered with occlusion-free multi-view supervisions.

of the hand-occluded objects misleads the network to reconstruct patchy geometric surfaces. Thus, the predicted hand coverage maps from the pre-trained 2D head are leveraged on real-world data (Fig. 2 (T-3)) to construct a soft constraint and introduce the amodal-mask-weighted geometric supervision for reconstructing full hand-held objects.

Synthetic Object Manipulation Video (SOMVideo) Dataset. Current video datasets capturing hand-object interactions [8, 22] are collected in real world. They typically contain limited object instances and are unfriendly for constructing occlusion-free supervisions for purely 2D-supervised methods. Therefore, following the generation pipeline of synthetic object manipulation scenes [23], we render the Synthetic Object Manipulation Video (SOMVideo) dataset for MOHO, boasting large-scaled hand-object images as well as corresponding occlusion-free multi-view supervisions (Fig. 3).

Generally, SOMVideo is synthesized by setting up the hand-object interaction scene and moving the camera. First, we select the same 2,772 objects as the ObMan [23] dataset from ShapeNet [7]. Objects’ textures are randomly sampled from the texture maps provided by ShapeNet. Then, the grasping generation procedure adopting GraspIt [37] and the body and hand models in the SMPL+H [35, 45] are used for the hand-object interaction scene setup. Finally, we add rotations and translations on the camera towards the hand-object interaction scene to capture hand-object multi-view images. Besides, to enable occlusion-free supervisions, we render corresponding video clips containing only objects without hands and bodies by setting the scene without SMPL+H models and keeping the camera parameters the same. SOMVideo consists of 141,550 scenes in total, in which each hand-object scene is captured by 10 views. Each corresponding occlusion-free video clip for supervision is also captured from 10 same view angles. For more rendering details, please refer to the supplementary material.

Synthetic Pre-Training with SOMVideo. MOHO is first pre-trained to remove hand-induced occlusion in 3D space by the rendered occlusion-free supervisions. Concretely, the input reference view fed to MOHO is a hand-object image, while the supervision views are corresponding rendered occlusion-free pictures from novel poses (Fig. 2 (T-1)). During each iteration in the pre-training, one hand-object reference image is fed to MOHO, while 8 novel views sampled

from the corresponding occlusion-free video are regarded as supervision.

Simultaneously, an auxiliary 2D amodal mask recovery head Γ is utilized in the pre-training. The 2D recovery head, whose architecture refers to [10], predicts the hand-covered object’s parts \hat{M}_I^{ho} in the input reference view \mathcal{I} by $\hat{M}_I^{ho} = \Gamma(\mathcal{F}_c^I)$ (Fig. 2 (T-2)), where \mathcal{F}_c^I is the color feature maps defined in Sec. 3.2. The supervision of this head is enforced by the binary cross-entropy loss between \hat{M}_I^{ho} and $M_I^{co} \ominus M_I$, where M_I^{co} means the rendered complete mask of the input reference view \mathcal{I} , M_I refers to the input hand-occluded object mask, and \ominus means bitwise subtraction. The benefits of incorporating such 2D perception are twofold. First, 2D hand coverage perception strengthens the ability of MOHO to handle hand-induced occlusion patterns. Second, considering more cross-domain consistency of the 2D neural network [6], we exploit the predictions of this 2D head for the real-world finetuning stage to promote the knowledge transfer about hand-induced occlusion removal learned in the pre-training stage. To this end, we freeze the 2D recovery head during the real-world finetuning stage, and infer the probabilistic hand coverage maps (Fig. 2 (T-3)). These maps are regarded as the relaxed constraints for the proposed amodal-mask-weighted geometric supervision.

Real-World Finetuning with Amodal-Mask-Weighted Supervision. After pre-training, MOHO is finetuned on hand-object videos from real-world datasets which typically suffer from partial observations caused by hand-induced occlusion. Therefore, we introduce the amodal-mask-weighted geometric supervision, taking the probabilistic hand coverage maps predicted by the pre-trained 2D amodal mask recovery head in real world into consideration (Fig. 2 (T-3)). The amodal-mask-weighted loss is defined by

$$\mathcal{L}_{amw} = BCE(\hat{M}_T^{ho} \oplus M_T, \hat{O}_T), \quad (2)$$

where T refers to a novel target view, \hat{M}_T^{ho} means the recovered amodal mask, \hat{O}_T is the predicted object mask by the volume rendering heads and \oplus means bitwise addition. Having gained knowledge about how to handle hand occlusion with such supervision, MOHO is capable of inferring the shape of the complete object in real world.

Volume Rendering Losses for Synthetic-to-real Training. Several losses are designed for supervising the 3D volume rendering heads of MOHO to get geometric consistent surfaces as well as photorealistic texture results in the whole synthetic-to-real framework. The overall loss function is defined as

$$\mathcal{L} = \mathcal{L}_{color} + \lambda_1 \mathcal{L}_{eik} + \lambda_2 \mathcal{L}_{mask} + \lambda_3 \mathcal{L}_{nori} + \lambda_4 \mathcal{L}_{nsmo}. \quad (3)$$

Thereby, the color loss \mathcal{L}_{color} is derived by $\mathcal{L}_{color} = |\hat{C}_T - C_T|$, where \hat{C}_T and C_T mean the predicted and ground-truth color maps of a novel view T respectively.

The Eikonal term [20] $\mathcal{L}_{eik} = \frac{1}{n} \sum_i (|\nabla \psi_S(P_i)|_2 - 1)^2$ is added for geometric regularization, in which P_i refers to the sampled points for volume rendering and n is the number of sampling points. The mask loss \mathcal{L}_{mask} is defined differently in the pre-training stage and the finetuning stage, for exploiting the occlusion-free supervisions of SOMVideo and adopting the amodal-mask-weighted supervision in real world respectively. At the pre-training stage, the mask loss is defined as

$$\mathcal{L}_{mask} = BCE(M_T^{co}, \hat{O}_T), \quad (4)$$

where M_T^{co} refers to the occlusion-free object mask of a novel view T in SOMVideo. At the finetuning stage, the mask loss is substituted by $\mathcal{L}_{mask} = \mathcal{L}_{amw}$ defined in Eq. (2). Two additional losses regularizing the predicted surface normals are used for restricting the orientation of visible normals towards the camera ($\mathcal{L}_{n_{ori}}$) [54] and making the predictions smoother ($\mathcal{L}_{n_{smo}}$) [48], which are detailed in the supplementary material. The weighted factors are set to $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 10^3$, $\lambda_4 = 10^{-2}$ during implementation. All factors are kept the same in both the synthetic pre-training and real-world finetuning.

4. Experiment

4.1. Experimental Setup

Datasets. We conduct experiments on two representative real-world datasets capturing hand-object interactions, HO3D [22] and DexYCB [8]. HO3D [22] contains 77,558 images from 68 sequences with 10 different persons manipulating 10 different objects. The pose annotations are yielded by multi-camera optimization pipelines. We follow [63] to split training and testing sets. DexYCB [8] is currently one of the largest real-world hand-object video datasets [13]. We follow [12, 61] to concentrate on right-hand samples and use the official *s0* split. 29,656 training samples and 5,928 testing samples are downsampled referring to the setting of [13]. Note that for both datasets, MOHO only utilizes the RGB pictures, segmentations, and poses for training, but without any need for the 3D ground-truth meshes.

Baselines. 3D-supervised baselines including Atlas-Net-based [21] HO [23], implicit-field-based GF [28], SDF-based IHOI [63], AlignSDF [12] and gSDF [13] are adopted for geometric comparisons with MOHO. We mainly compare the reconstructed meshes with them to demonstrate the ability of MOHO for surface reconstruction. Moreover, several 2D-supervised object-agnostic NeRF-based baselines are implemented, including PixelNeRF [65] and the more recent SSDNeRF [9]. We follow their single-view reconstruction setting and use the same training data as MOHO. We report both geometric reconstruction and novel view synthesis metrics against the NeRF-based baselines.

Method	F-5 \uparrow	F-10 \uparrow	CD \downarrow
HO [23]	0.11	0.22	4.19
GF [28]	0.12	0.24	4.96
IHOI [63]	0.28	0.50	1.53
PixelNeRF [65]	0.17	0.32	6.91
SSDNeRF [9]	0.25	0.40	2.60
Ours	0.31	0.50	0.91

Table 1. Geometric results on HO3D [22] compared with 3D supervised methods (top) and 2D supervised methods (bottom).

Method	F-5 \uparrow	F-10 \uparrow	CD \downarrow
HO [23]	0.38	0.64	0.42
GF [28]	0.39	0.66	0.45
AlignSDF [12]	0.41	0.68	0.39
gSDF [13]	0.44	0.71	0.34
PixelNeRF [65]	0.25	0.46	0.94
SSDNeRF [9]	0.27	0.49	0.58
Ours w/o SYN	0.52	0.74	0.18
Ours	0.60	0.81	0.15

Table 2. Geometric results on DexYCB [8] compared with 3D supervised methods (top) and 2D supervised methods (bottom).

Evaluation Metrics. For geometric metrics, we follow [13, 63] to uniformly sample 30,000 points on the reconstructed mesh, and report mean Chamfer Distance (CD, mm) and F-score at thresholds of 5mm (F-5) and 10mm (F-10). For metrics of novel view synthesis, we randomly sample 10 images in each video as the input reference views and another 10 views as the target views for each input reference. We report average PSNR, SSIM [58], and LPIPS [68] of the whole video dataset. Only the region within the object mask is considered for the aim of object reconstruction.

Implementation Details. We train MOHO on a single NVIDIA A100 GPU using an Adam optimizer with a learning rate of 10^{-3} for synthetic pre-training and 4×10^{-4} for real-world finetuning. The learning rate is scheduled by the cosine decay to the minimum of 5×10^{-5} . In the pre-training, we randomly select one hand-object reference view as the network input and 8 occlusion-free target views for supervision at each iteration. In the finetuning, the reference view and target views are selected in real-world video data. We pre-train MOHO on SOMVideo for 300K iterations and the real-world finetuning stage continues for another 300K iterations. For volume rendering, we use the same coarse-to-fine ray sampling technique as [57] by first uniformly sampling 40 points along the ray and then upsampling another 40 points near the coarsely predicted surface. During training, we randomly sample 150 rays in the object bounding box for each picture, following the protocol of ray origin and direction sampling strategy of [65]. Our SOMVideo data will be released along with our codes. For more details about the network architecture and synthetic data generation, please refer to the supplementary material.

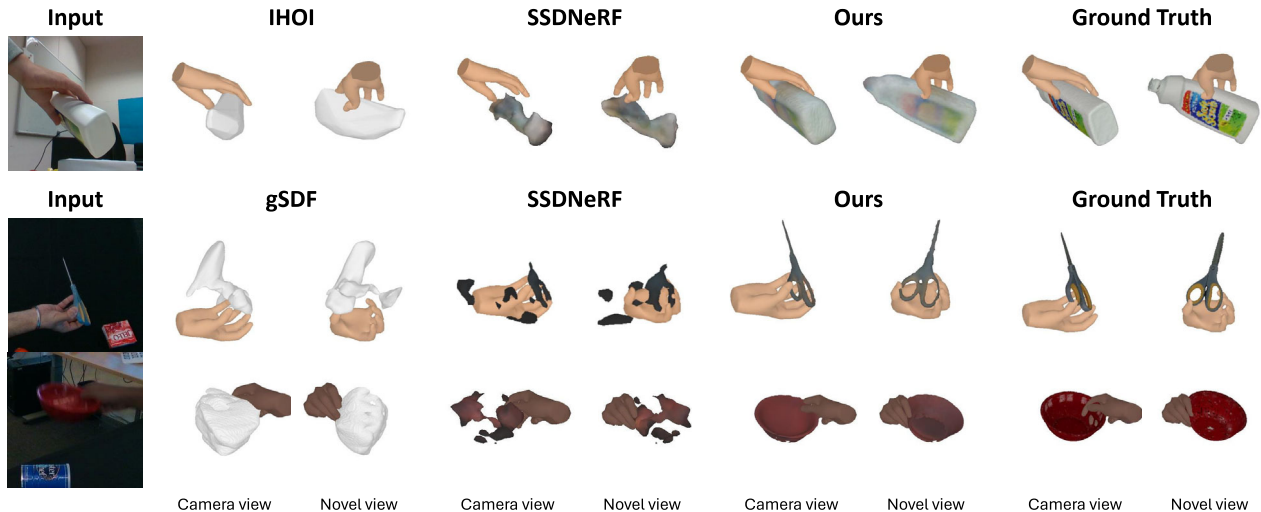


Figure 4. Visualization of textured meshes reconstructed by several baselines IHOI [63], gSDF [13], SSDNeRF [9] and MOHO on HO3D [22] (top) and DexYCB [8] (bottom). The reconstruction results are exhibited on the camera view and one novel view.

4.2. Geometric Reconstruction

We compare the quality of the geometric reconstruction ability of MOHO with two lines of methods including 3D-supervised baselines (typically SDF-based) and 2D-supervised baselines (typically object-agnostic NeRF-based). Tab. 1 and Tab. 2 exhibit the geometric metrics on HO3D [22] and DexYCB [8] respectively. In addition, we analyze the **efficiency** of MOHO in the supplementary material.

In Tab. 1, we report all baselines’ metrics following the setting of [63] which utilizes the synthetic-to-real paradigm to release the problem of data scarcity and lack of diversity in real world. All 3D-supervised methods in the top block strictly follow [63] to initialize with models pre-trained on ObMan [23] and then finetune on HO3D [22]. For 2D-supervised methods in the bottom block, we mimic the pre-training above using the hand-object multi-view images in SOMVideo for fair comparison. Compared to 2D-supervised baselines, MOHO exceeds PixelNeRF by 82.3% and the more recent SSDNeRF by 24.0% on the F-5 metric. This superiority demonstrates that due to the well-designed synthetic-to-real framework and the incorporation of the occlusion-aware features, MOHO better handles the extreme occlusion met when reconstructing hand-held objects from a single view. In contrast, the previous object-agnostic NeRF-based methods are only illustrated to be effective under the ideal occlusion-free condition. Compared with 3D-supervised methods, MOHO yields a significantly lower CD against the current top-performing approach IHOI by 40.5%, while leads by more against previous HO and GF. This indicates that the geometric surfaces reconstructed by MOHO contain much fewer outliers.

Experiments on the larger dataset DexYCB [8] shown in Tab. 2 further demonstrate the superiority of MOHO.

Method	HO3D [22]			DexYCB [8]		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
PixelNeRF [65]	24.82	0.955	0.055	32.77	0.986	0.019
SSDNeRF [9]	21.08	0.943	0.070	32.83	0.985	0.022
Ours	26.01	0.960	0.049	35.80	0.989	0.013

Table 3. Novel view synthesis results.

Due to the relatively sufficient data of DexYCB, the previous work [13] reports metrics of prior 3D supervised baselines [12, 13, 23, 28] directly trained on the real-world data. Thus, for fair comparison, we adopt this setting in Tab. 2 for both 3D and 2D supervised baselines and also report metrics of MOHO without the synthetic pre-training stage using SOMVideo (Ours w/o SYN). Thanks to the effective geometric volume rendering guided by the domain-consistent occlusion-aware features, MOHO w/o SYN achieves higher F-5 by 18.2% than the state-of-the-art 3D-supervised gSDF, while obtaining 92.6% superiority against the top-performing 2D supervised SSDNeRF on F-5 metric. When considering our proposed synthetic-to-real training framework, the lead margin is further extended to 36.4% and 122.2% respectively. This demonstrates that by considering the imposed features and training strategy together, MOHO is endowed with stronger robustness for handling various occlusion scenarios in real world.

4.3. Novel View Synthesis

As MOHO adopts the volume rendering technique, we also report its performance of novel view synthesis with other counterparts in Tab. 3, to demonstrate the capability to recover the object texture from the single view input. Notably, the previous predominant SDF-based methods [12, 13, 23, 28, 63] in the field cannot generate reconstruction results with object texture. However, MOHO can not only get geometrically coherent object surfaces,

Method	Pre-training	HO3D [22]		DexYCB [8]	
		F-5 \uparrow	CD \downarrow	F-5 \uparrow	CD \downarrow
IHOI [63]	\times	0.16	2.06	-	-
IHOI [63]	\checkmark	0.28	1.53	-	-
gSDF [13]	\times	-	-	0.44	0.34
gSDF [13]	\checkmark	-	-	0.46	0.30
PixelNeRF [65]	\times	0.13	14.07	0.25	0.94
PixelNeRF [65]	\checkmark	0.20	6.02	0.30	0.54
Ours	\times	0.22	1.18	0.52	0.18
Ours	\checkmark w/o AMW	0.29	0.99	0.57	0.16
Ours	\checkmark	0.31	0.91	0.60	0.15
		PSNR \uparrow	LPIPS \downarrow	PSNR \uparrow	LPIPS \downarrow
Ours	\times	24.96	0.052	35.41	0.014
Ours	\checkmark	26.01	0.049	35.80	0.013

Table 4. Ablations for the synthetic-to-real training framework.

\mathcal{F}_s -3	\mathcal{F}_s -16	\mathcal{F}_h -G	\mathcal{F}_h -L	F-5 \uparrow	F-10 \uparrow	CD \downarrow
\times	\times	\times	\times	0.52	0.72	0.18
\checkmark	\times	\times	\times	0.54	0.75	0.17
\times	\checkmark	\times	\times	0.55	0.76	0.16
\checkmark	\times	\checkmark	\times	0.59	0.79	0.15
\checkmark	\times	\times	\checkmark	0.60	0.81	0.15

Table 5. Ablations for the domain-consistent occlusion-aware features on DexYCB [8].

but also yield photorealistic object texture, which enables it to adapt in more application scenarios. Concretely, for novel view synthesis, MOHO leads by 4.8% on PSNR and 10.9% on LPIPS against PixelNeRF on HO3D and exceeds by 9.0% and 40.9% respectively compared with SSDNeRF on DexYCB, which illustrates the superior performance of MOHO than the NeRF-based competitors.

Visualization. To demonstrate both the abilities to get geometric surface as well as photorealistic texture, we visualize the textured meshes predicted by MOHO and compare them with 3D-supervised IHOI / gSDF and 2D-supervised SSDNeRF in Fig. 4. The results show that the 2D-supervised baseline typically fails and only yields incomplete and patchy reconstructed meshes when given heavily occluded real-world supervision (Row 2, 3). The 3D-supervised baselines obtain oversmoothed geometric surfaces without object texture. Moreover, they fail to reconstruct tiny objects like the scissors in Row 2. In contrast, MOHO is able to reconstruct geometrically coherent and photorealistic meshes. More visualization results are shown in the supplementary material.

4.4. Ablation Studies

We conduct ablation studies from two main aspects, *i.e.*, the effectiveness of the proposed synthetic-to-real training scheme (Tab. 4), and the generic semantic cues as well as the hand-articulated geometric embeddings included in the domain-consistent occlusion-aware features (Tab. 5). Additionally, we also analyze the zero-shot performance of MOHO and the sensitivity of the input hand pose predictions in the supplementary material.

We first exhibit the effects of the ObMan-based [23] 3D-

supervised synthetic-to-real training proposed by [63] on the top block of Tab. 4. Results show that such a strategy enhances the quality of geometric reconstruction (0.12 of F-5 for IHOI on HO3D), even on the larger-scale DexYCB (0.02 of F-5 for gSDF). Then, we conduct our proposed synthetic-to-real framework on 2D-supervised PixelNeRF to demonstrate its effectiveness. PixelNeRF directly trained on HO3D performs much inferiorly. However, after adopting our proposed synthetic-to-real framework, PixelNeRF gains superior results (0.07 on F-5). The same result is observed on DexYCB for PixelNeRF (0.05 enhancement on F-5). As for MOHO, the synthetic-to-real framework brings a boost of 0.09 and 0.08 of F-5 on two datasets respectively. When the amodal-mask-weighted geometric supervision (AMW) is removed, the performance of F-5 decreases by 0.02 and 0.03 on the two datasets respectively. Moreover, we find that the proposed framework also improves the novel view synthesis results due to suitable knowledge transfer.

Tab. 5 presents the ablations of the imposed domain-consistent occlusion-aware features. \mathcal{F}_s -16 means extending the PCA dimension of the semantic cues to 16. We find the performance enhancement is limited (0.01 of F-5). Thus, \mathcal{F}_s -3 setting is used for the MOHO implementation. Additionally, we compare the global hand-articulated embeddings adopting all hand joints with the local ones adopted in MOHO (Sec. 3.2). The local \mathcal{F}_h -L contributes to the performance improvement, since for the specific self-occluded part, the most credible heuristics implied by the holding hand come from the nearest joints.

5. Conclusion

This work has presented MOHO for single-view reconstruction of the hand-held object with multi-view occlusion-aware supervision from hand-object videos, tackling two predominant challenges of hand-induced occlusion and object’s self-occlusion. MOHO presents a novel synthetic-to-real paradigm to unleash hand-induced occlusion by adopting occlusion-free supervisions of SOMVideo in the synthetic pre-training and the amodal-mask-weighted geometric supervision in the real-world finetuning. Meanwhile, MOHO incorporates domain-consistent occlusion-aware features in order to overcome object’s self-occlusion in the whole synthetic-to-real process. Extensive experiments on HO3D and DexYCB datasets demonstrate that 2D-supervised MOHO gains superior results against 3D-supervised methods. In the future, we aim to adopt MOHO for robotic grasping in human-robot interaction scenes. Limitations are discussed in the supplementary material.

Acknowledgement This work was supported by the National Key R&D Program of China under Grant 2018AAA0102801. Moreover, we appreciate Zerui Chen, Yufei Ye and Haowen Sun for kind help.

References

- [1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. On the effectiveness of vit features as local semantic descriptors. In *ECCV*, pages 39–55. Springer, 2022. 3, 4
- [2] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. RenderDiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, pages 12608–12618, 2023. 2
- [3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 2
- [4] Samarth Brahmhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, pages 361–378. Springer, 2020. 2
- [5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pages 12417–12426, 2021. 2
- [6] Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Exploiting the complementarity of 2d and 3d networks to address domain-shift in 3d semantic segmentation. In *CVPR*, pages 98–109, 2023. 5
- [7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5
- [8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021. 1, 2, 5, 6, 7, 8
- [9] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 2, 6, 7
- [10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5
- [11] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE TIP*, 30:4008–4021, 2021. 2
- [12] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, pages 231–248. Springer, 2022. 1, 2, 6, 7
- [13] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *CVPR*, pages 12890–12900, 2023. 1, 2, 4, 6, 7, 8
- [14] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, pages 5031–5041, 2020. 2
- [15] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *ICCV*, pages 12396–12405, 2021. 2
- [16] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *CVPR*, pages 6781–6791, 2022. 2
- [17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, pages 409–419, 2018. 2
- [18] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pages 8359–8367, 2018. 2
- [19] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *CVPR*, pages 1471–1481, 2021. 2
- [20] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 6
- [21] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, pages 216–224, 2018. 6
- [22] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 2, 5, 6, 7, 8
- [23] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevtykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1, 2, 5, 6, 7, 8
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4
- [25] 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. 1
- [26] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 2
- [27] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *ICCV*, pages 12949–12958, 2021. 2
- [28] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, pages 333–344. IEEE, 2020. 1, 2, 6, 7
- [29] Zhiying Leng, Jiaying Chen, Hubert PH Shum, Frederick WB Li, and Xiaohui Liang. Stable hand pose estimation under tremor via graph neural network. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 226–234. IEEE, 2021. 1

- [30] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabriel Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *CoRL*, pages 80–93. PMLR, 2023. 1
- [31] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *NeurIPS*, 33:5011–5022, 2020. 1
- [32] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, pages 8456–8465, 2023. 2
- [33] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, pages 806–815, 2023. 2
- [34] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021. 2
- [35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6), 2015. 5
- [36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [37] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 5
- [38] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Generated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018. 2
- [39] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019.
- [40] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, pages 436–445. IEEE, 2018. 2
- [41] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *CVPR*, pages 4391–4401, 2023. 2
- [42] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE TPAMI*, 40(12):2883–2896, 2017. 2
- [43] Grégory Rogez, James S Supancic III, Maryam Khademi, Jose Maria Martinez Montiel, and Deva Ramanan. 3d hand pose detection in egocentric rgb-d images. *arXiv preprint arXiv:1412.0065*, 2014. 2
- [44] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, pages 3889–3897, 2015. 2
- [45] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2, 4, 5
- [46] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 2, 3, 4
- [47] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, pages 9869–9878, 2020. 2
- [48] Rajat Sharma, Tobias Schwanndt, Christian Kunert, Steffen Urban, and Wolfgang Broll. Point cloud upsampling and normal estimation using deep learning for robust surface reconstruction. *arXiv preprint arXiv:2102.13391*, 2021. 6
- [49] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, pages 3213–3221, 2015. 2
- [50] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, pages 294–310. Springer, 2016. 2
- [51] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019. 2
- [52] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *ECCV*, pages 530–546. Springer, 2020. 2
- [53] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118:172–193, 2016. 2
- [54] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, pages 5481–5490. IEEE, 2022. 6
- [55] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *CVPR*, pages 16611–16621, 2021. 2
- [56] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, pages 2642–2651, 2019. 2
- [57] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 34:27171–27183, 2021. 2, 3, 4, 6
- [58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6
- [59] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *ECCV*, pages 736–753. Springer, 2022. 2, 3

- [60] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *CVPR*, pages 4737–4746, 2023. [1](#)
- [61] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, pages 2750–2760, 2022. [6](#)
- [62] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, pages 20953–20962, 2022. [1](#)
- [63] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What’s in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, pages 3895–3905, 2022. [1](#), [2](#), [4](#), [6](#), [7](#), [8](#)
- [64] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19728, 2023. [1](#)
- [65] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. [2](#), [4](#), [6](#), [7](#), [8](#)
- [66] Chenyangguang Zhang, Yan Di, Ruida Zhang, Guangyao Zhai, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ddf-ho: Hand-held object reconstruction via conditional directed distance field. *arXiv preprint arXiv:2308.08231*, 2023. [1](#), [2](#)
- [67] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, pages 34–51. Springer, 2020. [2](#)
- [68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. [6](#)
- [69] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. [2](#)
- [70] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *CVPR*, pages 585–594, 2023. [1](#)
- [71] 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 *CVPR*, pages 5346–5355, 2020. [2](#)
- [72] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. [2](#)