Reconstructing Hand-Held Objects in 3D from Images and Videos

Jane Wu[1] Georgios Pavlakos [2] Georgia Gkioxari [3] Jitendra Malik [1]
[1] UC Berkeley [2] UT Austin [3] Caltech

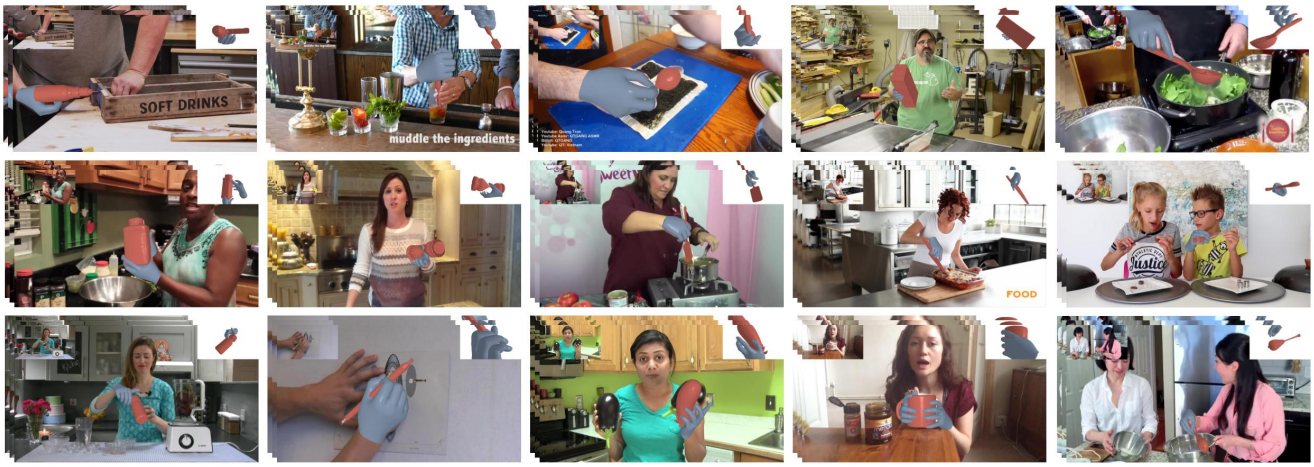

Figure 1. We present a scalable approach to hand-held object reconstruction from monocular RGB images or videos that is guided by object recognition and retrieval. Results demonstrate that our method is able to generate realistic object geometry that is also faithful to visual observations and consistent across frames. Please see the supplementary materials for evaluation of temporal consistency.

Abstract

Objects manipulated by the hand (i.e., manipulanda) are particularly challenging to reconstruct from Internet videos. Not only does the hand occlude much of the object, but also the object is often only visible in a small number of image pixels. At the same time, two strong anchors emerge in this setting: (1) estimated 3D hands help disambiguate the location and scale of the object, and (2) the set of manipulanda is small relative to all possible objects. With these insights in mind, we present a scalable paradigm for hand-held object reconstruction that builds on recent breakthroughs in large language/vision models and 3D object datasets. Given a monocular RGB video, we aim to reconstruct hand-held object geometry in 3D, over time. In order to obtain the best performing single frame model, we first present MCC-Hand-Object (MCC-HO), which jointly reconstructs hand and object geometry given a single RGB image and inferred 3D hand as inputs. Subsequently, we prompt a text-to-3D generative model using a VLM to retrieve a 3D object model that matches the object in the image(s); we call this alignment Retrieval-Augmented Reconstruction (RAR). RAR provides unified object geometry across all frames, and the result is rigidly aligned with both the input images and 3D MCC-HO observations in a temporally consistent manner. Experiments demonstrate that our approach achieves state-of-the-art performance on lab and Internet image/video datasets. We make our code and models available on the project website: https://janehwu.github.io/mcc-ho

1. Introduction

Recovering 3D hand-object interactions from visual data is an important problem in both computer vision and robotics. While there has been significant progress in reconstructing hands [55, 57, 93] and objects [41, 42, 45, 87] separately, joint reconstruction of hand-object interactions presents unique challenges that cannot currently be solved by general reconstruction approaches. Existing image-to-3D models [41, 42, 45, 87] are trained on data where the object is usually fully visible in the image; thus, these models are not trained to infer geometry primarily from hand/scene context. Procedural methods such as Structure from Motion (SfM) [76] will also struggle because the object is mostly occluded, and thus it is difficult to obtain enough feature points to achieve accurate reconstruction. Instead, state-

of-the-art approaches in hand-object reconstruction have relied on human assistance for data collection, whether it be lab datasets [6, 17, 43] or Internet videos [4, 54]. Manually scanned or retrieved object models naturally improve reconstruction quality and can provide better initialization, but this paradigm is not scalable.

In this paper, we present a novel approach to recovering hand-object interactions in 3D from monocular RGB images and videos, focusing in particular on reconstructing hand-held object geometry. See Figure 1 and the supplementary videos. Given an RGB image and estimated 3D hand, we train a transformer-based model, MCC-Hand-Object (MCC-HO), that jointly infers 3D hand and object geometry. This geometry is represented as a neural implicit surface composed of occupancy, color, and hand-object segmentation. By adapting our model architecture from an existing object reconstruction model, MCC [84], we make use of a learned object prior by fine-tuning an MCC model that is pretrained on the CO3Dv2 dataset [66]. Experiments show that our model outperforms existing methods for hand-held object reconstruction on available datasets containing 3D labels, including DexYCB, MOW, and HOI4D [4, 6, 43] (Table 1, Figures 4 and 5).

Given the limited size of existing 3D hand-object datasets, we additionally leverage vision-language models [11, 50] and 3D generative models [9, 45, 47, 56] in order to improve upon a domain-specific approach. In conjunction with MCC-HO, we propose Retrieval-Augmented Reconstruction (RAR), a method for automatically retrieving object models using large language/vision models to improve network-inferred geometry. Specifically, we prompt a VLM [50] to recognize the hand-held object in the image and provide a detailed text description. This description is passed to a text-to-3D generative model [47] to obtain a 3D object, which is then rigidly aligned with the input image or video using a combination of DINOv2 [51] and inferred 3D cues from MCC-HO. We demonstrate that RAR leads to quantitative improvements in 3D reconstruction (Tables 2 and 5), as well as more realistic object geometries compared to existing image-to-3D methods (Figures 5 and 6).

Note that while RAR works well on average, it will fail sometimes if no close enough 3D match is found by the generative model. But the good news is that it does not need to work well on every video for the results of this framework to be useful as a "data engine" for robotics. There is a great deal of interest in getting hand-object interaction trajectories for imitation learning by robots [72] and typically for this data is collected in lab settings [1, 20] which does not result in sufficient diversity. With our paradigm, we can use a quick "thumbs-up, thumbs-down" check from a human observer to screen for all videos where the human-object interactions are reconstructed good enough. We have applied this approach with great success to Internet videos from the 100DOH dataset [71]; see the supplementary materials. While these videos will be of slightly lower quality than those in laboratory collected datasets, they have the advantage of scale and diversity and thus can constitute the pre-training phase of a robot data pipeline with fine-tuning provided by the laboratory datasets.

2. Related Work

Object reconstruction. Reconstructing objects from images is a longstanding problem in computer vision, and the bulk of research in this area has focused on scenarios where the object is fully or almost fully unoccluded in the input view(s). Structure from motion (SfM) [26, 70, 73, 76] and SLAM [5, 49, 75] remain robust solutions particularly for multiview reconstruction and camera parameter estimation. Learning-based object reconstruction methods typically represent 3D geometry as point clouds [60, 61], meshes [21, 83], voxel grids [80, 85], or implicit surfaces [48, 53, 84]. Following the release of Objaverse [14] and Objaverse-XL [15], a number of works have used this large-scale 3D object dataset to train monocular object reconstruction models [31, 41, 42, 44]. The CO3Dv2 dataset [67] contains objects from 50 MS-COCO categories [40] and has also been used to train monocular object reconstruction models [84]. In practice, we find that off-the-shelf object reconstruction methods do not perform well on images of hand-object interactions due to occlusion from the hand, larger object pose diversity, etc.

Hand reconstruction. Recently, there has been a growing body of literature on estimating hand pose from a monocular RGB image [52, 55, 59, 69] or monocular video [34, 82, 86]. These approaches either estimate skeletal joint locations of the hands or use parametric models such as MANO [68]. Since humans have two hands, other methods aim to reconstruct the left and right hands jointly [79, 94]. We use HaMeR [55] in this work as it has achieved state-of-the-art performance on in-the-wild images.

Model-based hand-object pose estimation. When 3D object models are available, the problem of recovering hand-object interactions can be reduced to estimating the pose and shape of the hand and the rigid pose of the object [2, 4, 17, 28, 29, 54, 62, 88]. The estimated hands can then be used to rigidly align the object with images or videos [4, 54, 96] via inverse rendering [46, 64]. Due to the specific challenges arising from hand-object interactions, e.g. highly occluded geometry, optimization-based approaches are considered the current state-of-the-art in this area. RHO [4] studied the problem of recovering hand-object pose from in-the-wild images, whereas prior work was centered around datasets captured in lab settings. RHOV [54] extended RHO to video data by adding, among other terms, temporal and optical flow losses. Aiming to add more physics-inspired constraints, HandyPriors [96] leverages a differen-

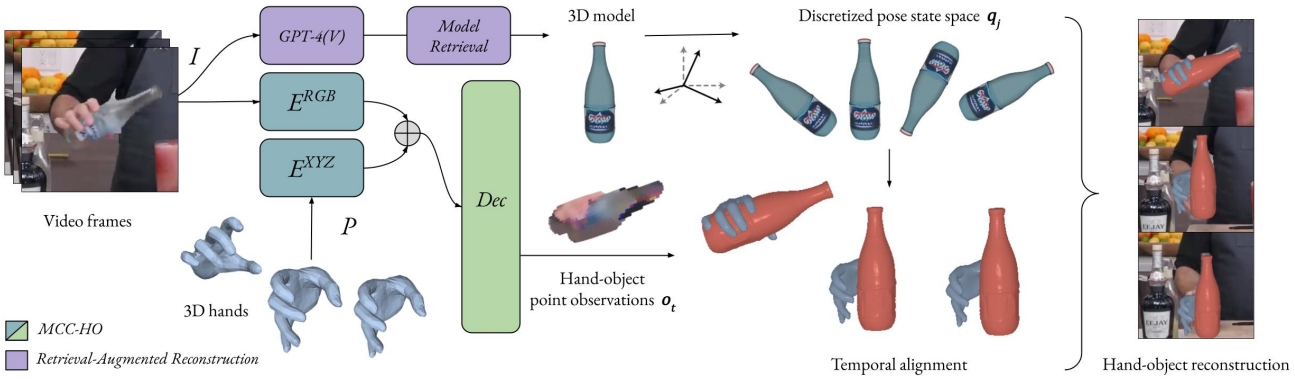

Figure 2. Given an RGB video and estimated 3D hands, our method reconstructs 3D hand-held object trajectories. First, MCC-HO is used to predict hand and object point clouds for each frame (Section 4). Then, a single 3D model for the object is obtained using Retrieval-Augmented Reconstruction (Section 5). The 3D object model is rigidly aligned with DINOv2 [51] image features and network-inferred geometry in a temporally consistent manner via our Viterbi algorithm (Section 6).

tiable contact module to minimize relative sliding between frames. G-HOP [92] guides optimization with a generative hand-object prior.

Model-free hand-object reconstruction. Reconstructing hand-held objects without corresponding 3D models poses additional challenges compared to general object reconstruction due to frequent occlusions by the hand and significantly less publicly available 3D data of hand-object interactions. Existing approaches are either tailored to reconstruction from a single image [7, 8, 10, 19, 27, 58, 90, 95], RGB video [25, 32, 91, 92], or RGB-D video [81]. Our approach combines the respective benefits of model-based and model-free reconstruction; we first reconstruct 3D objects from a single image, and the output is used to guide the registration of a retrieved 3D object model.

Hand-object datasets. Datasets containing hand-object interactions are either collected in lab environments [2, 6, 24, 37, 43], through large-scale capture [22, 23], or from Internet videos [4, 13, 71]. A number of hand-held object reconstruction models [7, 8, 10] were trained on either the HO-3D [24] or DexYCB [6] datasets. Since the combined number of object instances in these two datasets is 31, it is difficult for such approaches to generalize to unseen objects. HOI4D [43] is currently the largest available dataset with 3D hand and object labels (also used in [91]). There is also a comparably larger amount of data containing videos of hand-object interactions without 3D labels. Epic Kitchens [13] contains 55 hours of egocentric video of people in their personal kitchens. Ego4D [22] and Ego-Exo4D [23] are also egocentric datasets, and 100 Days of Hands (100DOH) [71] contains 131 days of footage collected from Internet videos. Following the release of [71], MOW [4] and RHOV [54] labeled a subset of images and video clips from 100DOH, respectively, with 3D hands and objects.

3. Method Overview

Our approach to reconstructing hand-object interactions from monocular RGB video combines feedforward network inference (MCC-HO) with object model retrieval (RAR) and subsequent rigid alignment. See Figure 2. First, our transformer-based model, MCC-HO, estimates 3D hand-object geometry from single images (Section 4). Simultaneously, a 3D object model is automatically synthesized by a text-to-3D generative model [47] through a technique we coin Retrieval-Augmented Reconstruction (Section 5). Finally, the 3D object model is rigidly aligned in a temporally consistent manner (Section 6).

4. Estimating Hand-Object Geometry

In the first stage, hand and object geometry are jointly inferred by training a model adapted from MCC [84] for the specific task of hand-object reconstruction. Our model, MCC-Hand-Object (MCC-HO), has to deal with two major challenges. First, MCC assumes RGB-D images as input, whereas our model needs to work with RGB input, since our ultimate goal is to reconstruct from Internet videos. Second, we have to find a way to cope with the significant domain shift between the training datasets. MCC was trained on the CO3D dataset [67], which consists of clean, 360-degree videos where each object is unoccluded and placed on a solid surface. For MCC-HO, our training data is a relatively small collection of in-the-wild and lab datasets of hand-object interactions, where the hand-held object is at least partially occluded in the vast majority of images.

The inputs to MCC-HO are a single RGB image and 3D hand geometry. Hand and object segmentation masks can be obtained either from ground truth labels or Segment Anything (SAM/SAM2) [35, 65], and the combined hand-object mask determines the bounding box that is used to

crop/resize the input image. Hand geometry is parameterized by the MANO hand model [68], which is composed of pose parameters $\theta \in \mathbb{R}^{48}$ and shape parameters $\beta \in \mathbb{R}^{10}$. The MANO hand skeleton contains 21 joints. At training time, the ground truth 3D hand is used. At test time, HaMeR [55] is used to infer the hand geometry.

The inferred geometry is represented as a neural implicit function $\rho(x)$ modeling the hand and object jointly, such that for any 3D location x, ρ returns the occupancy probability $\sigma(x)$, RGB color $c(x)$, and segmentation label $m(x)$:

$$\rho(x) = (\sigma(x), c(x), m(x)) \tag{1}$$

where $\sigma(x) \in [0, 1]$, $c(x) \in [0, 1]^3$, and $m(x) \in \{0, 1, 2\}$ to indicating a background, hand, or object point. The scale of the object geometry is normalized with respect to the input 3D hand.

4.1. Architecture

MCC-HO has an encoder-decoder architecture that is composed of transformers [77] and outputs an implicit representation of hand-object geometry. The input images and 3D hand are passed to the encoder to compute a feature map that is a function of both image and hand embeddings. These features are then used to condition the decoder, which maps any 3D location to occupancy, color, and hand-object segmentation. During training, a set of query 3D points Q are uniformly sampled within the bounds of a normalized volume. At test time, vertices of a voxel grid are passed as query points.

Encoder. The encoder takes as input a single RGB image $I \in \mathbb{R}^{H \times W \times 3}$ and a posed MANO hand mesh defined in camera world space with vertices $v_h \in \mathbb{R}^{778 \times 3}$ and faces $f_h \in \mathbb{R}^{1538 \times 3}$. The camera intrinsics are either known or consistent with the estimated 3D hand (*e.g.*, the camera parameters used during HaMeR inference [55]). In order to jointly encode the image and hand geometry, the hand mesh is rasterized and per-pixel 3D points for the hand are sampled. For each pixel (i, j) in the image, a ray is cast from the camera aperture to the pixel center to determine the first intersection point with the hand mesh at a point $P_0(i, j)$ in camera world space. If an intersection exists, the 3D location of each point is barycentrically interpolated from the intersecting camera world space triangle vertices, denoted $\{v_{h,1}, v_{h,2}, v_{h,3}\}$, such that:

$$P_0(i, j) = \alpha_1 v_{h,1} + \alpha_2 v_{h,2} + \alpha_3 v_{h,3} \in \mathbb{R}^3, \tag{2}$$

with barycentric weights α. The points P_0 are used to define a normalized camera world space, such that the 3D coordinates passed to the network are $P = s(P_0 - \mu(P_0))/\sigma(P_0)$ and geometry is inferred in this same coordinate system. Importantly, s is a constant scaling factor used to ensure the object geometry can be contained by the bounds of this hand-normalized sampling space.

The image I and hand points P are encoded into a single representation R, where:

$$R := f(E^{RGB}(I), E^{XYZ}(P)) \in \mathbb{R}^{N^{enc} \times C}, \tag{3}$$

and E^{RGB} and E^{XYZ} are identical to the transformers proposed in MCC [84], though we define P differently, *i.e.*, without an input depth image. The image transformer E^{RGB} has a Vision Transformer (ViT) architecture [16] with 16×16 patch embeddings. The point transformer E^{XYZ} uses a self-attention-based patch embedding design that differentiates between seen and unseen pixel points. In particular, for points in P where no ray-mesh intersection was found, a special C-dimensional embedding is learned in lieu of embedding 3D points. The function f concatenates the output of the two transformers, each of which has dimensions $N^{enc} \times C$. N^{enc} is the number of tokens used in the transformers and C is the output channel size.

Decoder. The decoder takes as input the encoder output, R, as well as a set of query points $Q \in \mathbb{R}^{N_q \times 3}$ (we choose $N_q = 1024$). These inputs are passed to a transformer with architecture adopted from MCC and inspired by Masked Autoencoders (MAE) [30]. Each query token is passed to three output heads that infer binary occupancy σ, RGB color c, and hand-object segmentation m, such that:

$$Decoder(R, Q) := (\sigma(Q), c(Q), m(Q)). \tag{4}$$

The decoder layers that output occupancy and color align with MCC, but MCC-HO has an additional linear layer that infers multiclass segmentation of each point into either background, hand, or object labels.

As discussed in further detail in MCC, a point-query-based decoder presents a number of advantages with respect to computational efficiency and scalability. Since the encoder is independent of the point queries, any number of points can be sampled using the same encoder structure and feature map size. The decoder architecture is relatively lightweight and also enables dynamic changes to the point sampling resolution and query size at training and/or test time.

Losses. The training loss is a combination of occupancy, color, and segmentation:

$$\mathcal{L} = \mathcal{L}_\sigma + \mathcal{L}_c(\sigma_{gt}) + \mathcal{L}_m \tag{5}$$

The occupancy loss $\mathcal{L}_\sigma$ is a binary cross-entropy loss comparing the predicted and ground truth occupancies of each query point. The color loss $\mathcal{L}_c$ is a 256-way classification loss computed for query points that have ground truth occupancy of 1. The segmentation loss $\mathcal{L}_m$ is 3-way classification loss for each query point corresponding to background (empty space), hand, and object labels.

5. Retrieval-Augmented Reconstruction

Like any other deep learning model, how well MCC-HO is expected to perform is related to the amount of training data available, *e.g.*, the size of available hand-object datasets. These are currently all quite small, orders of magnitude smaller than the "big data" on which computer vision foundation models have been trained. How might we find a way to exploit these other models to our end? Specifically, recognition of objects is very well developed [50], and large collections of 3D object models [14, 15] exist and have facilitated advancements in 3D generative modeling [9, 45, 47, 56]. Thus, in order to further improve upon network-inferred object geometry, we exploit these "big data" models by using GPT-4(V) [50] and a text-to-3D generative model [47] to automatically detect and "retrieve" a 3D object corresponding to the hand-held object in an image. We call this technique Retrieval-Augmented Reconstruction (RAR), somewhat in analogy to Retrieval-Augmented Generation in natural language processing [38].

Given a representative frame of the video, we prompt GPT-4(V) to provide a detailed description of the object the hand is holding. This text caption is passed to Genie, a text-to-3D model from Luma [47], to retrieve realistic 3D object geometry and appearance (Genie is also used in [39]). Since Genie outputs geometry normalized to fit inside a unit cube, the scale of the 3D model must be adjusted with respect to the input 3D hand.

5.1. Rigid Alignment

After a 3D object model is retrieved, it is rigidly aligned with the input visual data and MCC-HO predictions. We begin by considering the single image case. First, the object scale is estimated using the predicted MCC-HO object point cloud, which is a reasonable approximation because the prediction is grounded by the input 3D hand. If the MCC-HO point cloud is $X \in \mathbb{R}^{N\times 3}$ and a sampled point cloud of the retrieved object model is $Y \in \mathbb{R}^{N\times 3}$, then the scale is defined as $s = \sqrt{\lambda_X^0}/\sqrt{\lambda_Y^0}$ where λ_X^0 is the largest eigenvalue of $X^T X$ and λ_Y^0 is similarly defined for Y.

The remaining parameters to be optimized are object rotation R and translation T. In order to (1) guarantee a globally optimal solution while (2) maintaining fixed compute time, we discretize the space of all possible parameter values to construct finite state spaces $\{\mathcal{R}, \mathcal{T}\}$. The 3D rotation group, SO(3), is discretized using the unit cube subdivision strategy proposed in [36]. Translation is defined as an offset from the mean location of the MCC-HO point cloud $\mu(X)$, and the space is discretized by constructing a small voxel grid centered at $\mu(X)$. In order to avoid intractably large state space sizes (*e.g.*, $\mathbb{R}^3 \times \mathbb{R}^3$), we first optimize for rotation and then translation. For each parameter, we select the state that results in the lowest error using the sum of the following two metrics.

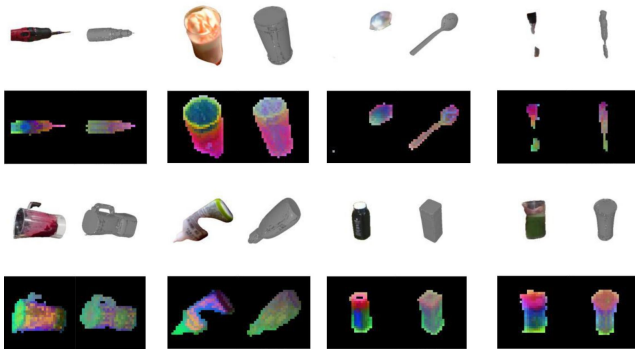

Figure 3. We compute a PCA basis of DINOv2 features using all frames masked by the object silhouettes (one frame and its first three PCA components are shown on the left side of each example). The first three components of this basis are used to determine the maximum likelihood Genie object rotation for each frame (shown on the right side).

MCC-HO Chamfer Distance. The Chamfer Distance (CD) between the MCC-HO point cloud and a transformed object model is the usual 3D evaluation metric. Note that when optimizing for rotation, the object model (centered at the origin) is first rotated and then translated to the mean location of the MCC-HO point cloud.

DINOv2 PCA Similarity. The MCC-HO point cloud gives us a reasonable estimate for object pose, but using this signal alone may not result in poses that align well with the input image (particularly for geometry with symmetries and in-the-wild objects). Thus, we additionally use DINOv2 [51] features to determine visual similarities between a transformed object model and the input image. For each rotation/translation state, the transformed object mesh is rendered using Pytorch3D [64]. A PCA basis is constructed using the DINOv2 features of the input image, and the first three PCA components are computed for both the rendered features $\mathcal{F}_j$ and input image features $\mathcal{F}_0$. We compute the cosine similarity between the two as

$$E_{DINO} = 1 - \frac{1}{2}\left(\frac{\text{PCA}(\mathcal{F}_j)\cdot\text{PCA}(\mathcal{F}_0)}{\max(||\text{PCA}(\mathcal{F}_j)||_2 \cdot ||\text{PCA}(\mathcal{F}_0)||_2), \epsilon)} + 1\right) \quad (6)$$

where $\mathcal{F}_j$ and $\mathcal{F}_0$ are both masked by the input image object silhouette. When ground truth object masks are not available, we use SAM 2 [65]. See Figure 3 for rotation results using example images from 100DOH [71].

6. From Images to Videos

In the previous section, we showed how given a single image, we select the rotation and translation of the 3D object model that minimizes error metrics based on CD and DINOv2 similarity. Now, if we have as input a video stream, we can further exploit the additional cue of temporal smoothness—we expect adjacent frames to have rotation and translation states that are close to each other.

Given a sequence of video frames with timesteps t, we jointly optimize for per-frame rotations $R(t)$ and subsequently per-frame translations $T(t)$. The typical approach for temporal alignment is to use gradient-based optimization [54, 91], which is prone to issues with local minima and has unbounded computation time. Aiming for robustness and efficiency, we propose using the Viterbi algorithm [18, 33, 78] to solve for a global solution to rigid object alignment across all frames of a video sequence. The Viterbi algorithm is a dynamic programming approach to obtaining the maximum *a posteriori* probability of the most likely sequence of hidden states $Q = q_1, q_2, \ldots, q_N$ that results in a sequence of observed events $O = o_1, o_2, \ldots, o_N$. The hidden states q_t can take on values from a finite set of parameters $\mathcal{Q}$ (where in our case, $\mathcal{Q} = \mathcal{R}$ or $\mathcal{Q} = \mathcal{T}$). Following the single image case, the observations o_t are a combination of the MCC-HO inferred object point clouds and the DINOv2 image features.

Each cell of the Viterbi trellis, $v_t(j)$, is the probability of ending up in state j after seeing the first t observations and passing through the most probable state sequence $q_1, \ldots, q_{t-1}$. Given the probability of being in every state at time $t-1$, the Viterbi probabilities are computed via:

$$v_t(j) = \max_{i=1}^{|\mathcal{Q}|} v_{t-1}(i) a_{ij} b_j(o_t), \tag{7}$$

where a_{ij} is the transition probability from the previous state q_i to the current state q_j, and $b_j(o_t)$ is the state observation likelihood given the current state j. In particular, a_{ij} is Rodrigues' rotation error when optimizing for rotations and Euclidean distance when optimizing for translations. The emission cost $b_j(o_t)$ is the sum of the two metrics defined in the single image discussion, *e.g.*, CD with respect to the MCC-HO point clouds and DINOv2 PCA similarity with respect to the input images.

7. Experiments

Section 7.1 details the datasets that were used for training and evaluation. Section 7.2 defines the evaluation metrics used for comparisons. We also quantitatively (Section 7.3) and qualitatively (Section 7.4) evaluate against existing work and find that our approach achieves state-of-the-art performance on all three datasets used [4, 6, 43]. Also in Section 7.4.1, we provide qualitative results for our Retrieval-Augmented Reconstruction (RAR) approach applied to in-the-wild Internet videos from 100DOH [71] to demonstrate the feasibility and scalability of our automated model retrieval and fitting. Section 7.5 includes ablation studies for MCC-HO, improvements attributed to RAR, and the effect of conditioning hand-held object reconstruction on 3D hands.

7.1. Datasets

We experiment on four different datasets for hand-object reconstruction.

HOI4D. The HOI4D dataset [43] consists of 4,000 lab videos of hands interacting with specified categories of common objects. We selected all sequences containing one of 6 hand-held object categories, and divide the sequences into training, validation and test sets. In total, we used 1142 training sequences, 6 validation sequences, and 12 test sequences consistent with DiffHOI [91]. For the validation set, we selected one video per category and favored object instances that appear the least frequently in the dataset.

DexYCB. The DexYCB dataset [6] contains 1,000 lab videos of hands interacting with 21 YCB objects. We used the "s0" dataset split, which contains 800 training, 40 validation, and 160 test sequences.

MOW/RHOV. MOW [4] is a dataset of 512 images from 100 Days of Hands [71] labeled with 3D hands and objects. We use the same splits as [90] with 350 training images and 92 test images. Additionally, we use 110 videos from RHOV [54], each of which is a 2 second clip centered at the corresponding MOW frame. In the RHOV dataset, 19 videos overlap with the MOW test set and thus we do not use them during training.

100DOH. The 100 Days of Hands dataset [71] contains 131 days of footage collected from Internet videos showing hands interacting with objects. The vast majority of videos do not have associated 3D hand or object labels, so we evaluate how well our approach generalizes using examples from this unlabeled subset.

For DexYCB [6], MOW [4], and HOI4D [43], all results generated using our approach use the ground truth 3D hand provided by each dataset. For 100DOH [71] results (*e.g.*, examples not contained in MOW), we use HaMeR [55] to estimate the 3D hand.

7.2. Evaluation Metrics

Since the focus of our approach is on the object manipulated by the hand, we quantitatively evaluate object reconstruction quality using standard 3D metrics: CD (cm^2) and F-score reported at 5mm and 10mm. Both metrics are computed on point sets. The 3D hand provided by each dataset is used as input; note that MOW uses FrankMocap [69] and HOI4D optimizes hand pose using 2D keypoint annotations. For each example, we sample 10k points on the surface of the ground truth mesh consistent with [90, 91] (no transformations are applied to the ground truth geometry). The MCC-HO predicted point clouds are supersampled/subsampled to obtain 10k points as well. In order to account for unknown ground truth camera parameters, ICP with scaling [64] is used to align all predicted geometry with the ground truth mesh. Median metrics are reported unless otherwise noted.

	DexYCB				MOW				HOI4D			
	F-5 (↑)	F-10 (↑)	CD (↓)	Vol (↓)	F-5 (↑)	F-10 (↑)	CD (↓)	Vol (↓)	F-5 (↑)	F-10 (↑)	CD (↓)	Vol (↓)
HO [27]	0.24	0.48	4.76	11.8	0.03	0.06	49.8	78.2	0.28	0.51	3.86	12.7
IHOI [90]	-	-	-	-	0.13	0.24	23.1	20.4	0.42	0.70	2.7	**0.91**
HORSE [58]	0.23	0.42	6.97	3.63	0.11	0.23	24.5	17.0	0.26	0.45	6.69	1.39
MCC-HO (Ours)	**0.36**	**0.60**	**3.74**	**2.42**	**0.15**	**0.31**	**15.2**	**13.8**	**0.52**	**0.78**	**1.36**	1.94

Table 1. We compare our method, MCC-HO, to prior works on held-out test images from DexYCB, MOW, and HOI4D. Chamfer Distance (cm^2), F-score (5mm, 10mm), and Intersection Volume using 0.5cm voxels (as in HO [27], IHOI [90]) are reported.

Method	F-5 (↑)	F-10 (↑)	CD (↓)
G-HOP [92]	0.61	0.89	0.8
MCC-HO + RAR	**0.74**	**0.91**	**0.64**

Table 2. Quantitative comparison to a state-of-the-art video-based hand-held object reconstruction method using the HOI4D test dataset. Mean metrics are reported for consistency with G-HOP.

7.3. Quantitative Evaluation

We quantitatively evaluate MCC-HO on a variety of labeled hand-object datasets and in comparison to existing model-free reconstruction approaches [27, 58, 90]. See Table 1. In order to sample points more densely in hand-object regions, the dimensions and granularity of the voxel grid used to query MCC-HO are determined from an initial coarse prediction using a default voxel grid. For DexYCB and HOI4D, we use every 5th frame of each test video sequence. The results for IHOI [90] and HO [27] evaluated on the MOW dataset are obtained from the IHOI and DiffHOI [91] papers; otherwise, we used available public repositories of prior work. Our method achieves state-of-the-art performance for the Chamfer Distance and F-score metrics on all three datasets, implying that our technique is not overfitted to any particular dataset. We conjecture that increasing the number of unique object instances seen during training is important to model generalization. Thus, the pretrained MCC model provides a strong object prior for initialization.

We also compare to G-HOP [92], a state-of-the-art video-based approach for RGB hand-held object reconstruction in Table 2. Our method (using generated Genie objects for RAR) outperforms the generalist G-HOP model (that was trained on seven hand-object datasets [3, 6, 12, 43, 74, 89]) in all three metrics.

7.4. Qualitative Evaluation

We qualitatively evaluate (1) the reconstruction quality of MCC-HO and (2) how well retrieved object models using RAR can be rigidly aligned to the predicted point clouds in 3D (a necessary and sufficient condition for downstream tasks). Figure 4 shows predicted point clouds for a number of MOW test examples, as well as the result of RAR. Since MCC-HO is trained with monocular images, the predicted RGB colors are only valid when viewed from the given camera. Because the object geometry is conditioned on the 3D hand, our approach tends to predict reasonable hand-object contact. Additionally, the object in contact with the hand is correctly identified by MCC-HO without input segmentation maps. Figure 5 shows qualitative comparisons to existing work, where the combination of MCC-HO + RAR allows us to obtain more realistic object geometry.

Figure 4. MCC-HO results for MOW test examples. The input image (top, left), network-inferred hand-object point cloud (top, right), RAR (bottom, left), and an alternative view of the point cloud (bottom, right) are shown.

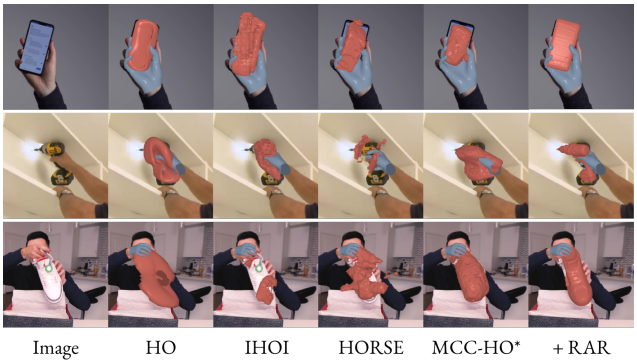

Figure 5. Qualitative comparisons using MOW test examples. *Note that the MCC-HO point clouds are rendered as a mesh via Poisson surface reconstruction, which can introduce artifacts not attributed to our method. The last column is MCC-HO + RAR.

7.4.1 Retrieval-Augmented Reconstruction

The ultimate goal of RAR is to provide a scalable paradigm for creating large hand-object 3D datasets; to this end, we explore the potential to scale up our approach on unlabeled images and videos. Figure 6 shows our results for 9 images from 100DOH [71] that do not have 3D ground truth (*e.g.*, not in MOW). Estimated 3D hands for each image are obtained using HaMeR [55]. At test time, our method does not require an input hand-object segmentation mask [4, 54], depth image [84], or hand joint locations [10, 90]. Please see the supplementary materials for additional results.

Figure 6. Our approach applied to in-the-wild images from the 100 Days of Hands dataset [71]. For each image, we estimate 3D hands using HaMeR [55], inference object geometry using MCC-HO and retrieve a 3D model using RAR (*e.g.*, GPT-4(V) and Genie).

Method	F-5 (↑)	F-10 (↑)	CD (↓)
MCC [84]	0.10	0.21	31.6
MCC-HO	**0.15**	**0.31**	**15.2**

Table 3. Analysis of MCC-HO performance gains over our object reconstruction baseline, MCC, using the MOW test dataset.

Model Parameters	F-5 (↑)	F-10 (↑)	CD (↓)
MCC [84]	0.10	0.21	31.6
MCC-HO (w/o hand)	0.03	0.12	71.6
MCC-HO	**0.15**	**0.31**	**15.2**

Table 4. Analysis of how hand-held object reconstruction benefits from conditioning on the hand using the MOW test dataset.

7.5. Ablation Studies

MCC Baseline. Because our model is built on top of MCC [84], we begin by evaluating whether MCC-HO leads to performance gains for the task of hand-held object reconstruction. For MCC, we use the released model that was trained on the CO3Dv2 dataset [67] and apply an existing monocular depth estimator [63] to obtain depth maps (as MCC takes an RGB-D image as input). Table 3 compares 3D metrics between MCC and MCC-HO computed on the MOW test set. We find that using MCC-HO, which was trained on the specific task of hand-held object reconstruction, leads to performance improvements on test images.

Effectiveness of Input Hand. Table 4 analyzes the effectiveness of hand pose prediction input by evaluating MCC-HO with and without the input hand mesh. As we expect, the model performs significantly worse without the input hand geometry. The experimental results further corroborate the hypothesis that grounding the 3D object prediction with respect to the hand coordinate system significantly constrains the reconstruction problem.

RAR Reconstruction Improvements. RAR leverages text-to-3D generative models to synthesize realistic geometry, and we quantitatively evaluate whether using RAR in conjunction with MCC-HO leads to improved results. Table 5 demonstrates that RAR (with Genie objects) can indeed help improve the object reconstruction quality of monocular reconstruction methods such as MCC-HO.

Method	F-5 (↑)	F-10 (↑)	CD (↓)
MCC-HO	0.15	0.31	15.2
MCC-HO + RAR	**0.16**	**0.44**	**9.41**

Table 5. Analysis of how RAR improves object reconstruction. Median Chamfer Distance (cm^2) and F-score (5mm, 10mm) metrics are computed on the MOW test dataset.

8. Conclusion

In this paper, we present a novel paradigm for reconstructing hand-held objects that combines the respective benefits of model-free and model-based prediction. Our transformer-based model, MCC-HO, is trained to predict hand-object geometry given a single RGB image and estimated 3D hand. Experiments conducted using the DexYCB, MOW, and HOI4D datasets all demonstrate that MCC-HO achieves state-of-the-art performance in hand-held object reconstruction. We additionally present Retrieval-Augmented Reconstruction (RAR), an automatic method for object model retrieval that leverages recent advances in large language/vision models and 3D object datasets. These two approaches can be combined to scale the amount of labeled hand-object interaction data, as suggested by our results using unlabeled 100DOH videos [71].

Limitations. As 3D generative models improve, RAR will naturally be able to retrieve more accurate and realistic geometry. Our approach also relies on the ability to obtain accurate hand and object masks which are tracked across frames. We expect that both MCC-HO and our rigid alignment algorithm will reap the benefits of any future image segmentation/tracking improvements.

Acknowledgements

We would like to thank Ilija Radosavovic for helpful discussions and feedback. This work was supported by ONR MURI N00014-21-1-2801. J. W. was supported by the NSF MSPF and the UC PPFP.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 2

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2, 3

[3] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 7

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 2, 3, 6, 7

[5] Jose A Castellanos, José MM Montiel, José Neira, and Juan D Tardós. The spmap: A probabilistic framework for simultaneous localization and map building. *IEEE Transactions on robotics and Automation*, 15(5):948–952, 1999. 2

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 3, 6, 7

[7] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision*, pages 231–248. Springer, 2022. 3

[8] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12890–12900, 2023. 3

[9] Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21401–21412, 2024. 2, 5

[10] Hongsuk Choi, Nikhil Chavan-Dafle, Jiacheng Yuan, Volkan Isler, and Hyunsoo Park. Handnerf: Learning to reconstruct hand-object interaction scene from a single rgb image. *arXiv preprint arXiv:2309.07891*, 2023. 3, 7

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2

[12] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 7

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 3

[14] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 5

[15] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2

[18] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. 6

[19] Bowen Fu, Yan Di, Chenyangguang Zhang, Gu Wang, Ziqin Huang, Zhiying Leng, Fabian Manhardt, Xiangyang Ji, and Federico Tombari. Hacd: Hand-aware conditional diffusion for monocular hand-held object reconstruction. *arXiv preprint arXiv:2311.14189*, 2023. 3

[20] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17461–17474, 2025. 2

[21] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 2

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activ-

ity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 3

[24] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 3

[25] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17079–17088, 2023. 3

[26] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[27] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 3, 7

[28] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 2

[29] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021. 2

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4

[31] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[32] 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. 3

[33] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000. 6

[34] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12802–12813, 2023. 2

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[36] Gerhard Kurz, Florian Pfaff, and Uwe D Hanebeck. Discretization of so (3) using recursive tesseract subdivision. In *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 49–55. IEEE, 2017. 5

[37] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 3

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 5

[39] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 5

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[41] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[42] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2

[43] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 3, 6, 7

[44] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2

[45] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 2, 5

[46] Matthew M Loper and Michael J Black. Opendr: An approximate differentiable renderer. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 154–169. Springer, 2014. 2

[47] Luma AI. Capture 3D. https://lumalabs.ai. Luma ai. capture 3d. https://lumalabs.ai., 2023. 2, 3, 5

[48] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[49] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[50] OpenAI. Gpt-4v(ision) system card. Openai. gpt-4v(ision) system card., 2023. 2, 5

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 5

[52] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 2

[53] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[54] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022. 2, 3, 6, 7

[55] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. *arXiv preprint arXiv:2312.05251*, 2023. 1, 2, 4, 6, 7, 8

[56] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5

[57] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 1

[58] Aditya Prakash, Matthew Chang, Matthew Jin, Ruisen Tu, and Saurabh Gupta. 3d reconstruction of objects in hands without real world 3d supervision. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 7

[59] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[60] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[61] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[62] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15100–15111, 2023. 2

[63] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 8

[64] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 2, 5, 6

[65] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 5

[66] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2

[67] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 2, 3, 8

[68] 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

[69] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv:2008.08324*, 2020. 2, 6

[70] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 2

[71] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2, 3, 5, 6, 7, 8

[72] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. *arXiv preprint arXiv:2409.08273*, 2024. 2

[73] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 2

[74] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 7

[75] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9(1):1–11, 2017. 2

[76] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9:137–154, 1992. 1, 2

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[78] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967. 6

[79] Congyi Wang, Feida Zhu, and Shilei Wen. Memahand: Exploiting mesh-mano interaction for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 564–573, 2023. 2

[80] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5731, 2021. 2

[81] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023. 3

[82] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21243–21253, 2023. 2

[83] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6187–6197, 2022. 2

[84] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023. 2, 3, 4, 7, 8

[85] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 2

[86] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17048–17058, 2023. 2

[87] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1

[88] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2750–2760, 2022. 2

[89] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 7

[90] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. 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*, pages 3895–3905, 2022. 3, 6, 7

[91] 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. 3, 6, 7

[92] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. *arXiv preprint arXiv:2404.12383*, 2024. 3, 7

[93] Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J Black. Predicting 4d hand trajectory from monocular videos. *arXiv preprint arXiv:2501.08329*, 2025. 1

[94] Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P. Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12955–12964, 2023. 2

[95] Chenyangguang Zhang, Guanlong Jiao, Yan Di, Ziqin Huang, Gu Wang, Ruida Zhang, Bowen Fu, Federico Tombari, and Xiangyang Ji. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. *arXiv preprint arXiv:2310.11696*, 2023. 3

[96] Shutong Zhang, Yi-Ling Qiao, Guanglei Zhu, Eric Heiden, Dylan Turpin, Jingzhou Liu, Ming Lin, Miles Macklin, and Animesh Garg. Handypriors: Physically consistent perception of hand-object interactions with differentiable priors. *arXiv preprint arXiv:2311.16552*, 2023. 2