Articulated Objects in Free-form Hand Interaction

Zicong Fan¹², Omid Taheri², Dimitrios Tzionas², Muhammed Kocabas¹², Manuel Kaufmann¹, Michael J. Black², and Otmar Hilliges¹

¹ ETH Zürich, Switzerland
² Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract. We use our hands to interact with and to manipulate objects. Articulated objects are especially interesting since they often require the full dexterity of human hands to manipulate them. To understand, model, and synthesize such interactions, automatic and robust methods that reconstruct hands and articulated objects in 3D from a color image are needed. Existing methods for estimating 3D hand and object pose from images focus on rigid objects. In part, because such methods rely on training data and no dataset of articulated object manipulation exists. Consequently, we introduce ARCTIC – the first dataset of free-form interactions of hands and articulated objects. ARCTIC has 1.2M images paired with accurate 3D meshes for both hands and for objects that move and deform over time. The dataset also provides hand-object contact information. To show the value of our dataset, we perform two novel tasks on ARCTIC: (1) 3D reconstruction of two hands and an articulated object in interaction; (2) an estimation of dense hand-object relative distances, which we call interaction field estimation. For the first task, we present ArcticNet, a baseline method for the task of jointly reconstructing two hands and an articulated object from an RGB image. For interaction field estimation, we predict the relative distances from each hand vertex to the object surface, and vice versa. We introduce InterField, the first method that estimates such distances from a single RGB image. We provide qualitative and quantitative experiments for both tasks, and provide detailed analysis on the data. Code and data will be available at https://arctic.is.tue.mpg.de.

1 Introduction

Articulated objects are ubiquitous in our daily lives: We open and close our laptops’ lid many times a day, we open and close doors as we move through buildings, we cut paper with scissors – countless other mechanisms exist and are essential to our daily life because of the dexterity of our hands. To understand the rich interactions that humans carry out with such objects, computers need an automatic method to reconstruct hands and articulated objects in 3D from images. To develop such a method, we need training data consisting of images paired with 3D annotations of hands and articulated objects. To reason about fine-grained interactions additional annotations such as contact and 3D surface information are needed. Such a dataset must have accurate ground-truth meshes together with realistic images and sufficient variation in object shapes and hand poses, making the acquisition non-trivial. Existing datasets, focus on grasping of rigid objects [8, 21, 23, 33] or on articulated objects but without human actors [40, 79].
Fig. 1. We present ARCTIC, the first dataset with images and 3D meshes of hands interacting with articulated objects. ARCTIC contains sequences of 9 people that dynamically and freely interact with 10 articulated objects. We train ArcticNet on this data, a model that infers 3D meshes for hands (MANO) and articulated objects from an image. We also provide contact annotations and show that the distance between hands and objects can be estimated from images alone.

Bridging these two strands of research, we collect a novel dataset, called ARCTIC (ARticulated objeCTs in InteracTion), featuring two hands that interact with articulated objects in a fluid and unconstrained manner. ARCTIC consists of sequences of multi-view RGB frames, and each frame is paired with accurate 3D meshes for hands and articulated objects. Additionally, it contains detailed hand-object contact labels. ARCTIC contains data stemming from 9 people interacting with 10 articulated objects, resulting in a total of 1.2M RGB images. Images are captured from multiple synchronized and calibrated views, including 8 static allocentric views and 1 moving egocentric view. To reconstruct accurate meshes for each frame, we draw inspiration from GRAB [67] and sync our color cameras with a Vicon MoCap setup. We use 54 high-resolution Vantage 16 cameras [76], ensuring that small MoCap markers on the subjects and objects are accurately reconstructed despite strong occlusions and the distance to the cameras. We then fit pre-captured human and object meshes to the observed markers [39, 67]. The objects in ARCTIC consist of two rigid parts that rotate about a shared axis. For example, the flip phone in Fig. 1 consists of the main body and a cover that rotates about a hinge. Figure 6 illustrates objects in the ARCTIC dataset. ARCTIC can be used for a variety of hand related tasks. To formalize the study of articulated hand-object interaction, we define two novel tasks: (1) joint 3D reconstruction of two hands and an articulated object from an RGB image; (2) an estimation of dense relative vertex distances of hands and objects from an RGB image — we call this interaction field estimation.

For 3D reconstruction, given an RGB image, the task is to estimate the 3D meshes of two hands and an articulated object in the camera coordinate. This is a non-trivial step beyond estimation of hands only or hands grasping a rigid object. Regressing two hands interacting with articulated objects is inherently a much more challenging problem. We identify three key reasons for this: (1) The former involves significantly more degrees of freedom and images contain much stronger occlusions. (2) We go beyond
static grasps [8, 21] as subjects are instructed to freely and dynamically interact with objects, including in-hand manipulation. This causes a much higher variance in hand poses than existing datasets; see Fig. 2. (3) During interaction with articulated objects, small parts, e.g., lids are often severely occluded and thus hard to observe; see Fig. 1.

As an initial step towards addressing these challenges and to provide a baseline for future work, we introduce ArcticNet to jointly reconstruct two hands with an articulated object in free-form interaction. ArcticNet uses an encoder-decoder architecture to estimate parameters of the MANO hand model for the two hands, and our articulated object model. We experiment with two variations of decoders for ArcticNet: a one-shot decoder and an iterative decoder. We provide different experimental protocols for evaluation, and evaluate ArcticNet both qualitatively and quantitatively.

When interacting with objects, contact is clearly important. Some classical approaches [52, 56] explore the task of binary contact estimation from RGB images with either synthetic data or on limited real data. We go beyond binary contact estimation and propose a new task of interaction field estimation from a single RGB image, in which for each hand vertex, we want to estimate the shortest distance to the object mesh and vice-versa for the object. Note that when the estimated distance is close to zero, the problem formulation is equivalent to binary contact estimation. To provide a baseline, we introduce InterField, the first method that estimates interaction fields from an RGB image. As input InterField takes the RGB image and 3D meshes of hands and an articulated object, both in a canonical pose. For each hand it produces the distances of each hand vertex to the object, and the distances of the object to the hand. See Fig. 10 for examples of prediction and ground-truth. We provide both qualitative and quantitative evaluation of InterField in Sec. 8.

To summarize: (1) We collect ARCTIC, the first large-scale dataset of two hands that freely interact with articulated objects, with multi-view RGB images paired with accurate 3D meshes. (2) We introduce a novel task of monocular RGB reconstruction for two hands and an articulated object in interaction. We establish a baseline ArcticNet for this task for future comparison. (3) We propose the task of interaction field estimation, namely estimating the dense relative distances between the two hands and the object. We also provide a baseline for this task. Our dataset (ARCTIC), models (ArcticNet and InterField) and code will be available at https://arctic.is.tue.mpg.de.

2 Related Work

Estimating 3D hands from an RGB image: Estimating 3D hand pose and shape from images has seen much research attention [66, 81]. Existing methods can be categorized into optimization-based approaches [25, 26, 34, 37, 48, 50], search-based approaches [1, 9, 57], and data-driven approaches to infer single-hand skeletons or meshes from images [4, 23, 27, 42, 45, 61, 63, 64, 84, 85]. More recent methods focus on inferring both hands sometimes with an emphasis on interacting hands. [2, 14, 24, 43, 46, 49, 59, 62, 73, 77, 82]. For example, Moon et al. [43] propose InterHand2.6M, a large-scale multi-view dataset of interacting hands, and use this to train InterNet, a deep net for estimating the 3D pose of two hands. Zhang et al. [82] extract features for each hand separately with a pose-aware attention module and a cascaded refinement stage, while
Rong et al. [59] first predict rough MANO [58] meshes with a CNN, and gradually resolve collisions via a factorized refinement module. The above methods, however, focus on hands alone and not on hand-object interaction. Here we focus on such related works, which introduce additional challenges.

**Estimating 3D hands and objects from an RGB image:** Recently, focus has shifted from optimization [2, 18, 19, 20, 48, 49, 52, 65, 72, 74, 75, 78] and classification [54, 55, 56, 57] based methods towards data-driven ones [10, 11, 22, 23, 36, 69, 80]. Doosti et al. [11] use two graph convolutional neural nets, one for detecting 2D hand joints and object corners, and one for lifting these to 3D. Tekin et al. [68] infer 3D control points for both the hand and the object in videos, using a temporal model to propagate information across time. Several works leverage MANO [58] to go beyond keypoints and model interactions more precisely. Hasson et al. [23] render synthetic images and train a neural network to regress a static grasp of a 3D hand and a rigid object, using full supervision together with contact losses. Later, Hasson et al. [22] use a dataset with sparse ground-truth annotations for hand-object interaction, and “propagate” annotations across frames via photometric consistency. Corona et al. [10] estimate MANO grasps for objects from an image, by first inferring the object shape and a rough hand pose, which is refined via contact constraints and an adversarial prior. Liu et al. [36] use a transformer-based contextual-reasoning module that encodes the synergy between hand and object features, and has higher responses at contact regions. Yang et al. [80] infer rough hand-object poses that are refined with a contact potential field with a “spring-mass” formulation. None of the above methods, however, deal with articulated objects, which result in more complex hand-object interactions.

**Hand-object datasets:** Several datasets [2, 7, 65, 75] contain images of hand-object interaction, but here we focus on the large-scale data [17, 21, 23, 86] that facilitates machine learning. Some datasets contain synthetic images [23] while others capture real images [17, 21, 86]. Garcia-Hernando et al. [17] instrument hands and rigid objects with magnetic sensors to recover pseudo ground-truth pose, and also capture monocular RGB-D images. Zimmermann et al. [86] use a multi-view RGB system and recover ground-truth hand shape and pose by frame-level fitting of MANO to multi-view images. Hampali et al. [21] use a multi-kinect system and fit both MANO and YCB object meshes with sequence-level fitting and contact constraints. Chao et al. [8] use a similar multi-RGB-D system and fit MANO and YCB object meshes using a multi-view setup, human annotations, and frame-level fitting. All of the above datasets capture interactions with one hand. However, recently focus has shifted to two-hands interacting with an object. Kwon et al. [33] target egocentric action recognition, and capture interactions with multiple kinect cameras, one of which is placed on the subject’s head. Brahmbhatt et al. [6] use MoCap and multiple RGB-D cameras to capture hand-object interactions paired with contact heatmaps [5]. However, the grasps in [6] are mostly static and yield relatively little hand pose changes over time. Taheri et al. [67] use MoCap to capture full-body interactions with objects, with dynamic fingers, but capture no images.

**Estimating 3D articulated objects:** Michel et al. [40] capture a dataset of 4 objects with prismatic or revolute joints; object articulation is fixed for each sequence. They then recover the configuration of a known kinematic chain from an RGB-D image through 3D-3D correspondences. Li et al. [35] use a deep net for inferring category-
level pose from a depth image, by predicting an articulation-aware normalized coordinate space hierarchy. Mu et al. [44] regress an articulated signed-distance function from depth images, disentangling shape and articulation encoding. Tseng et al. [71] regress a category-level articulated neural radiance field from a few RGB views of an object.

**Contact detection:** Contact has been shown important for: pose taxonomies [3, 16, 28], pose estimation [18, 21, 23, 65, 72, 75, 80], in-hand scanning [74, 83], grasp synthesis [18, 31, 67, 80]. Many methods [18, 21, 65, 72, 75] use the proximity between the 3D hand and object meshes to estimate contacts and regularize pose estimation based on these. Three main categories for contact estimation exist: 1) directly from meshes; 2) on the image pixel space from RGB images; 3) binary contact in 3D space from RGB images. Grady et al. [18] use off-the-shelf regressors to estimate grasping hand and object meshes. They use these meshes to predict contacts on the objects provided by [6], and use contacts to refine the grasp. Narasimhaswamy et al. [47] infer bounding boxes for all depicted hands in contact on the input RGB image. Fang et al. [15] take an RGB video and infer interaction heatmaps on a separate image of the object. Rogez et al. [56] learn to infer contacts from the image using synthetic data, while Pham et al. [52] use real contact data captured with instrumented objects. Unlike others, [52, 56] estimate 3D binary contact from RGB images but the former suffers from generalizing to real images and the latter uses a classical approach due to the limited amount of data. In contrast, our task in interaction field estimation goes beyond binary contact by using a CNN-based approach and learning from our large dataset. In fact, we consider a richer and harder task: the estimation of dense relative distances between hands and objects.

## 3 ARCTIC Dataset

Since there is no existing dataset featuring two hands and articulated objects with accurate 3D annotations, we construct ARCTIC to enable the task of 3D reconstruction for hands and articulated objects in free-form interaction. ARCTIC contains 242 sequences of free-form manipulation of 10 articulated objects by 9 subjects (4 males and 5 females). It consists of 133k multi-view frames for a total of 1.2M RGB images from 8 static views and 1 egocentric view, paired with accurate 3D MANO and object meshes.

**Comparison to existing datasets:** Table 1 shows a comparison with existing datasets of hand-object interaction and of articulated object pose estimation methods. In particular, ARCTIC is the only dataset containing articulated objects with interacting hands. Further, most datasets [6, 8, 21, 23] study largely static grasps of objects, while subjects in ARCTIC freely interact with objects and perform in-hand manipulation. To minimize ambiguities due to occlusion, we capture our dataset using an accurate Vicon MoCap setup with 54 high-end infrared Vantage 16 cameras [76], whereas HO-3D and DexYCB are captured with 5 – 8 commodity RGB-D cameras. GRAB has a similar MoCap setup as ours, but contains only rigid objects and does not contain images. Michel et al. [40] describe the only articulated object dataset with real RGB images, but it only contains 7K images, too little data to train deep learning models. Another commonly used articulated-object dataset is the PartNet Mobility dataset [41], which consists of 3D models of articulated objects which does not contain images. In com-
### Table 1. Comparison of our ARCTIC dataset with existing datasets.

| Dataset          | Real images | # number of: | Ego-centric image resol. | Articulated objects | Both hands | Freeform interaction | Annot. type |
|------------------|-------------|--------------|--------------------------|---------------------|------------|----------------------|-------------|
| FreiHand [86]    | ✓           | 37k          | 8 35 27                 | ✓                   | ✓          | ✓                   | semi-auto   |
| ObMan [23]       | ✓           | 154k         | 1 20 2.7K               | ✓                   | ✓          | ✓                   | synthetic   |
| FHPA [17]        | ✓           | 105k         | 1 6 4                   | ✓                   | ✓          | ✓                   | magnetic    |
| HO3D [21]        | ✓           | 78k          | 1-5 10 10               | ✓                   | ✓          | ✓                   | multi-kinect|
| ContactPose [6]  | ✓           | 2.9M         | 3 50 25                 | ✓                   | ✓          | ✓                   | multi-kinect|
| GRAB [67]        | -           | -            | 10 51                   | -                   | ✓          | ✓                   | mocap       |
| DexYCB [8]       | ✓           | 582k         | 8 10 20                 | ✓                   | ✓          | ✓                   | manual      |
| H2O [33]         | ✓           | 571k         | 5 4 8                   | ✓                   | ✓          | ✓                   | multi-kinect|
| Michel et al. [40]| ✓           | 7k           | 1 0 4                   | ✓                   | ✓          | ✓                   | -           |
| PartNet-Mobility [79]| -     | -            | -                       | -                   | ✓          | ✓                   | -           |
| ARCTIC (Ours)    | ✓           | 1.2M         | 9 9 10                  | ✓                   | ✓          | ✓                   | mocap       |

### Free-form interaction vs. static grasps:
Existing datasets [8, 21, 23, 67] are mostly limited to stable grasps, which lead to quasi-static hand poses when in contact with the object, whereas our dataset contains free-form interaction in which the subject dexterously and dynamically manipulates the object (see video on project website). Figure 2 compares DexYCB [8], HO-3D [21] and ARCTIC in terms of hand pose variations. Figure 2a quantifies hand pose variation by showing a T-SNE clustering [38] of hand poses in axis-angle representation, revealing that our dataset (shown with red color) has a significantly larger hand pose diversity than DexYCB (shown with green color) and HO-3D (shown with blue color). Figure 2b shows how the object manipulation in ARCTIC produces diverse poses (joint angles are obtained from MANO [58]) throughout interaction sequences. Specifically, we align each sequence of each dataset such that the frames with the first moment of contact (dashed red line) coincide. We use the final frame of each sequence as a reference frame and compute the absolute difference between each frame and the reference in joint angles. We plot the mean of the joint angle distances and their standard deviation. The plot shows that DexYCB (shown with blue color) has very similar hand poses to the final frame hand pose after the first moment of contact. In contrast, our dataset (shown in orange) has much more dynamic hand poses and does not converge as fast to the final pose after the first moment of contact. Our subjects freely interact with objects, including in-hand manipulation. Sequences in DexYCB are a lot shorter than ours.

### Hand-object contact:
Figure 3 shows frequently contacted regions on hands and objects in our ARCTIC dataset. We generate the contact heatmaps following GRAB’s [67] approach, by integrating per-frame binary contact labels for vertices over all sequences. “Hotter” regions denote a higher chance of being in contact while “cooler” regions denote lower chance of contact. Similarly to HO-3D [21] and GRAB [67], finger tips in our dataset are most likely to be in contact with objects. However, thanks to the free-form interactions it contains, ARCTIC has higher contact likelihood in the palm region than the other datasets; note that for ARCTIC the heatmaps appear more “spread out”. For regular-sized everyday objects, such as the ketchup bottle, the contact regions
Fig. 2. **Hand pose and motion variation.** (a): T-SNE clustering of hand poses in different datasets. ARCTIC contains a significantly larger range of poses. (b): Differences in hand pose joint angles in comparison to last frame in time. Compared to other datasets, ARCTIC’s hand poses are more dynamic after the moment contact is established, due to object manipulation.

Fig. 3. **Contact heatmaps for hands and objects.** (Left): Frequently contacted regions for hands in HO-3D [21] and GRAB [67], for the right hand only. (Right): Contacted regions for both hands (right and left) and for objects in our ARCTIC dataset. ARCTIC has higher contact diversity, as seen with the broader heatmap spread on the hands.

“agree” with our usual interaction with them; one hand firmly grasps the bottle’s body by its mid points, while the other one grasps the lid and opens it. For toy objects of (unusually) smaller size, such as the waffle iron, subjects are more likely to pick up the object and support it with one hand, leading to “hot” regions at the bottom of the object.

### 4 Capture Setup

Here we detail our motion capture (MoCap) setup to acquire 3D surfaces of strongly interacting hands and articulated objects, outlined in Fig. 4. We synchronize a MoCap system with a multi-view RGB system, shown in Fig. 5. With the latter we capture synchronized and calibrated RGB images from 8 static allocentric views and 1 moving egocentric view at 30 FPS. The data capture pipeline consists of five steps: (1) obtaining personalized geometry of the subjects and objects in a canonical pose, (2) estimating the rotation axis for each of our single-hinged articulated objects, shown in Appendix, (3) capturing human-object interaction using infrared optical MoCap, (4) solving for the poses of the body, hands, and objects from the observed MoCap markers using MoSh++ [39], following [67], and (5) computing hand-object contact based on geometric proximity, shown in Appendix.
Fig. 4. Markers for motion capture (MoCap). We put 1.5mm radius markers on objects, hands and the egocentric camera. For the body, we use 4.5mm radius markers. The markers are shown here to scale. Best viewed in color and zoomed-in.

Fig. 5. ARCTIC camera views. We use 8 static allocentric views and 1 moving egocentric view. Cameras are synchronized, calibrated, and high resolution. Best viewed in color and zoomed-in.

Obtaining canonical geometry: We obtain the ground-truth hand and body shape of each subject in a canonical T-Pose using 3D scans from a 3dMD [70] scanner. In particular, we register SMPL-X [51] to 3D scans at different time steps in varying poses for the coverage of hands and body shapes and construct a personalized template for each subject. See Appendix for details of the template creation. To obtain object geometries, we scan each object using an Artec 3D hand-held scanner in a pre-defined pose. We cut each scanned object mesh into two parts in Blender and fill mesh holes to ensure each rigid part is watertight. Figure 6 shows all 10 articulated objects in our dataset.

Capturing human-object interaction: Hand-object interaction exhibits severe occlusion, especially when considering dexterous manipulation of articulated objects. To ensure high accuracy, we perform full-body, hand and object tracking using a Vicon MoCap system with 54 infrared Vantage 16 cameras [76] to minimize the issues with occlusion. To capture usable RGB images alongside the MoCap data, we balance the trade-off between accuracy and marker intrusiveness by using minimal-sized hemispherical markers with 1.5mm in radius for hands and objects (see Fig. 4). The markers are placed on the dorsal side of the hand to not encumber participants during natural hand-object interaction, similarly to GRAB [67]. While we focus on hands here, full-body pose estimates provide more reliable global rotations and translations for each
hand. Furthermore, in contrast to MANO which has a rigid wrist, we fit SMPL-X [51] to the observed MoCap markers and thus attain more realistic wrist articulations.

**Obtaining surfaces from MoCap:** To associate MoCap marker positions with their corresponding subject/object vertices in the geometries obtained in canonical spaces (the personalized SMPL-X templates and the scanned object meshes), following [39, 67], we first pick initial guesses of marker-to-vertex correspondence on the subject/object meshes and use MoSh++ [39] to refine the correspondence. To obtain the full body and hand surface that explain the MoCap data, we optimize SMPL-X pose using each subject’s SMPL-X template to minimize the distance between the markers and their correspondences in the SMPL-X mesh. The articulated object surface is parameterized by the 6D pose of each object’s base part and an 1D articulation relative to a canonical pose, for a total of seven degrees of freedom. We obtain the 6D pose of the object base for each MoCap frame by solving the rigid transformation between the MoCap markers for the object base at that frame, and the object vertices corresponding to the markers in the object canonical space. The 1D articulation is computed according to the estimated rotation axis (see Appendix) and a pre-defined rest pose.

5 Method: 3D Hand-Object Reconstruction

The ARCTIC dataset supports the training and evaluation of ArcticNet, the first data-driven method that reconstructs 3D meshes for two hands and the articulated object they interact with from a single RGB image.

**Parametric models:** To simplify our notation, we use $l$, $r$, and $o$ to denote the left hand, the right hand and the object respectively. For hands, we use MANO [58] to represent the hand pose and shape by $\Theta = \{\theta, \beta\}$, which consists of parameters for the pose $\theta \in \mathbb{R}^{48}$ and the shape $\beta \in \mathbb{R}^{10}$. The MANO model maps $\Theta$ to a shaped and posed 3D mesh $H(\theta, \beta) \in \mathbb{R}^{778 \times 3}$ in a differentiable way. The 3D joint locations $J_{3D} = WH \in \mathbb{R}^{7 \times 3}$ are obtained using a pre-trained linear regressor $W$, where $H$ is in canonical pose. For each object, we construct a 3D model $O(\cdot)$ using the scanned object mesh, the estimated rotation axis, and the marker-vertex correspondences estimated in Sec. 4. The function takes as inputs the articulated object pose, $\Omega$, and outputs a posed 3D mesh, $O(\omega, R_o, T_o) \in \mathbb{R}^{V \times 3}$, where $V$ denotes the object’s number of vertices. The object pose, $\Omega \in \mathbb{R}^{7}$, consists of the 1D rotation (radians) for articulation, $\omega \in \mathbb{R}$, and the 6D object rigid pose, i.e., its rotation, $R_o \in \mathbb{R}^{3}$, and translation, $T_o \in \mathbb{R}^{3}$.

**Model architecture:** Figure 7 shows our framework for reconstructing two hands and articulated objects from a single RGB image. Inspired by Hasson et al. [22, 23], we use an encoder-decoder architecture. In particular, the CNN encoder takes in the input image and produces image features $x$. The image features are used by the hand decoders...
The CNN encoder takes a single image as input and produces an image feature vector $x$. The hand decoders predict MANO parameters $\Theta_l, \Theta_r$ and their translation $T_l, T_r$. The object decoder estimates the articulated object pose $\Omega$. Using parametric models for hands $\mathcal{H}(\Theta)$ and for articulated objects $\mathcal{O}(\Omega)$, we obtain posed 3D meshes for the two hands and the articulated object corresponding to the image. For more details see Appendix.

to estimate the parameters for the left and right hands, $\Theta_l$ and $\Theta_r$, as well as the translations for the two hands, $T_l$ and $T_r$. Similarly, the object decoder predicts the articulated object pose, $\Omega$. We use axis-angle as the rotation representation. More details of the network can be found in Appendix. To estimate the translation for hands and objects $(T_l, T_r, T_o)$, we predict weak-perspective camera parameters $(s, t_x, t_y)$ consisting of the scale $s \in \mathbb{R}$ and translation $(t_x, t_y) \in \mathbb{R}^2$ in pixel space. Using a fixed focal length $f$ and image patch width $w$, we then convert the weak-perspective camera to a perspective camera to reconstruct the translation via:

$$T = (t_x, t_y, \frac{2f}{ws}) \in \mathbb{R}^3,$$

We do this for each $(T_l, T_r, T_o)$, following previous work on body and hand surface reconstruction [4, 29, 32, 60, 84].

**Training losses:** Our loss $\mathcal{L}$ is defined as a summation of the left hand, right hand and object losses: $\mathcal{L} = \mathcal{L}_l + \mathcal{L}_r + \mathcal{L}_o$. In particular, the hand losses are defined as

$$\mathcal{L}_h = \lambda^{h}_{3D}\mathcal{L}_3^{h} + \lambda^{h}_{2D}\mathcal{L}_2^{h} + \lambda^{\theta}_{\Theta}\mathcal{L}_\Theta^{h} + \lambda^{\beta}_{\beta}\mathcal{L}_\beta^{h} + \lambda^h_{T}\mathcal{L}_T^{h},$$

where $h = \{l, r\}$ denotes the handedness. We fully supervise the 3D joints (after subtracting the roots), the 2D re-projection of the predicted 3D joints, the MANO pose and shape parameters and the weak-perspective camera parameters. Similarly, we pre-define 3D landmarks for objects using farthest point sampling [12, 13] on the object mesh. Using these pre-defined landmarks, we formulate the object losses as

$$\mathcal{L}_o = \lambda^{o}_{3D}\mathcal{L}_3^{o} + \lambda^{o}_{2D}\mathcal{L}_2^{o} + \lambda_{\omega}\mathcal{L}_\omega^{o} + \lambda_{R}\mathcal{L}_R^{o} + \lambda^o_{T}\mathcal{L}_T^{o},$$

where $\mathcal{L}_\omega, \mathcal{L}_R$ and $\mathcal{L}_T^{o}$ supervise the articulation angle in radians, the global orientation and the weak perspective camera parameters. All $\lambda$ variables are steering weights for each respective loss and are set empirically. All losses above use the MSE criterion.

## 6 Experiment: 3D Hand-Object Reconstruction

**Metrics:** We use four metrics to measure hand-object reconstruction. First, we use the “mean per joint position error” (MPJPE) in millimeter to measure hand pose estimation performance. The error is defined as the L2 distance between the 21 predicted and
ground-truth joints for each hand after subtracting its root. Second, we use the “average articulation error” (AAE) in degrees to measure the performance of articulation estimation for the objects. We compute the average absolute error between the predicted degree of articulation and the ground-truth. Third, to measure the overall object reconstruction, we use the root-relative “vertex-to-vertex” (V2V) distance error in millimeter by computing the average L2 distance between the predicted object vertices and the ground-truth ones. We consider the center of the object bottom part as its root. Lastly, following Moon et al. [43], we use the “mean relative-root position error” (MRRPE) metric to measure the root translation of $a$ relative to $b$, defined as:

$$MRRPE_{a\rightarrow b} = \sqrt{(J_a^0 - J_b^0) - (\hat{J}_a^0 - \hat{J}_b^0)}^2,$$  

where $a \in \{l, r, o\}$ and $b \in \{l, r, o\}$, $J^0 \in \mathbb{R}^3$ is the ground-truth root joint location and $\hat{J}^0$ the predicted one. We report MRRPE$_{l\rightarrow r}$ and MRRPE$_{o\rightarrow r}$ in experiments.

**Evaluation protocols:** Our dataset consists of 8 static allocentric views and 1 egocentric view, and 9 subjects. We recommend the following splits for evaluation:

- **P1 (static):** For reconstruction from allocentric views, all subjects and allocentric views partake in training and evaluation. We withhold separate sequences for each object for the validation and test sets and use the remaining sequences for training.
- **P2 (egocentric):** For reconstruction from the egocentric view, we consider only the egocentric view for training and evaluation for this split, and we withhold separate sequences for each object for the validation and the test sets.
- **P3 (unseen subjects):** We use 6 subjects for training, 1 hold-out subject for validation and 2 hold-out subjects for testing. Only static views are used in this setting. Training, validation and test sets do not share sequences in the splits above.

**Quantitative results:** Table 2 shows evaluation results of ArcticNet on the test set in each split. For validation results, please refer to Appendix. We provide experiments from two variation of ArcticNet: ArcticNet-MLP and ArcticNet-iterative. The former method uses one-shot decoders to estimate hand and object parameters (similar to Has-son et al. [22]). The latter method uses decoders with refinement layers for the parameters (similar to HMR [29]). We observe that the two models have similar hand pose estimation performance for the left hand and the right hand in different splits (shown on the left and the right side of the slash). However, the relative translation prediction between the left and the right hand MRRPE$_{l\rightarrow r}$ has a lower error for ArcticNet-iterative in most protocols. Similar observation for relative translation between the object and the right hand. We hypothesize that ArcticNet-iterative provides better camera parameter estimates due to its refinement procedure. We also notice that MPJPE and MRRPE are lower in the egocentric view setting (P2) than in the allocentric view (P1), as hands are more visible in egocentric views. Object errors are lower in P1 than P2 because P1 has eight times more images. Lastly, as expected, unseen subjects in allocentric views (P3) are harder to estimate in comparison to seen subjects (P1).

**Qualitative results:** Figure 8 shows the prediction of ArcticNet-iterative on our test set. This indicates that ArcticNet performs reasonably well and the reconstruction results resemble the ground-truth even for small objects like the phone. Further, the results show that the our dataset makes it possible to estimate accurate and reasonable articulation of the objects (e.g., the laptop and the microwave) from images.
Fig. 8. Qualitative results of ArcticNet with iterative refinement. Best viewed in color and zoomed in. For qualitative results of ArcticNet with the one-shot decoder, see Appendix.

Table 2. Comparison of methods on the test set. We show MPJPE for left/right hands, MRRPE for between left-right hands and object-right hand (i.e., MRRPE\(l\rightarrow r\)/MRRPE\(o\rightarrow r\)), AAE for the object articulation in degrees, and V2V for top/bottom part of the object. All Euclidean metrics are in millimeter. ArcticNet-MLP uses a one-shot decoder for hands and object parameters where ArcticNet-iterative does an iterative refinement on the parameters. See Appendix for more details.

| Splits | Methods            | MPJPE [mm] | MRRPE [mm] | AAE [°] | V2V [mm] |
|--------|--------------------|------------|------------|---------|----------|
| P1     | ArcticNet-MLP      | 22.41/21.44| 53.85/47.40| 7.25    | 30.08/18.81 |
|        | ArcticNet-iterative| 22.76/21.87| 51.68/43.85| 7.46    | 31.23/19.28 |
| P2     | ArcticNet-MLP      | 19.93/18.62| 32.46/42.10| 8.43    | 40.71/28.31 |
|        | ArcticNet-iterative| 19.81/18.52| 30.79/41.71| 8.48    | 39.70/27.35 |
| P3     | ArcticNet-MLP      | 27.07/25.19| 61.47/49.38| 8.55    | 28.93/17.47 |
|        | ArcticNet-iterative| 28.70/25.90| 61.50/47.23| 8.80    | 29.62/17.51 |

7 Method: Hand-object Interaction Estimation

Classical approaches estimate binary contact from RGB images with synthetic data or limited real data [52, 56]. In two-handed free-form interactions, hands can be near the object surface, but not necessarily in contact. Thus, we need a richer understanding of the hand-object interaction that goes beyond binary contacts. We do so by modeling interaction fields that measure closest distances from the hand to the object and vice-versa. We show how ARCTIC enables estimation of interaction fields from RGB images.

Problem Formulation: We define an interaction field \(F_{a \rightarrow b} \in \mathbb{R}^{V_a}\) as the distance to the closest vertex on the mesh \(M_b\) for all vertices in mesh \(M_a\) where \(V_a\) (or \(V_b\)) is the number of vertices in mesh \(M_a\) (or \(M_b\)). Formally,

\[
F_{i \rightarrow b} = \min_{1 \leq j \leq V_b} ||v_i^a - v_j^b||_2, \quad 1 \leq i \leq V_a
\]

where \(v_k^m \in \mathbb{R}^3\) represents the \(k\)-th vertex of mesh \(M_m\). We define our task to estimate the interaction fields \(F^{l \rightarrow o}, F^{r \rightarrow o}, F^{o \rightarrow l}, \) and \(F^{o \rightarrow r}\) given an RGB image. I.e., for each vertex of each hand we aim to infer the closest distance to the object and vice-versa.

Estimating Interaction Fields: Figure 9 outlines how we estimate interaction fields from an RGB image. Suppose that we estimate the field \(\hat{F}^{l \rightarrow o}\). We first extract image features \(x \in \mathbb{R}^D\) via a CNN backbone. Next, we concatenate \(x\) to each vertex of the
Articulated Objects in Free-form Hand Interaction

**Fig. 9.** Estimating interaction fields. We concatenate image features $x$ to each vertex of the left hand, right hand, and the object in canonical pose. The concatenated vectors are passed through a PointNet and then regressed to distance values. The interaction field is visualized as a heatmap for each entity (dark: closest vertex is far, bright: closest vertex is near).

**Fig. 10.** Qualitative results of estimating interaction fields from a single RGB image. “Left Pred.” means that we show the predicted fields $\hat{F}_{l \rightarrow o}$ (overlaid on the hand) and $\hat{F}_{o \rightarrow l}$ (overlaid on the object). “Left GT” shows the respective ground-truth. Each field is visualized as a heatmap over the meshes (dark: closest vertex is far, bright: closest vertex is near). Ground-truth meshes are only used for visualization.

For tractability, we threshold the interaction field distances at 10cm for training and evaluation in our experiments. For implementation details, please refer to Appendix.

Qualitative Results: Figure 10 shows qualitative samples of the predicted and the ground-truth interaction fields. Their values are visualized as heatmaps over the meshes of the respective hands or objects. A “hotter” region denotes closer distances. Note that the ground-truth meshes are only used for visualization purposes (not network inputs). We see that the predicted fields correlate well with the ground-truth data.

8 Experiment: Hand-object Interaction Estimation

For tractability, we threshold the interaction field distances at 10cm for training and evaluation in our experiments. For implementation details, please refer to Appendix.

Qualitative Results: Figure 10 shows qualitative samples of the predicted and the ground-truth interaction fields. Their values are visualized as heatmaps over the meshes of the respective hands or objects. A “hotter” region denotes closer distances. Note that the ground-truth meshes are only used for visualization purposes (not network inputs). We see that the predicted fields correlate well with the ground-truth data.
Quantitative Results: To evaluate our results quantitatively, we introduce the Percentage of Correct Distances (PCD) metric. For a given threshold level $\alpha$ and interaction field $F_{a\rightarrow b}$ in millimeter, we compute the percentage of absolute errors that are less than $\alpha$, i.e., $|\hat{d}_i - d_i| < \alpha$ for $1 \leq i \leq V_a$. We plot the resulting PCD curves for all four interaction fields with varying $\alpha$ in Fig. 11. Shown are curves for left hand-interactions and right hand-interactions (dashed lines). The plots are computed under the P1 evaluation protocol (cf. Sec. 6).

9 Discussion and Conclusions

We introduce ARCTIC, the first dataset for hands interacting with articulated objects and baseline methods for the task of hand and object reconstruction and for interaction field estimation. Being a first step, our work is not without limitations. One limitation of our method is that it assumes that the object model is known. We view articulated 3D shape estimation of unknown objects as an orthogonal problem on which the field is making progress. Now that we have demonstrated the feasibility of inferring hand-object interaction for such objects, future work should bring together our method with 3D articulated object inference. This is challenging and we believe it is critical to make progress on sub-problems, for which ARCTIC can be leveraged.

Furthermore, our dataset only considers objects with a single revolute joint whereas articulated objects may have multiple joints of different types. Additionally, some of our objects are toys, which are not to scale and lack some of the visual complexity of real objects. Our work is the first step towards hand interaction with articulated objects from RGB images. Future work should expand the number and complexity of objects to further study the problems of depth ambiguity and hand-object occlusion in this setting. Another limitation is that we use optical marker-based capture to provide accurate hand and object poses, thus potentially introducing label noise. However, our hand markers are 1.5mm in radius, which are minimally intrusive and barely visible when the images are resized for network input.

In conclusion, ARCTIC includes high-quality 3D ground-truth for hands, and objects with a total of 1.2M RGB images from 8 static views and 1 egocentric view of 9 subjects interacting with 10 articulated objects. We introduce two novel tasks on ARCTIC: 1) 3D reconstruction of hands and articulated object; 2) interaction field estimation. We introduce ArcticNet, the first method that jointly reconstructs strongly interacting hands and articulated objects from an RGB image, and InterField, the first method that estimates interaction fields from an RGB image. ARCTIC potentially enables more tasks. For example, there is significant work on generating grasps of novel objects [30, 31, 67]. Currently these works focus on static, single-handed grasps and rigid objects. ARCTIC can support extending such work to multiple hands and articulated objects. Our code, data, and models will be available for research purposes.
Disclosure: MJB has received research gift funds from Adobe, Intel, Meta/Facebook, Nvidia, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While MJB was a part-time Amazon employee during this project, his research was performed solely at, and funded solely by, the Max Planck Society.

Acknowledgements: The authors deeply thank: Tsvetelina Alexiadis (TA) for trial coordination; Markus Höschle (MH), Senya Polikovsky, Matvey Safroshkin, Tobias Bauch (TB) for the capture setup; MH, TA and Galina Henz for data capture; Nima Ghorbani for MoSh++; Leyre Sánchez Vinuela, Andres Camilo Mendoza Patino, Mustafa Alperen Ekinci for data cleaning; TB for Vicon support; MH and Jakob Reinhardt for object scanning; Taylor McConnell for Vicon support, and data cleaning coordination; Benjamin Pellkofer for IT/web support. We also thank Adrian Spurr and Xu Chen for insightful discussion. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B

References

1. Athitsos, V., Sclaroff, S.: Estimating 3D hand pose from a cluttered image. In: Computer Vision and Pattern Recognition (CVPR). pp. 432–442 (2003)
2. Ballan, L., Tanuja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: European Conference on Computer Vision (ECCV). pp. 640–653 (2012)
3. Bernardin, K., Ogawara, K., Ikeuchi, K., Dillmann, R.: A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. Transactions on Robotics (T-RO) 21(1), 47–57 (2005)
4. Boukhayma, A., de Bem, R., Torr, P.H.S.: 3D hand shape and pose from images in the wild. In: Computer Vision and Pattern Recognition (CVPR). pp. 10843–10852 (2019)
5. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: Computer Vision and Pattern Recognition (CVPR). pp. 8709–8719 (2019)
6. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: European Conference on Computer Vision (ECCV). vol. 12358, pp. 361–378 (2020)
7. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: International Conference on Computer Vision (ICCV). pp. 12417–12426 (2021)
8. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., Kautz, J., Fox, D.: DexYCB: A benchmark for capturing hand grasping of objects. In: Computer Vision and Pattern Recognition (CVPR). pp. 9044–9053 (2021)
9. Chen, C., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: Computer Vision and Pattern Recognition (CVPR). pp. 5759–5767 (2017)
10. Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F., Rogez, G.: GanHand: Predicting human grasp affordances in multi-object scenes. In: Computer Vision and Pattern Recognition (CVPR). pp. 5030–5040 (2020)
11. Doosti, B., Naha, S., Mirbagheri, M., Crandall, D.J.: HOPE-Net: A graph-based model for hand-object pose estimation. In: Computer Vision and Pattern Recognition (CVPR). pp. 6607–6616 (2020)
12. Eldar, Y., Lindenbaum, M., Porat, M., Zeevi, Y.Y.: The farthest point strategy for progressive image sampling. In: International Conference on Pattern Recognition (ICPR). pp. 93–97 (1994)
13. Eldar, Y., Lindenbaum, M., Porat, M., Zeevi, Y.Y.: The farthest point strategy for progressive image sampling. Transactions on Image Processing (TIP) 6(9), 1305–1315 (1997)
14. Fan, Z., Spurr, A., Kocabas, M., Tang, S., Black, M.J., Hilliges, O.: Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In: International Conference on 3D Vision (3DV). pp. 1–10 (2021)
15. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: Computer Vision and Pattern Recognition (CVPR). pp. 2139–2147 (2018)
16. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The GRASP taxonomy of human grasp types. Transactions on Human-Machine Systems (THMS) 46(1), 66–77 (2016)
17. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In: Computer Vision and Pattern Recognition (CVPR). pp. 409–419 (2018)
18. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: Computer Vision and Pattern Recognition (CVPR). pp. 1471–1481 (2021)
19. Hamer, H., Gall, J., Weise, T., Van Gool, L.: An object-dependent hand pose prior from sparse training data. In: Computer Vision and Pattern Recognition (CVPR). pp. 671–678 (2010)
20. Hamer, H., Schindler, K., Koller-Meier, E., Van Gool, L.: Tracking a hand manipulating an object. In: International Conference on Computer Vision (ICCV). pp. 1475–1482 (2009)
21. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: HOannotate: A method for 3D annotation of hand and object poses. In: Computer Vision and Pattern Recognition (CVPR). pp. 3193–3203 (2020)
22. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Computer Vision and Pattern Recognition (CVPR). pp. 568–577 (2020)
23. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Computer Vision and Pattern Recognition (CVPR). pp. 11807–11816 (2019)
24. He, Y., Yan, R., Fragkiadaki, K., Yu, S.I.: Epipolar transformers. In: Computer Vision and Pattern Recognition (CVPR). pp. 7779–7788 (2020)
25. Heap, T., Hogg, D.C.: Towards 3D hand tracking using a deformable model. In: International Conference on Automatic Face & Gesture Recognition (FG). pp. 140–145 (1996)
26. Iqbal, U., Doering, A., Yasin, H., Krüger, B., Weber, A., Gall, J.: A dual-source approach for 3D human pose estimation from single images. Computer Vision and Image Understanding (CVIU) 172, 37–49 (2018)
27. Iqbal, U., Molchanov, P., Gall, T.B.J., Kautz, J.: Hand pose estimation via latent 2.5D heatmap regression. In: European Conference on Computer Vision (ECCV). pp. 118–134 (2018)
28. Kamakura, N., Matsuo, M., Ishii, H., Mitsuboshi, F., Miura, Y.: Patterns of staticprehension in normal hands. American Journal of Occupational Therapy 34(7), 437–445 (1980)
29. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
30. Karunratanakul, K., Spurr, A., Fan, Z., Hilliges, O., Tang, S.: A skeleton-driven neural occupancy representation for articulated hands. In: International Conference on 3D Vision (3DV). pp. 11–21 (2021)
31. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping Field: Learning implicit representations for human grasps. In: International Conference on 3D Vision (3DV). pp. 333–344 (2020)
32. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: International Conference on Computer Vision (ICCV). pp. 11127–11137 (2021)
33. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2O: Two hands manipulating objects for first person interaction recognition. In: International Conference on Computer Vision (ICCV). pp. 10138–10148 (2021)
34. de La Gorce, M., Fleet, D.J., Paragios, N.: Model-based 3D hand pose estimation from monocular video. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33(9), 1793–1805 (2011)
35. Li, X., Wang, H., Yi, L., Guibas, L.J., Abbott, A.L., Song, S.: Category-level articulated object pose estimation. In: Computer Vision and Pattern Recognition (CVPR). pp. 3703–3712 (2020)
36. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3D hand-object poses estimation with interactions in time. In: Computer Vision and Pattern Recognition (CVPR). pp. 14687–14697 (2021)
37. Lu, S., Metaxas, D.N., Samaras, D., Oliensis, J.: Using multiple cues for hand tracking and model refinement. In: Computer Vision and Pattern Recognition (CVPR). pp. 443–450 (2003)
38. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
39. Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision (ICCV). pp. 5441–5450 (2019)
40. Michel, F., Krull, A., Brachmann, E., Yang, M.Y., Gumhold, S., Rother, C.: Pose estimation of kinematic chain instances via object coordinate regression. In: British Machine Vision Conference (BMVC). pp. 181.1–181.11 (2015)
41. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Computer Vision and Pattern Recognition (CVPR) (June 2019)
42. Moon, G., Lee, K.M.: I2L-MeshNet: Image-to-Lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In: European Conference on Computer Vision (ECCV). vol. 12352, pp. 752–768 (2020)
43. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In: European Conference on Computer Vision (ECCV). vol. 12365, pp. 548–564 (2020)
44. Mu, J., Qiu, W., Kortylewski, A., Yuille, A., Vasconcelos, N., Wang, X.: A-SDF: Learning disentangled signed distance functions for articulated shape representation. In: International Conference on Computer Vision (ICCV). pp. 13001–13011 (2021)
45. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: GANerated hands for real-time 3D hand tracking from monocular RGB. In: Computer Vision and Pattern Recognition (CVPR). pp. 49–59 (2018)
46. Mueller, F., Davis, M., Bernard, F., Sotnychenko, O., Verschoor, M., Otaduy, M.A., Casas, D., Theobalt, C.: Real-time pose and shape reconstruction of two interacting hands with a single depth camera. Transactions on Graphics (TOG) 38(4), 1–13 (2019)
47. Narasimhaswamy, S., Nguyen, T., Hoai, M.: Detecting hands and recognizing physical contact in the wild. In: Conference on Neural Information Processing Systems (NeurIPS). vol. 33, pp. 7841–7851 (2020)
48. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: International Conference on Computer Vision (ICCV). pp. 2088–2095 (2011)
49. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: Computer Vision and Pattern Recognition (CVPR). pp. 1862–1869 (2012)
50. Panteleris, P., Oikonomidis, I., Argyros, A.A.: Using a single RGB frame for real time 3D hand pose estimation in the wild. In: Winter Conference on Applications of Computer Vision (WACV). pp. 436–445 (2018)
51. Pavlakos, G., Choutas, V., Ghorbani, N., Bulkat, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
52. Pham, T., Kyriazis, N., Argyros, A.A., Kheddar, A.: Hand-object contact force estimation from markerless visual tracking. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 40(12), 2883–2896 (2018)
53. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: Computer Vision and Pattern Recognition (CVPR). pp. 77–85 (2017)
54. Rogez, G., Khademi, M., Supančič III, J.S., Montiel, J.M.M., Ramanan, D.: 3D hand pose detection in egocentric RGB-D images. In: European Conference on Computer Vision Workshops (ECCVw). vol. 8925, pp. 356–371 (2014)
55. Rogez, G., Supančič III, J.S., Ramanan, D.: First-person pose recognition using egocentric workspaces. In: Computer Vision and Pattern Recognition (CVPR). pp. 4325–4333 (2015)
56. Rogez, G., Supančič III, J.S., Ramanan, D.: Understanding everyday hands in action from RGB-D images. In: International Conference on Computer Vision (ICCV). pp. 3889–3897 (2015)
57. Romero, J., Kjellström, H., Kragic, D.: Hands in action: Real-time 3D reconstruction of hands in interaction with objects. In: International Conference on Robotics and Automation (ICRA). pp. 458–463 (2010)
58. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. Transactions on Graphics (TOG) 36(6), 245:1–245:17 (2017)
59. Rong, Y., Wang, J., Liu, Z., Loy, C.C.: Monocular 3D reconstruction of interacting hands via collision-aware factorized refinements. In: International Conference on 3D Vision (3DV). pp. 432–441 (2021)
60. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: Metric-scale truncation-robust heatmaps for 3D human pose estimation. In: International Conference on Automatic Face & Gesture Recognition (FG) (2020)
61. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Computer Vision and Pattern Recognition (CVPR). pp. 4645–4653 (2017)
62. Smith, B., Wu, C., Wen, H., Peluse, P., Sheikh, Y., Hodgins, J.K., Shiratori, T.: Constraining dense hand surface tracking with elasticity. Transactions on Graphics (TOG) 39(6), 1–14 (2020)
63. Spurr, A., Dahiya, A., Wang, X., Zhang, X., Hilliges, O.: Self-supervised 3D hand pose estimation from monocular rgb via contrastive learning. In: International Conference on Computer Vision (ICCV) (2021)
64. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: Computer Vision and Pattern Recognition (CVPR). pp. 89–98 (2018)
65. Sridhar, S., Mueller, F., Zollhoefer, M., Casas, D., Oulasvirta, A., Theobalt, C.: Real-time joint tracking of a hand manipulating an object from RGB-D input. In: European Conference on Computer Vision (ECCV). vol. 9906, pp. 294–310 (2016)
66. Supančič III, J.S., Rogez, G., Yang, Y., Shotton, J., Ramanan, D.: Depth-based hand pose estimation: Data, methods, and challenges. In: International Conference on Computer Vision (ICCV). pp. 1868–1876 (2015)
67. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV). vol. 12349, pp. 581–600 (2020)

68. Tekin, B., Bogo, F., Pollefeys, M.: H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 4511–4520 (2019)

69. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6D object pose prediction. In: Computer Vision and Pattern Recognition (CVPR). pp. 292–301 (2018)

70. 3dMDhand system series. https://3dmd.com/products/

71. Tseng, W., Liao, H., Lin, Y., Sun, M.: CLA-NeRF: Category-level articulated neural radiance field. In: International Conference on Robotics and Automation (ICRA) (2022)

72. Tsoli, A., Argyros, A.A.: Joint 3D tracking of a deformable object in interaction with a hand. In: European Conference on Computer Vision (ECCV). vol. 11218, pp. 504–520 (2018)

73. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision (IJCV) 118(2), 172–193 (2016)

74. Tzionas, D., Gall, J.: 3D object reconstruction from hand-object interactions. In: International Conference on Computer Vision (ICCV). pp. 729–737 (2015)

75. Tzionas, D., Gall, J.: Reconstructing articulated rigged models from RGB-D videos. In: European Conference on Computer Vision Workshops (ECCVw). vol. 9915, pp. 620–633 (2016)

76. Vicon Vantage: Cutting edge, flagship camera with intelligent feedback and resolution. https://www.vicon.com/hardware/cameras/vantage

77. Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M.A., Casas, D., Theobalt, C.: RGB2Hands: Real-time tracking of 3D hand interactions from monocular RGB video. Transactions on Graphics (TOG) 39(6), 1–16 (2020)

78. Wang, Y., Min, J., Zhang, J., Liu, Y., Xu, F., Dai, Q., Chai, J.: Video-based hand manipulation capture through composite motion control. Transactions on Graphics (TOG) 32(4), 43:1–43:14 (2013)

79. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., Yi, L., Chang, A.X., Guibas, L.J., Su, H.: SAPIEN: A simulated part-based interactive environment. In: Computer Vision and Pattern Recognition (CVPR). pp. 11094–11104 (2020)

80. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: International Conference on Computer Vision (ICCV). pp. 11097–11106 (2021)

81. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J.Y., Lee, K.M., Molchanov, P., Kautz, J., Honari, S., Ge, L., Yuan, J., Chen, X., Wang, G., Yang, F., Akiyama, K., Wu, Y., Wan, Q., Madadi, M., Escalera, S., Li, S., Lee, D., Oikonomidis, I., Argyros, A.A., Kim, T.: Depth-based 3D hand pose estimation: From current achievements to future goals. In: Computer Vision and Pattern Recognition (CVPR). pp. 2636–2645 (2018)

82. Zhang, B., Wang, Y., Deng, X., Zhang, Y., Tan, P., Ma, C., Wang, H.: Interacting two-hand 3D pose and shape reconstruction from single color image. In: International Conference on Computer Vision (ICCV). pp. 11354–11363 (2021)

83. Zhang, H., Zhou, Y., Tian, Y., Yong, J., Xu, F.: Single depth view based real-time reconstruction of hand-object interactions. Transactions on Graphics (TOG) 40(3), 29:1–29:12 (2021)

84. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular RGB image. In: International Conference on Computer Vision (ICCV). pp. 2354–2364 (2019)

85. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single RGB images. In: International Conference on Computer Vision (ICCV). pp. 4913–4921 (2017)
86. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In: International Conference on Computer Vision (ICCV). pp. 813–822 (2019)