ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation

Zicong Fan[1,3] Omid Taheri[3] Dimitrios Tzionas[2] Muhammed Kocabas[1,3]
Manuel Kaufmann[1] Michael J. Black[3] Otmar Hilliges[1]
[1]ETH Zürich, Switzerland [2]University of Amsterdam [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany

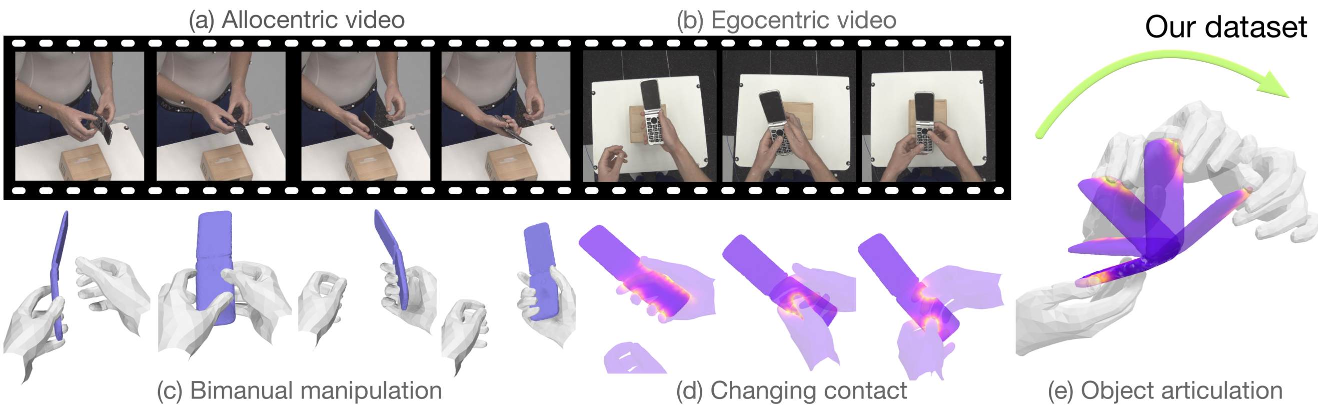


Figure 1. ARCTIC is a dataset of hands dexterously manipulating *articulated* objects. The dataset contains videos from both eight 3^{rd}-person allocentric views (a) and one 1^{st}-person egocentric view (b), together with accurate ground-truth 3D hand and object meshes, captured with a high-quality motion capture system. ARCTIC goes beyond existing datasets to enable the study of dexterous bimanual manipulation of articulated objects (c) and provides detailed contact information between the hands and objects during manipulation (d-e).

Abstract

*Humans intuitively understand that inanimate objects do not move by themselves, but that state changes are typically caused by human manipulation (*e.g.*, the opening of a book). This is not yet the case for machines. In part this is because there exist no datasets with ground-truth 3D annotations for the study of physically consistent and synchronised motion of hands and articulated objects. To this end, we introduce ARCTIC – a dataset of two hands that dexterously manipulate objects, containing* 2.1*M video frames paired with accurate 3D hand and object meshes and detailed, dynamic contact information. It contains bi-manual articulation of objects such as scissors or laptops, where hand poses and object states evolve jointly in time. We propose two novel articulated hand-object interaction tasks: (1) Consistent motion reconstruction: Given a monocular video, the goal is to reconstruct two hands and articulated objects in 3D, so that their motions are spatio-temporally consistent. (2) Interaction field estimation: Dense relative hand-object distances must be estimated from images. We introduce two baselines ArcticNet and InterField, respectively and evaluate them qualitatively and quantitatively on ARCTIC. Our code and data are available at* https://arctic.is.tue.mpg.de.

1. Introduction

Humans constantly manipulate complex objects: we open our laptop's cover to work, we apply spray to clean, we carefully control our fingers to cut with scissors – rigid and articulated parts of objects move *together* with our hands. Inanimate objects only move or deform if external forces are applied to them. The study of the physically consistent dynamics of hands and objects during manipulation has so far been under-researched in the hand pose estimation literature. This is partly because existing hand-object datasets [8, 18, 19, 21, 30, 34] are mostly limited to grasping of rigid objects and contain few if any examples of rich and dexterous manipulation of articulated objects.

To enable the study of dexterous articulated hand-object manipulation, we collect a novel dataset called ARCTIC (**AR**ticulated obje**CT**s in **I**ntera**C**tion). ARCTIC consists of video sequences of multi-view RGB frames, and each frame is paired with accurate 3D hand and object meshes. ARCTIC contains data from 10 subjects interacting with 11 articulated objects, resulting in a total of 2.1M RGB images. Images are captured from multiple synchronized and calibrated views, including 8 static allocentric views and 1 moving egocentric view. To capture accurate 3D meshes during manipulation, we synchronize color cameras with 54 high-resolution Vicon MoCap cameras [66]. These allow the use of small MoCap markers that do not interfere with hand-object interaction and are barely visible in the images. We then fit pre-scanned human and object meshes to the observed markers [35,56]. The objects consist of two rigid parts that rotate about a shared axis such as the flip phone in Fig. 1 (for all objects, see SupMat).

Our dataset enables two novel tasks: (1) consistent motion reconstruction, (2) interaction field estimation. For *consistent motion reconstruction*, given a monocular video, the task is to reconstruct the 3D motion of two hands and an articulated object. In particular, the reconstructed hand-object meshes should have spatio-temporally consistent hand-object contact, object articulation, and smooth motion during interaction. This task has several challenges: (1) Spatio-temporal consistency requires precise hand-object 3D alignment for all frames; (2) This precision is hard to achieve due to depth ambiguity and severe occlusions during dexterous manipulation; (3) The unconstrained interaction causes more variations in hand pose and contact than in existing datasets [8, 18, 19, 34] (see Fig. 2).

As an initial step towards addressing these challenges, and to provide baselines for future work, we introduce ArcticNet to reconstruct the motions of two hands and an articulated object from a video. ArcticNet uses an encoder-decoder architecture to estimate parameters of the MANO hand model [45] for the two hands, and our articulated object model. We experiment with two variations of ArcticNet: a single-frame model and a temporal model with a recurrent architecture inspired by [28]. We provide qualitative and quantitative results for future comparison.

When studying hand-object interaction, contact is important [17, 67]. Some approaches [22, 67] explore the task of binary contact estimation from a single RGB image. In the two-handed manipulation setting, hands can be near the object but not in contact. To understand the dynamic, relative spatial configuration between hands and objects in more detail, even when not in contact, we propose the general task of *interaction field estimation* from RGB images. The goal is to estimate, for each hand vertex, the shortest distance to the object mesh and vice versa (see Fig. 6 for a visualization). We introduce a baseline, InterField, for this task and benchmark both a single-frame and a recurrent version of InterField on ARCTIC for future comparison.

In summary, our contributions are as follows: (1) We present ARCTIC, the first large-scale dataset of two hands that *dexterously* manipulate *articulated* objects, with multi-view RGB images paired with accurate 3D meshes; (2) We introduce two novel tasks of consistent motion reconstruction and interaction field estimation to study the physically consistent motion of hands and articulated objects; (3) We provide baselines for both tasks on ARCTIC.

2. Related Work

Human-object datasets: Several datasets [1, 7, 38, 53, 61, 64] contain images of human-object interaction, but here we focus on large-scale data [3, 15, 18, 21, 23, 47, 78] that facilitates machine learning. There are three categories. *(1) Human body with rigid objects:* Bhatnagar *et al.* [3] and Huang *et al.* [23] introduce image datasets for human body interaction with big objects. Compared to ours, [3] do not capture the hands. Huang *et al.* [23] capture hands and body using a multi-view RGB-D setup while ours is captured using a MoCap setup for more accurate 3D data. Compared to both, we have dexterous bimanual manipulation, dynamic hand-object contact, and articulated objects. GRAB [56] contains detailed human-object interaction but no images, while BEDLAM [4] contains videos with ground-truth humans but no object interaction. *(2) Single hand with rigid objects:* Most hand-object datasets [6, 8, 15, 18, 21, 34] consist of single-hand grasping interaction. However, hand poses in grasping interaction are mostly static, with relatively little pose variation over time. Hampali *et al.* [18] use a multi-RGB-D system and fit both MANO and YCB object meshes with sequence-level fitting and contact constraints. *(3) Two hands with rigid objects:* Kwon *et al.* [30] and Hampali *et al.* [19] present two-hand datasets interacting with rigid objects. Compared to (2) and (3), our dataset has 3D annotations of the full human body, both hands, and articulated objects. We go beyond grasping and focus on less constrained dexterous bimanual manipulation. We discuss the comparison between ours (ARCTIC) and existing hand-object datasets [8, 18, 19, 30, 34] in Sec. 3.1.

Estimating 3D hands and objects from RGB images: Monocular RGB 3D hand reconstruction has a long history since Rehg and Kanade [43]. Most work in the literature focuses on hand-only reconstructions [5, 13, 21, 24, 31, 36, 37, 49–52, 62, 70, 73, 76, 76, 77]. Zimmermann *et al.* [77] use a deep convolutional network for 3D hand pose estimation via a multi-stage approach. Spurr *et al.* [51] introduce biomechanical constraints to regularize hand pose prediction. Ziani *et al.* [76] use a self-supervised time-contrastive formulation to improve smoothness for hand motion reconstruction. Recently, there has been increased interest in hand-object reconstruction from RGB images [12, 17, 20, 21, 33, 57, 67, 75]. Tekin *et al.* [57] infer 3D control points for both the hand and the object in videos, using a temporal model to propagate information across time. Hasson *et al.* [21] 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. Corona *et al.* [12] 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.* [33] use a transformer-based contextual-reasoning module that encodes the synergy between hand

dataset	real images	# number of: img	view	ego-centric	image resol.	articulated objects	both hands	human body	dexterous manipulation	annot. type
FreiHand [78]	✓	37k	8	✗	224×224	✗	✗	✗	✗	semi-auto
ObMan [21]	✗	154k	1	✗	256×256	✗	✗	✗	✗	synthetic
FHPA [15]	✓	105k	1	✓	1920×1080	✗	✗	✗	✗	magnetic
HO3D [18]	✓	78k	1-5	✗	640×480	✗	✗	✗	✗	multi-kinect
ContactPose [6]	✓	2.9M	3	✗	960×540	✗	✗	✗	✗	multi-kinect
GRAB [56]	-	-	-	-	-	✗	✓	✓	✗	mocap
DexYCB [8]	✓	582k	8	✗	640×480	✗	✗	✗	✗	multi-manual
H2O [30]	✓	571k	5	✓	1280×720	✗	✓	✗	✗	multi-kinect
H2O-3D [19]	✓	76k	5	✗	640×480	✗	✓	✗	✗	multi-kinect
HOI4D [34]	✓	2.4M	1	✓	1280×800	✓	✗	✗	✗	single-manual
ARCTIC (Ours)	✓	2.1M	9	✓	2800×2000	✓	✓	✓	✓	mocap

Table 1. **Comparison of our ARCTIC dataset with existing datasets**. The keyword "single/multi-manual" denotes whether single or multiple views being used to annotate manually.

and object features, and has higher responses at contact regions. Zhou *et al*. [74] learn an interaction motion prior to denoise motion predicted from an off-the-shelf single-frame hand-object reconstruction method. None of these methods deal with articulated objects, which result in complex hand-object interactions.

Human-object contact detection: Contact has been shown important for: pose taxonomies [2, 14, 25], pose estimation [17, 18, 21, 53, 60, 64, 67], in-hand scanning [63, 72], and grasp synthesis [17, 27, 56, 67]. Many methods [17, 18, 53, 60, 64] use the proximity between the 3D hand/body 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*. [17] take off-the-shelf regressors to estimate grasping hand and object meshes, use these meshes to predict contacts on the objects provided by [6], and leverage contacts to refine the grasp. Their recent dataset [16] contains both contact and pressure between a hand and a flat sensor surface. Tripathi *et al*. [59] infer pressure from body-scene contact. Narasimhaswamy *et al*. [39] and Shan *et al*. [48] infer bounding boxes for hands in contact on the input RGB image. Chen *et al*. [9] infer human-scene contact on pixels. Rogez *et al*. [44] learn to infer contacts from the image using synthetic data, while Pham *et al*. [41] use real contact data captured with instrumented objects. Unlike others, [44] and [41] estimate 3D binary contact from RGB images but the former does not generalize well to real images and the latter uses a classical approach due to the limited amount of data. BSTRO estimates contact on the 3D body from an image but does not estimate 3D hand or object pose [22]. Hi4D [68] provides ground-truth contact for close human interaction. In contrast, our task of interaction field estimation goes beyond binary contact to model the dense relative distances between hands and objects. Thanks to our dexterous manipulation, ARCTIC contains fast changing hand-object contact.

3. ARCTIC Dataset

Overview: To allow the study of object articulation with hands in motion, we construct ARCTIC, a video dataset with accurate 3D annotation for hands and articulated objects. ARCTIC contains 339 sequences of dexterous manipulation of 11 articulated objects by 10 subjects (5 fe-/males). The dataset consists of 2.1M RGB images from 8 static views and 1 egocentric view, paired with 3D hand and object meshes. To capture different interaction modes, we ask our subjects to either "use" (1.7M images) or "grasp" (457K images) the objects. Depth images of the two hands, the human body, and objects can be rendered from ARCTIC (see SupMat).

3.1. Data Characteristics

Dataset features comparison: Table 1 compares ARCTIC with existing hand-object datasets. ARCTIC is the only dataset that contains both hands, the full human body (in SMPL-X [40]) and articulated objects. ARCTIC provides calibrated cameras (8 allocentric and 1 egocentric) with high-resolution images, enabling the study of monocular, multi-view and egocentric reconstruction settings. Importantly, ARCTIC is a motion dataset that focuses on bimanual dexterous manipulation, meaning that subjects can freely interact with objects using both hands. In contrast, existing hand-object datasets focus single-hand grasping [8, 18, 21] and the movement is often controlled [19, 30]. GRAB [56] has fast motion by using a similar MoCap setup but captures only rigid objects and does not have images. HOI4D [34] is the only hand-object dataset that contains articulated objects, but it contains only a single view, does not capture the full human body, has a single hand, and mainly focuses on grasping. Crucially, their hand data is captured from only a single egocentric view, which introduces ambiguity for the occluded fingers.

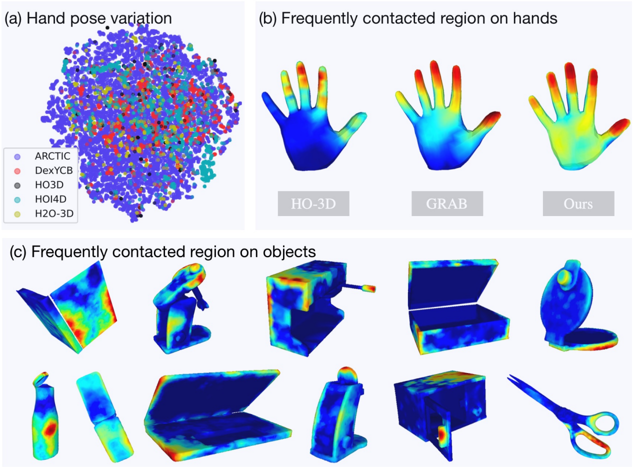


Figure 2. **Hand pose and contact variations in datasets**. (a) T-SNE clustering of hand poses in different datasets. The plot shows that ARCTIC has a significantly larger range of poses than all existing datasets. (b) Frequently contacted regions for hands in HO-3D [18], GRAB [56], and ARCTIC. As seen with the broader heatmap spread on the hands, ARCTIC has higher contact diversity. (c) Frequently contacted areas on our objects.

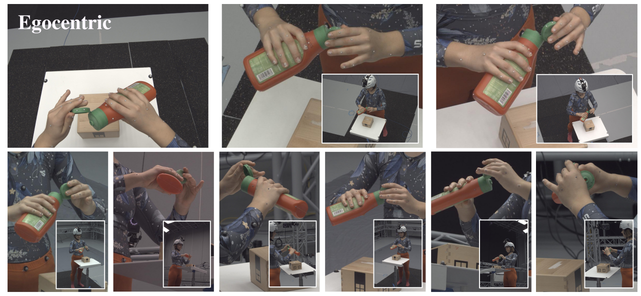


Figure 3. **Our camera views**. We capture high resolution images in 8 static allocentric and 1 moving egocentric views. Here we show zoomed-in crops and the original images.

Capture setup comparison: Capturing dexterous manipulation while maintaining the quality of 3D annotation is extremely challenging due to fast motion and heavy occlusion during the interaction. In particular, the joints of a hand often have significant self-occlusion. The occlusion is even more severe when a hand interacts with objects and when there are multiple hands [36]. Existing hand-object datasets [8,18,19,30,34] are captured with $1-8$ commodity RGB-D cameras, which is insufficient to eliminate occlusion. As a result, their hand-object motion is often slow and they mainly focus on grasping interaction. To reduce occlusion and to enable the capture of dexterous manipulation, we construct our dataset using an accurate Vicon MoCap setup with 54 high-end infrared Vantage-16 cameras [66]. To show our dexterous motion, and to compare 3D annotation quality between datasets, see our project page video.

Hand pose and contact variations: Figure 2a compares different hand-object datasets [8,18,19,34] in terms of hand pose variations by showing a T-SNE clustering [65] of 3D hand joints. The plot reveals that our dataset (shown in blue) has a significantly larger hand pose diversity than others. This is due to the unconstrained nature of ARCTIC in which the subjects dexterously and dynamically *manipulate* the object (see project page video). The figure also shows frequently in-contact regions on hands (b) and objects (c) in the ARCTIC dataset. We generate the contact heatmaps following GRAB's [56] 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. Similar to HO-3D [18] and GRAB [56], finger tips in our dataset are most likely to be in contact with objects. However, thanks to the dexterous manipulation it contains, ARCTIC has higher contact likelihood in the palm region than other datasets, hence the heatmaps appear more "spread out". For regular-sized everyday objects, such as the ketchup bottle, the contact regions "agree" with our usual interaction with them. For smaller toy objects like the waffle iron, subjects are likely to pick up the object and support it with one hand, leading to "hot" regions on the bottom of the object.

3.2. Acquisition Setup

We detail our motion capture (MoCap) setup to acquire 3D surfaces of strongly interacting hands and articulated objects. We synchronize a MoCap system with a multi-view RGB system. See SupMat for the marker sets. With the latter we capture RGB videos from 8 static allocentric views and 1 moving egocentric view at 30 FPS (see Fig. 3). The capture pipeline has five steps: (1) obtaining the 3D template geometry of the subjects and objects, (2) estimating the rotation axis for articulated objects, shown in SupMat, (3) capturing interaction using marker-based MoCap together with calibrated and synchronized video, (4) solving for the poses of the body, hands, and objects from MoCap markers following [35, 56], and (5) computing hand-object contact based on proximity, shown in SupMat.

Obtaining canonical geometry: We obtain the ground-truth (GT) hand and body shape of each subject in a canonical T-Pose using 3D scans from a 3dMD [58] scanner. We register SMPL-X [40] to 3D scans at different time steps in varying poses and construct a personalized 3D template of each subject. See the SupMat 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 separate each scanned object mesh into two articulated parts in Blender. See SupMat for all 11 articulated objects.

Capturing human-object interaction: To ensure accuracy, we perform full-body, hand and object tracking using a Vicon MoCap system with 54 infrared Vantage-16 cameras [66] 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 small hemispherical markers with 1.5mm radius on the hands and objects. The markers are placed on the dorsal side of the hand to not encumber participants during natural hand-object interaction, similar to GRAB [56]. While our focus is on hands, we retrieve full-body pose estimates as they provide more reliable global rotations and translations for each hand. Therefore, we fit SMPL-X [40] to the observed markers to attain realistic wrist articulations, as MANO contains no wrist articulation.

Obtaining surfaces from MoCap: Following [35, 56], we associate MoCap marker positions with their corresponding subject/object vertices in the geometries obtained in canonical spaces. We first pick initial guesses of marker-to-vertex correspondence on the subject/object meshes and use MoSh++ [35] 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 on 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. We obtain the 6D pose of the object base for each MoCap frame by solving for the rigid transformation between the MoCap markers of 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 SupMat) and a pre-defined rest pose.

4. Evaluation Protocol

Data split: We split the data by subjects, 8 subjects for training, 1 for validation (male) and 1 for testing (female). To ensure gender balance in evaluation, we use one male and one female subject. With this same split, we establish two protocols: an allocentric protocol (**allo**) and an egocentric protocol (**ego**). The former protocol lets us study our tasks in the 3rd-person, while the latter is similar to 1st-person views in a mixed-reality setting. In the allocentric protocol, during training and evaluation, the model only has access to images from the allocentric views. In the egocentric protocol, to provide additional training images, we allow models access to images from all views of the training split, but in evaluation, only egocentric images are used. Further information can be found in SupMat.

Metrics for consistent motion reconstruction: Our goal is to reconstruct the 3D motion of the hands and an articulated object during dexterous manipulation from a video. Importantly, our focus extends beyond hand-object poses and we require the reconstructed meshes to have accurate hand-object contact (CDev), and smooth motion (ACC). Further, when a hand moves or articulates an object, vertices of the hand and the object in stable contact should move together (MDev). To this end, we define the following metrics:

- **Contact Deviation (CDev):** For a frame, suppose $\{(\mathbf{h}_i, \mathbf{o}_i)\}_{i=1}^{C}$ are C pairs of in-contact hand-object vertices ($< 3mm$ distance in ground-truth), and $\{(\hat{\mathbf{h}}_i, \hat{\mathbf{o}}_i)\}_{i=1}^{C}$ are the corresponding predictions. CDev is defined as the average distance between $\hat{\mathbf{h}}_i$ and $\hat{\mathbf{o}}_i$ in millimeters:

$$\frac{1}{C}\sum_{i=1}^{C}||\hat{\mathbf{h}}_i - \hat{\mathbf{o}}_i|| \tag{1}$$

  This metric reflects how much the hand vertices deviate from the supposed contact vertices on the object in the prediction.
- **Motion Deviation (MDev):** Given a ground-truth sequence of a hand and an object, we denote vertex i of the hand and vertex j of the object at frame t as $\mathbf{h}_i^t$, $\mathbf{o}_j^t$ respectively. We use (i, j, m, n) to denote $\mathbf{h}_i^t$ has stable contact with $\mathbf{o}_j^t$ during a window from frame m to frame n, and they do not have contact at time $m-1$ and $n+1$ (*i.e.*, longest contact window). Hand-object vertex indices (i, j) have stable contact in a window (m, n) if they are close within a threshold α for every frame in the window:

$$\forall t \in \{m, \cdots, n\}, \left\|\mathbf{h}_i^t - \mathbf{o}_j^t\right\| \le \alpha\,. \tag{2}$$

  Given the above definition, we extract a set of tuples $\{(i, j, m, n)\}$ from each GT sequence. When two hand-object vertices $\mathbf{h}_i^t$, $\mathbf{o}_j^t$ are in stable contact within a window, they should move in the same direction in consecutive frames. To measure this, we define the motion deviation for a tuple (i, j, m, n) of the predicted hand-object sequence $\hat{\mathbf{h}}$ and $\hat{\mathbf{o}}$ as

$$\frac{1}{n-m}\sum_{t=m+1}^{n}||\delta\hat{\mathbf{h}}_i^t - \delta\hat{\mathbf{o}}_j^t|| \tag{3}$$

  where $\delta\hat{\mathbf{h}}_i^t = \hat{\mathbf{h}}_i^t - \hat{\mathbf{h}}_i^{t-1}$ and $\delta\hat{\mathbf{o}}_j^t = \hat{\mathbf{o}}_j^t - \hat{\mathbf{o}}_j^{t-1}$. Intuitively, this measures the disagreement in the moving direction between consecutive frames of in-contact hand-object vertices in the window (m, n). We only consider longer motions by using windows with at least 0.5 second or 15 frames (*i.e.*, $n - m + 1 \ge 15$) and we choose $\alpha = 3mm$ to detect a sufficient number of windows. We compute this metric for all detected windows and average over them.
- **Acceleration Error (ACC):** Following [28], we report acceleration error in m/s^2 to measure the smoothness of the reconstruction, calculated as the difference in acceleration between the ground-truth and predicted vertex sequences for each hand and the object. We subtract the root for each entity before computing the acceleration [28]. The root for the object is defined as the center of an object's base. Note that we report this error in m/s^2, while [28] reports mm/s^2. See SupMat for more details.

Apart from motion and contact, we need metrics to measure hand and object poses, and their relative translations:

- **Mean Per-Joint Position Error (MPJPE):** the L2 distance (mm) between the 21 predicted and ground-truth joints for each hand after subtracting its root.
- **Average Articulation Error (AAE):** the average absolute error between the predicted degree of articulation and the ground-truth.
- **Success Rate:** Following [54, 69], to measure object reconstruction quality, we use a success rate metric that is independent of the object size. It is the percentage of predicted object vertices having L2 error to the ground-truth that is less than 5% of the object diameter:

$$\frac{1}{V_o}\sum_{i=1}^{V_o}\mathbb{1}(\|\mathbf{o}_i-\hat{\mathbf{o}}_i\|<0.05D)\times 100\% \tag{4}$$

  where D, V_o, $\mathbf{o}_i$, $\hat{\mathbf{o}}_i$ are the diameter, the number of object vertices, ground-truth and predicted object vertices, and $\mathbb{1}(\cdot)$ is the indicator function. To decouple the effect of root estimation, we subtract the predicted and the ground-truth vertices by their object roots respectively. The root is the center of each object's base.
- **Mean Relative-Root Position Error (MRRPE):** Following [13, 36], to measure the root translation of between hand-hand and hand-object, we use this metric to measure the relative root translation between two entities a and b in the scene,

$$\mathrm{MRRPE}_{a\to b}=\left\|\left(\mathbf{J}_0^a-\mathbf{J}_0^b\right)-\left(\hat{\mathbf{J}}_0^a-\hat{\mathbf{J}}_0^b\right)\right\|_2, \tag{5}$$

  where $a\in\{l,r,o\}$ and $b\in\{l,r,o\}$ and l,r,o denote the left hand, right hand, and the object, $\mathbf{J}_0\in\mathbb{R}^3$ is the ground-truth root joint location and $\hat{\mathbf{J}}_0$ the predicted one. A graphical illustration of this metric can be found in SupMat.

Metrics for interaction field estimation: In this task, given images from a video, for each hand vertex i, we estimate its shortest distance $\hat{\mathbf{F}}_i^{r\to o}\in\mathbb{R}$ to the object (*i.e.*, the distance field from a hand to the object) and vice versa. Taking the field from the right hand to the object as an example, to quantify, we measure the average error between the predicted distances $\hat{\mathbf{F}}_i^{r\to o}$ and the ground-truth distances $\mathbf{F}_i^{r\to o}$ in millimeters, which we call average distance error. The error is computed as:

$$\frac{1}{V_r}\sum_{i=1}^{V_r}|\mathbf{F}_i^{r\to o}-\hat{\mathbf{F}}_i^{r\to o}| \tag{6}$$

where V_r is the number of right-hand vertices. To measure smoothness, we estimate the distance field for every frame in each sequence. We then compute the acceleration sequence for the predicted field sequence. The acceleration error is computed as the average absolute difference between predicted and ground-truth acceleration sequences. See SupMat for the formula of acceleration error.

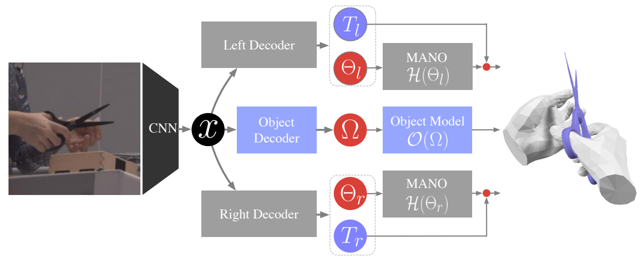


Figure 4. **ArcticNet-SF architecture**. The CNN encoder yields image features x. The hand decoders predict MANO parameters $\mathbf{\Theta}_l, \mathbf{\Theta}_r$ and their translation $\mathbf{T}_l, \mathbf{T}_r$ while the object decoder estimates the articulated object pose $\mathbf{\Omega}$ consisting of the articulation, rotation and translation. With parametric models of hands $\mathcal{H}(\mathbf{\Theta})$ and articulated objects $\mathcal{O}(\mathbf{\Omega})$, we obtain 3D meshes for the two hands and the articulated object .

5. Baselines and Experiments

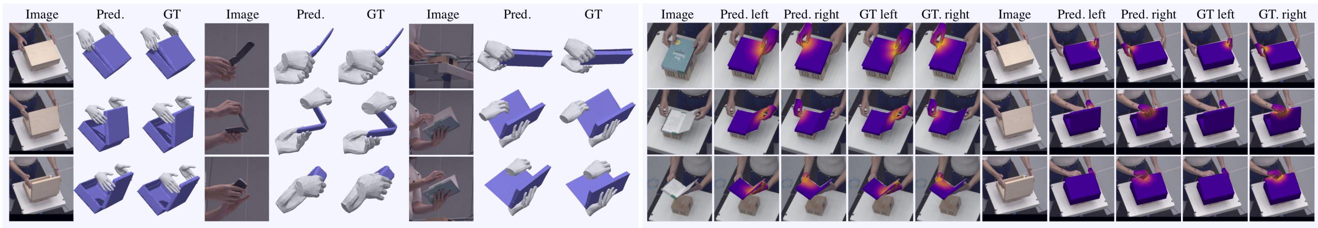
We present two tasks on ARCTIC: consistent motion reconstruction and interaction field estimation. For consistent motion reconstruction, we reconstruct the 3D motion of two hands and an articulated object from a video. For interaction field estimation, given a video, we estimate, for each hand vertex, the closest distance to the object and vice versa. Here we detail and evaluate our baselines in the two tasks to lay the foundation for future comparison.

5.1. Consistent motion reconstruction

Problem formulation: Given a video, our goal is to reconstruct the 3D motion of a subject's two hands and an articulated object in dexterous manipulation for every frame. Our emphasis is to require the reconstructed hand-object meshes to be in temporally-consistent hand-object contact and motion during object articulation and manipulation.

Parametric models: For brevity, we use l, r, and o to denote the left hand, the right hand and the object. For hands, we use MANO [45] to represent the hand pose and shape by $\mathbf{\Theta}=\{\theta,\beta\}$, which consists of parameters for the pose $\theta\in\mathbb{R}^{48}$ (with global orientation) and the shape $\beta\in\mathbb{R}^{10}$. The MANO model maps $\mathbf{\Theta}$ to a shaped and posed 3D mesh $\mathcal{H}(\theta,\beta)\in\mathbb{R}^{778\times 3}$. The 3D joint locations $\mathbf{J}=W\mathcal{H}\in\mathbb{R}^{\mathrm{J}\times 3}$ are obtained using a pre-trained linear regressor W. For each object, we construct a 3D model $\mathcal{O}(\cdot)$ using the scanned object mesh, the estimated rotation axis, and the marker-vertex correspondences estimated in Sec. 3.2. The function takes as inputs the articulated object pose, $\mathbf{\Omega}$, and outputs a posed 3D mesh, $\mathcal{O}(\mathbf{\Omega})\in\mathbb{R}^{\mathrm{V}\times 3}$, where V denotes the object's number of vertices. The object pose, $\mathbf{\Omega}\in\mathbb{R}^7$, consists of the 1D rotation (radians) for articulation, $\omega\in\mathbb{R}$, and the 6D object rigid pose, *i.e.*, its rotation, $\mathbf{R}_o\in\mathbb{R}^3$, and translation, $\mathbf{T}_o\in\mathbb{R}^3$.

Baselines: We introduce ArcticNet to estimate the poses of the two hands and the articulated object from RGB images.



(a) Consistent motion reconstruction

(b) Interaction field estimation

Figure 5. **Qualitative results of ArcticNet-LSTM (a) and InterField-LSTM (b)**. Best viewed in color and zoomed in. See SupMat for results of ArcticNet-SF and InterField-SF.

		Contact and Relative Position		Motion		Hand	Object	
Splits	Method	CDev$_{ho}$ $[mm]\downarrow$	MRRPE$_{rl/ro}$ $[mm]\downarrow$	MDev$_{ho}$ $[mm]\downarrow$	ACC$_{h/o}$ $[m/s^2]\downarrow$	MPJPE$_h$ $[mm]\downarrow$	AAE $[°]\downarrow$	Success Rate [%] $\uparrow$
Allo. Val	ArcticNet-SF	41.4	50.1/37.6	10.4	6.6/8.8	23.0	5.9	71.8
	ArcticNet-LSTM	**38.8**	**47.1/36.8**	**8.9**	**5.6/6.9**	**22.9**	**5.8**	**74.9**
Allo. Test	ArcticNet-SF	41.6	52.4/**37.5**	10.4	5.7/7.6	**21.5**	5.4	71.4
	ArcticNet-LSTM	**38.9**	**49.2**/37.7	**9.3**	**5.0/6.1**	**21.5**	**5.2**	**73.5**
Ego. Val	ArcticNet-SF	**44.1**	**33.9/36.8**	11.8	6.3/11.3	**22.9**	**8.0**	59.0
	ArcticNet-LSTM	44.5	39.3/39.0	**8.1**	**4.3/7.2**	23.8	**8.0**	**59.1**
Ego. Test	ArcticNet-SF	44.7	**28.3**/36.2	11.8	5.0/9.1	**19.2**	**6.4**	**53.9**
	ArcticNet-LSTM	**43.3**	31.8/**35.0**	**8.6**	**3.5/5.7**	20.0	6.6	53.5

Table 2. **Comparison of two reconstruction baselines**. Contact and relative position metrics measure hand-object contact and relative root position prediction. Motion metrics reflect motions with temporally-consistent contact and smoothness. Hand and object metrics show root-relative reconstruction error. See Sec. 4 for metric details. We use l, r, o to denote the left, the right hand, and the object. To simplify the results, we average left and right hand metrics into one hand (denoted by h). For example, CDev$_{ho}$ is the contact deviation between a hand and the object averaged over the two hands; MRRPE$_{rl/ro}$ denotes MRRPE$_{r\to l}$ and MRRPE$_{r\to o}$ between the slash.

We benchmark two versions of ArcticNet: a single-frame model (ArcticNet-SF), and a model with a recurrent architecture (ArcticNet-LSTM). The LSTM baseline is used to allow a joint reasoning of hand and articulated object motions. Figure 4 summarizes the architecture of ArcticNet-SF. Inspired by Hasson *et al*. [20, 21], we use an encoder-decoder architecture. In particular, the CNN encoder takes in the input image and produces image features $\mathbf{x}$. The features are used by the hand decoders to estimate the parameters for the left and right hands, $\mathbf{\Theta}_l$ and $\mathbf{\Theta}_r$, as well as the translations for the two hands, $\mathbf{T}_l$ and $\mathbf{T}_r$. Similarly, the object decoder predicts the articulated object pose, $\mathbf{\Omega}$. We use axis-angle for rotation and use the weak perspective camera model to estimate the translations [5, 26, 29, 46, 73]. The ArcticNet-LSTM model has the same architecture as ArcticNet-SF, except that it has an LSTM network to aggregate image features from multiple frames before passing them to the regression heads. We train the models with ground-truth 3D keypoints, 2D projected keypoints, and the parameters of the hand and the object models. We show details of the model and the training procedure in SupMat.

Results: Figure 5a shows the predictions of one of our baselines, ArcticNet-LSTM. To see qualitative results of ArcticNet-SF, we refer to the SupMat. Table 2 shows the quantitative evaluation of the two baseline models on ARCTIC. The results show that, overall, the ArcticNet-LSTM model has temporally more consistent contact (CDev), and motion (MDev) between the hands and objects. Further, it has smoother motion (ACC). This demonstrates that temporal modelling is important for spatio-temporally consistent hand-object motion and contact. See Sec. 4 for metric details.

5.2. Interaction field estimation

Existing contact detection methods mainly focus on binary contact estimation [17, 67]. In two-handed dexterous interactions, hands can be near the object, but not always in contact. We define a general task of interaction field estimation to capture the relative spatial relations between hands and the object even when not in contact.

Problem formulation: We define an interaction field $\boldsymbol{F}^{a\to b} \in \mathbb{R}^{V_a}$ as the distance to the closest vertex on the mesh $\mathbf{M}_b$ for all vertices in mesh $\mathbf{M}_a$ where V_a (or V_b) is the number of vertices in mesh $\mathbf{M}_a$ (or $\mathbf{M}_b$). Formally,

$$\boldsymbol{F}_i^{a\to b} = \min{}_{1\leqslant j\leqslant V_b} ||\mathbf{v}_i^a - \mathbf{v}_j^b||_2, \quad 1 \leqslant i \leqslant V_a \qquad (7)$$

where $\mathbf{v}_k^m \in \mathbb{R}^3$ represents the k-th vertex of mesh $\mathbf{M}_m$. We define our task to estimate the interaction fields $\boldsymbol{F}^{l\to o}$, $\boldsymbol{F}^{r\to o}$, $\boldsymbol{F}^{o\to l}$, and $\boldsymbol{F}^{o\to r}$ for each image. In other words, for each vertex of each hand we aim to infer the closest distance to the object and vice-versa.

Splits	Method	Average Distance Error $[mm]$↓	ACC $[m/s^2]$↓
Allo. Val	InterField-SF	9.6/9.9	3.0/2.9
	InterField-LSTM	**9.0/8.9**	**2.1/2.0**
Allo. Test	InterField-SF	9.0/10.0	2.7/2.7
	InterField-LSTM	**8.7/9.1**	**1.9/1.9**
Ego. Val	InterField-SF	8.8/9.2	2.4/2.3
	InterField-LSTM	**8.4/8.9**	**2.1/2.0**
Ego. Test	InterField-SF	8.2/9.2	2.1/2.0
	InterField-LSTM	**8.0/9.1**	**1.8/1.8**

Table 3. **Comparison of two field estimation baselines**. To simplify the evaluation, we average metrics for the two hands into one. The slashes denote the average distance error and the acceleration error for hand-to-object/object-to-hand.

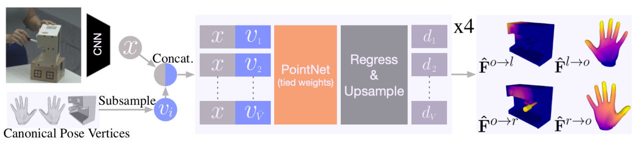


Figure 6. **InterField-SF architecture**. We concatenate image features $\mathbf{x}$ to each subsampled hand-object vertex in canonical pose. The concatenated vectors are passed through a PointNet and then regressed to distance values. The interaction field is visualized as a heatmap for each entity (bright: closest vertex is near).

Baselines: We present InterField to estimate the interaction field from RGB images. We benchmark two versions of InterField: a single-frame (InterField-SF) and a temporal baseline (InterField-LSTM). The temporal model lets us evaluate the benefits of temporal information. Figure 6 outlines the framework of InterField-SF. Suppose that we estimate the field $\hat{\boldsymbol{F}}^{l\to o}$. We first extract image features $\mathbf{x} \in \mathbb{R}^{\mathrm{d}}$ via a CNN backbone. Next, we concatenate $\mathbf{x}$ to each sub-sampled vertex of the left hand (l) in its canonical pose to obtain $\mathbf{p}_i = [\mathbf{x}; \mathbf{v}_i] \in \mathbb{R}^{\mathrm{d}+3}$ for all $1 \leqslant i \leqslant \bar{V}_l$ where $\bar{V}_l$ denotes the number of subsampled vertices. All points $\mathbf{p}_i$ are fed to a PointNet [42] followed by a regression head that estimates the distance. The predicted distances are upsampled to the full mesh. For efficiency, we use subsampled vertices for the PointNet and upsample for regression. The remaining interaction fields are estimated via the same network with a shared CNN and PointNet but different heads. InterField-LSTM follows the same formulation except it has an LSTM to aggregate image features in a temporal window to jointly reason about hand-object motion. See more training and baseline details in SupMat.

Results: Figure 5b shows qualitative samples of InterField-LSTM. The predicted 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; they are not network inputs. We find that the predicted fields correlate well with the ground truth. Table 3 shows the performance of our baselines. The results show that modeling the hand-object interaction field over time yields more accurate results (see distance error), and smoother predictions (ACC).

6. Conclusions

We introduce ARCTIC, the first dataset with two hands dexterously manipulating articulated objects that includes high-quality 3D ground-truth for hands, and objects together with synchronized video. ARCTIC has a total of 2.1M RGB images from 8 static views and 1 egocentric view of 10 subjects interacting with 11 articulated objects. We present two tasks on ARCTIC. First is *consistent motion reconstruction*. Given a video, we reconstruct two hands and an articulated object in 3D for every frame, such that their motions are spatio-temporally consistent. The second task is *interaction field estimation*, where we estimate dense relative hand-object distances from images in a video. We present two baselines ArcticNet and InterField for the two tasks respectively, and evaluate them on ARCTIC to lay the foundation for future work.

Future directions: ARCTIC can enable a range of tasks related to hand manipulation with object articulation. First, methods for generating hand-object interaction focus on generating grasps of rigid objects [11, 27], but less work has been done on generating dexterous bimanual manipulation motion with objects [10, 71] and prior work does not generate interaction with articulated objects (*e.g.*, "cutting with scissors"). ARCTIC can enable these new generation tasks, and extend them to the full-body [55] with our SMPL-X ground-truth. Second, we introduce tasks of consistent motion reconstruction and interaction field estimation. Future work could leverage the interaction field representation for pose estimation to improve hand-object contact in reconstruction. Finally, articulated object pose estimators [32] from depth images do not consider humans in the scene. The rendered depth images in ARCTIC can be used to benchmark such methods in more realistic settings.

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, Galina Henz for data capture; Giorgio Becherini, Nima Ghorbani for MoSh++; Priyanka Patel for alignment; Leyre Sánchez Vinuela, Andres Camilo Mendoza Patino, Mustafa Alperen Ekinci for data cleaning; TB for Vicon support; MH, Jakob Reinhardt for object scanning; Taylor McConnell for Vicon support, and data cleaning coordination; Benjamin Pellkofer for IT support. We also thank Xu Chen, Adrian Spurr, Jie Song for insightful discussion. OT and DT were partially funded by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B. DT's work was partially performed at the MPI-IS.

Disclosure: `https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt`

References

[1] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision (ECCV)*, pages 640–653, 2012. 2

[2] Keni Bernardin, Koichi Ogawara, Katsushi Ikeuchi, and Ruediger Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *Transactions on Robotics*, 21(1):47–57, 2005. 3

[3] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 15935–15946, 2022. 2

[4] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A dataset of bodies exhibiting detailed lifelike animated motion. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3D hand shape and pose from images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10843–10852, 2019. 2, 7

[6] 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 *European Conference on Computer Vision (ECCV)*, volume 12358, pages 361–378, 2020. 2, 3

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

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 1, 2, 3, 4

[9] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[10] Yuanpei Chen, Yaodong Yang, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuang Jiang, Stephen Marcus McAleer, Hao Dong, Zongqing Lu, and Song-Chun Zhu. Towards human-level bimanual dexterous manipulation with reinforcement learning. *arXiv preprint arXiv:2206.08686*, 2022. 8

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

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

[13] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J. Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *International Conference on 3D Vision (3DV)*, pages 1–10, 2021. 2, 6

[14] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *Transactions on Human-Machine Systems (THMS)*, 46(1):66–77, 2016. 3

[15] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 409–419, 2018. 2, 3

[16] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D. Twigg, Chengde Wan, James Hays, and Charles C. Kemp. PressureVision: Estimating hand pressure from a single RGB image. *European Conference on Computer Vision (ECCV)*, 13666:328–345, 2022. 3

[17] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021. 2, 3, 7

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020. 1, 2, 3, 4

[19] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3D pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11090–11100, 2022. 1, 2, 3, 4

[20] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020. 2, 7

[21] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 1, 2, 3, 7

[22] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, June 2022. 2, 3

[23] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, volume 13485, pages 281–299, 2022. 2

[24] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D

heatmap regression. In *European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2

[25] Noriko Kamakura, Michiko Matsuo, Harumi Ishii, Fumiko Mitsuboshi, and Yoriko Miura. Patterns of static prehension in normal hands. *American Journal of Occupational Therapy*, 34(7):437–445, 1980. 3

[26] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 7

[27] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 3, 8

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5253–5263, 2020. 2, 5

[29] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 7

[30] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 1, 2, 3, 4

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

[32] Xiaolong Li, He Wang, Li Yi, Leonidas J. Guibas, A. Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3703–3712, 2020. 8

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

[34] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 1, 2, 3, 4

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 2, 4, 5

[36] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, volume 12365, pages 548–564, 2020. 2, 4, 6

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

[38] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *International Conference on Computer Vision (ICCV)*, pages 1163–1172, 2017. 2

[39] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. 3

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3, 4, 5

[41] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A. Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):2883–2896, 2018. 3

[42] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. 8

[43] James M. Rehg and Takeo Kanade. Visual tracking of high DOF articulated structures: An application to human hand tracking. In *European Conference on Computer Vision (ECCV)*, volume 801, pages 35–46, 1994. 2

[44] Grégory Rogez, James Steven Supančič III, and Deva Ramanan. Understanding everyday hands in action from RGB-D images. In *International Conference on Computer Vision (ICCV)*, pages 3889–3897, 2015. 3

[45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 2, 6

[46] István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. Metric-scale truncation-robust heatmaps for 3D human pose estimation. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 407–414, 2020. 7

[47] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21064–21074, 2022. 2

[48] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9866–9875, 2020. 3

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

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

[51] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision (ECCV)*, volume 12362, pages 211–228, 2020. 2

[52] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 89–98, 2018. 2

[53] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *European Conference on Computer Vision (ECCV)*, volume 9906, pages 294–310, 2016. 2, 3

[54] Stefan Stevšič and Otmar Hilliges. Spatial attention improves iterative 6D object pose estimation. In *International Conference on 3D Vision (3DV)*, pages 1070–1078, 2020. 6

[55] Omid Taheri, Vassileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13253–13263, 2022. 8

[56] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, volume 12349, pages 581–600, 2020. 2, 3, 4, 5

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

[58] 3dMDhand system series. `https://3dmd.com/products/`. 4

[59] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[60] Aggeliki Tsoli and Antonis A. Argyros. Joint 3D tracking of a deformable object in interaction with a hand. In *European Conference on Computer Vision (ECCV)*, volume 11218, pages 504–520, 2018. 3

[61] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, 118(2):172–193, 2016. 2

[62] Dimitrios Tzionas and Juergen Gall. A comparison of directional distances for hand pose estimation. In *German Conference on Pattern Recognition (GCPR)*, volume 8142, pages 131–141, 2013. 2

[63] Dimitrios Tzionas and Juergen Gall. 3D object reconstruction from hand-object interactions. In *International Conference on Computer Vision (ICCV)*, pages 729–737, 2015. 3

[64] Dimitrios Tzionas and Juergen Gall. Reconstructing articulated rigged models from RGB-D videos. In *European Conference on Computer Vision Workshops (ECCVw)*, volume 9915, pages 620–633, 2016. 2, 3

[65] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008. 4

[66] Vicon Vantage: Cutting edge, flagship camera with intelligent feedback and resolution. `https://www.vicon.com/hardware/cameras/vantage`. 2, 4

[67] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *International Conference on Computer Vision (ICCV)*, pages 11097–11106, 2021. 2, 3, 7

[68] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[69] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6D pose object detector and refiner. In *International Conference on Computer Vision (ICCV)*, pages 1941–1950, 2019. 6

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

[71] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 8

[72] Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. Single depth view based real-time reconstruction of hand-object interactions. *Transactions on Graphics (TOG)*, 40(3):29:1–29:12, 2021. 3

[73] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *International Conference on Computer Vision (ICCV)*, pages 2354–2364, 2019. 2, 7

[74] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*, volume 13663, pages 1–19, 2022. 3

[75] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5345–5354, 2020. 2

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

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

[78] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 2, 3