Contact4D: A Video Dataset for Whole-Body Human Motion and Finger Contact in Dexterous Operations

Jyun-Ting Song[1] JungEun Kim[3] Jinkun Cao[1] Yu Lei[1] Takuma Yagi[2] Kris Kitani[1]
[1]Carnegie Mellon University [2]AIST [3]KAIST

Abstract

Understanding how humans interact with objects is key to building robust human-centric artificial intelligence. However, this area remains relatively unexplored due to the lack of large-scale datasets. Recent datasets focusing on this issue mainly consist of activities captured entirely in controlled lab environments, and contact annotations are mostly estimated using threshold clips. We introduce Contact4D, a multi-view video dataset for human-object interaction that provides detailed body poses and accurate contact annotations. We use a flexible multi-view capture system to record individuals performing furniture assembly tasks and provide annotations for human detection, tracking, 2D/3D pose estimation, and ground-truth contact. Additionally, we propose a novel processing pipeline to extract accurate hand poses even when they are severely occluded. Contact4D consists of 2M images captured from 19 synchronized cameras across 350 video sequences, spanning diverse environments, varioius furniture types, and unique subjects. We evaluate existing methods for human pose estimation and human-centric contact estimation, demonstrating their inability to generalize to our dataset. Lastly, we fine-tune a pretrained MultiHMR model on Contact4D and observe an improved performance of 56.6% body MPJPE and 26.4% hand MPJPE in scenarios under severe self-occlusion and object occlusion. Code and data are available at https://jyuntins.github.io/Contact4D.

1. Introduction

Modeling whole-body human motion in dexterous manipulation tasks is critical for creating accurate computational models of human activity. In particular, detecting the precise position of the fingers and their contact is one of the essential elements for understanding the motor skills of the human hand and for building visuomotor solutions.

To obtain precise and broadly applicable models of human motion and finger behavior in dexterous tasks, we need datasets that (i) include videos capturing whole-body motion during everyday dexterous activities and (ii) provide finger-level motion together with accurate contact labels. Most existing hand-motion datasets focus on the hands alone and lack full-body motion, whereas full-body datasets rarely capture interactions with objects. Moreover, many datasets infer contact by computing distances between hand- and object-mesh vertices and declaring contact when the minimum distance falls below a fixed threshold [12, 16, 39], which makes the labels sensitive to modeling and reconstruction errors. To mitigate this, some works employ 360-degree hand recordings or confine data collection to controlled environments[12, 15, 57], but such settings limit the generalization of trained models to real-world scenes.

Motivated by this, we introduce Contact4D. Contact4D is a multi-view video dataset for dexterous operations that offers whole-body human poses with precise contact annotations. The dataset is captured with 18 third-person (exocentric) cameras and 1 egocentric camera (Aria glasses[1]). It provides comprehensive 3D whole-body and hand pose/mesh annotations, alongside with ground truth finger contact. In total, Contact4D comprises 100 minutes of footage (over 2 million images) featuring individuals in real-world settings performing furniture assembly task - a complex task encountered in daily life that requires bending, kneeling, reaching, and repositioning around objects.

To annotate whole-body motion in our videos, we design a multi-stage annotation pipeline that lifts multi-view 2D estimations to 3D. We introduce a novel visibility-aware processing framework (Sec. 3.2.2) that leverages per-joint visibility to handle severe occlusion and truncation during hand-object manipulation to extract accurate finger keypoints. We further mitigate noise in keypoint detections to produce robust ground-truth labels through post-optimization (Sec. 3.2.3).

To obtain ground truth contact annotations, we use a custom finger-contact sensor and a commercial foot-pressure sensor. We intentionally avoid bulky sensors, such as sensing textiles [29, 48] or object-mounted pressure sensors [38] to minimize visual discrepancies with real-world scenar-

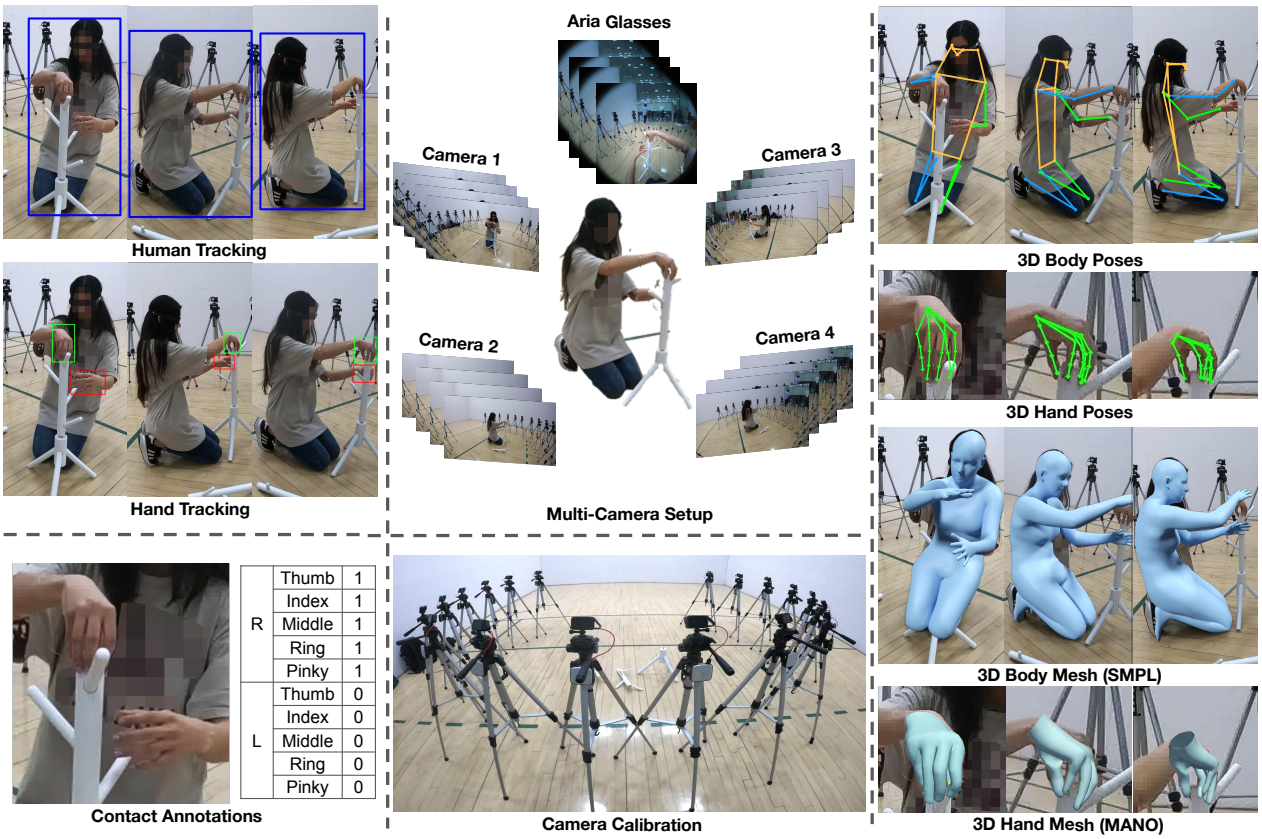

Figure 1. Contact4D is captured using multiple synchronized, calibrated cameras and provides accurate finger contact data collected with customized sensors. It includes comprehensive human annotations, such as human/hand bounding boxes, 2D/3D poses, and meshes. Contact4D supports a broad spectrum of computer vision tasks, including human and hand detection, tracking, pose estimation, reconstruction, and human-centric contact estimation.

ios. Instead, we designed a sticker-based pressure sensor (Fig. 3) that adheres directly to the fingertips, ensuring minimal appearance change.

We evaluate state-of-the-art human pose and human-centric contact estimation models on Contact4D and find that these existing methods achieve limited accuracy for both pose and contact estimation. This indicates that Contact4D poses a significant challenge for current models. Furthermore, we fine-tune a pose-estimation model on Contact4D and observe substantial improvements in both body- and hand-keypoint accuracy. We believe that the dataset's diversity, scale, and multi-modal annotations will serve as a valuable resource for advancing research in human motion estimation and contact-based human behavior analysis.

In summary, we present the dataset Contact4D for the study of whole-body human motion and finger contact. Compared to existing human motion datasets, Contact4D has the following advantages:

- We provide comprehensive whole-body motion annotations, including precise finger keypoints for dexterous manipulations along with multi-view videos.
- We develop a wearable finger contact sensor that accurately captures finger contact information, enabling data collection with only minor appearance changes for human hands.
- Contact4D features a diverse set of subjects and furniture, encompassing a wide range of real-world scenarios. It is the largest dataset that captures the ground-truth contact between humans and objects.

2. Related Work

2.1. Hand Pose and Motion

Hand pose and motion are crucial for understanding human behavior and driving realistic motion animation. Early efforts, such as the FreiHand dataset [10], provided 2D/3D hand pose annotations on static RGB images. More recently, the AssemblyHands dataset [34] introduced video data with dynamic hand-motion annotations, moving beyond static representations. Additional datasets [46, 52] collected in multi-camera motion capture studio further contribute hand motion annotations, with In-

terhand2.6M [31] being one of the largest and most recent. However, the controlled in-studio settings of these datasets limit the study of hand motion in natural, real-world scenarios, which are desirable for robotic tasks and high-fidelity human animation. Contact4D is the first large-scale dataset to provide multi-view hand motion annotations captured in diverse real-world environments.

2.2. Hand Contact in Dexterous Manipulation

Understanding dexterous hand-object manipulation has led to the development of several specialized datasets. For instance, ContactPose [7] focuses on hand grasps and provides static hand pose annotations. GRAB [47] offers object scans and full-body motion data but lacks images for vision models. More recent datasets such as ARCTIC [12], DexYCB [8], HO-3D [13], and OakInk [53] concentrate on hand motion during interactions but do not captured full-body movement. Some works derive contact information via manual image labeling [33, 49, 54], specialized hardware [7, 47], or by approximating contact with distance thresholds [12, 25, 39]. In contrast, Contact4D leverages a custom-designed wearable finger contact sensor to directly measure contact during dexterous manipulation, paired with synchronized multi-view videos to enable vision-based training and evaluation.

2.3. Whole-body Pose and Motion

Whole-body human motion analysis extends traditional body-pose studies by incorporating detailed hand-joint articulations and movements. While several large-scale human motion datasets exist, many lack fine-grained annotations for finger joints [18, 51]. For example, Human3.6M [18] provides indoor multi-view videos with 3D body poses but omits finger-joint annotations. In contrast, COCO-WholeBody [20] offers whole-body pose annotations but is confined to 2D static representations. The CMU Panoptic Studio [21], which has been extended in subsequent works [46, 52] to include hand motion captured from multi-camera setups, is limited by its studio environment, scale, and diversity. In comparison, Contact4D provides large scale, multi-view, whole-body motion with precise finger-joint annotations collected in diverse scenes. We believe this dataset will substantially advance research in human motion estimation, generation, and animation.

3. Contact4D: Data Collection and Processing

In this section, we describe our data collection setup, our proposed whole-body reconstruction pipeline, and the design of our contact sensors. Our objective is to extract accurate whole-body human poses from multi-view videos while providing precise finger-contact annotations.

3.1. Data Collection Setups

We aim to capture a multi-view video dataset for people performing various furniture assembly tasks.

Multi-View Setup, Scenes and Subjects Our multi-view capture system consists of 18 GoPro cameras and one pair of Aria glasses [1]. To better capture the subject's hands, the GoPros are arranged in a fan shape with equal spacing between each camera (see Fig. 1). All GoPro videos are recorded at 4K (3840×2160) resolution and 60 frames per second (FPS), then downsampled to 20 FPS. All cameras are synchronized to ensure temporal consistency across different views. Each sequence features one subject and one randomly selected piece of furniture. To increase diversity, we record 7 subjects across 6 distinct indoor scenes and 6 different pieces of furniture. All subjects are briefed on the research project but are not informed of the furniture item in advance, ensuring natural, unchoreographed behavior during capture.

Camera Calibration We compute the intrinsic and extrinsic parameters for all cameras using each sequence's structure-from-motion (SfM) [43]. The world coordinate system is scaled to metric units and aligned with gravity. This is achieved by computing a scaling factor and rotation matrix via Procrustes analysis [28], comparing the subject's egocentric camera trajectory [1] with the reconstructed camera trajectory.

3.2. Whole-body Human Motion

One key contribution of Contact4D is its large-scale, high-quality whole-body pose annotations. We designed a markerless multi-view reconstruction pipeline that accurately captures 3D joint positions for the body and hands. Because object manipulation often causes severe truncation and occlusion, we add a visibility-aware post-processing step to improve hand-joint accuracy. Finally, we apply refinement and temporal smoothing to produce temporally consistent, high-quality annotations.

3.2.1 Whole-body Pose Reconstruction

Our data processing pipeline builds upon EgoHumans [22] and extends it to accurately extract finger keypoints. The pipeline consists of two components: body reconstruction and hand reconstruction.

Human Body Reconstruction We adopt the 3D body reconstruction methods from EgoHumans [22, 24] to obtain 3D body poses and meshes. This pipeline performs reliably in our setting since the subject remains largely visible across most camera views.

Hand Localization In contrast to localizing a single human during the body reconstruction phase, obtaining con-

sistent hand locations across the video is considerably more challenging. To determine the position of human hands, we utilize the SMPL [27] mesh produced by our human body reconstruction pipeline. Following the parametric hand mesh model MANO [42], we select the 784 vertices nearest to the SMPL hand joints, project them into all exocentric cameras, and compute the corresponding bounding boxes for the subject's hands. We then run a YOLO [40] model specifically trained for hand detection to finally determine the hand bounding boxes by matching its predictions with SMPL outcomes.

2D and 3D Hand Poses We obtain 2D hand poses using an off-the-shelf hand pose estimation model WiLoR [40]. For each camera view, we pass the matched bounding boxes to WiLoR and estimate the MANO parameters. We then regress the output mesh to 3D hand keypoints and project them back onto the camera view to obtain 2D hand poses. To estimate the 3D hand poses from the 2D estimates, we follow the approach used in Egohumans, employing a confidence-weighted multi-view triangulation method [14]. This method utilized RANSAC [11] to identify inlier camera views for each hand keypoint. Since WiLoR does not output confidence scores for its estimates, we use the confidence scores from the corresponding YOLO bounding boxes for all of its 2D hand keypoints.

Mesh Registration Given the reconstructed 3D whole-body pose, we follow HybrIK-X [26] to fit the whole-body mesh to these 3D pose sequences to obtain the mesh registrations. Because human mesh reconstruction and hand mesh reconstruction are typically separate tasks, we provide the mesh of the body in SMPL [27] format and the mesh of hands in MANO [42] format to follow the usual convention of human body/hand reconstruction tasks. The SMPL and MANO mesh can be converted to SMPL-X [36] parametric model for whole-body reconstruction.

3.2.2 Visibility-Aware Post Processing

We aim to provide accurate 3D hand poses even under severe occlusion (e.g., hand-object manipulation). Although the straightforward 'capture-estimation' pipeline performs well for human body pose reconstruction [16, 17, 22], we observed that it fails to yield reliable hand poses under truncation and occlusion, which are common in dexterous operations. This is because, compared to human body poses, hand joint positions are highly sensitive to even minor errors (in either pixel measurements or physical 3D distances). Consequently, any noise in the triangulation process has a significant impact on accuracy, producing suboptimal 3D hand poses.

To overcome this issue, we trained a 2D hand keypoints visibility detector, which outputs binary visibility labels for each hand keypoint given an RGB image as input. Our visibility detector is trained on COCO-WholeBody [20] images, which include ground truth visibility labels. Then, we estimate the visibility of all hand joints from all camera views. Starting from the 3D hand poses obtained via the original triangulation, we design an optimization process to improve the 3D hand joint position. Fig. 2 provides a comparison of the hand joint position before and after our visibility-aware optimization.

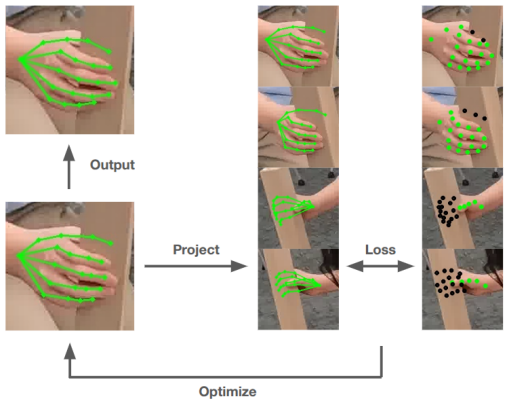

Figure 2. We compare the triangulation and visibility-aware optimization to derive more accurate keypoint annotation. The optimization loss is the distance between the projected 2D hand keypoints and the visible 2D keypoint estimations.

At each optimization iteration, we project the 3D hand keypoints onto all camera views and compute the reprojection loss between these projections and the visible 2D keypoints estimated from WiLoR on the camera images. We adopt the following loss function:

$$L_{\text{proj}}(y_j) = \sum_{c=1}^{C} V(y_j, c) \cdot D\big(P(y_j, c),\, W(j, c)\big), \quad (1)$$

where $y_j \in \mathbb{R}^3$ represents the location of the jth keypoint, with $j \in \{1, \ldots, J\}$ (where J is the total number of hand keypoints), and $c \in \{1, \ldots, C\}$ (with C being the total number of cameras). Here, $V(y_j, c)$ is an indicator function that outputs 1 if the jth keypoint is visible in camera c and 0 otherwise, $P(y_j, c)$ is the projection function that maps the 3D keypoint y_j to the 2D image plane of camera c, $W(j, c)$ is the 2D keypoint location estimated by WiLoR for the jth keypoint in camera c, and D is a distance function (e.g., the Euclidean distance) that measures the discrepancy between the projected keypoint and the 2d estimate from WiLoR.

3.2.3 Refinement and Temporal Smoothing

To obtain more plausible 3D hand pose sequences, we eliminate outlier joints by first computing the average and standard deviation of each knuckle length and discarding keypoints that exhibit excessive jitter by measuring the squared

Figure 3. Our hand contact sensors cause only minor changes to the appearance of the hands and are nearly imperceptible. The commercially available foot sensor is inside the shoes, which is completely invisible.

Euclidean distance between the positions on consecutive timesteps. We then interpolate [45] the discarded keypoints using adjacent 3D hand poses, followed by manual inspection. Lastly, following [50], we optimize the 3D hand poses using the following loss:

$$L_{pose3d}(y) = w_k L_{knuckle}(y) + w_s L_{symm}(y) + w_t L_{temporal}(y) + w_i L_{reg}(y), \quad (2)$$

where $L_{knuckle}$, L_{symm}, $L_{temporal}$, and L_{reg} represent the constant knuckle length, left-right hand symmetry, temporal smoothing, and regularization losses, respectively, and w_k, w_s, w_t, and w_i are constant weighting factors.

3.3. Contact Collection and Processing

Besides the whole-body motion annotations, Contact4D has ground-truth contact labels directly measured from wearable sensors. Although our primary focus is finger tip contact, we also annotate foot contact with a commercially available foot pressure sensor. These sensors are attached to feet and fingers without changing the appearance while providing accurate contact signals.

Finger Contact Sensing To capture accurate finger contact data, we developed wearable finger pressure sensors (Fig. 3) that remain nearly invisible. We attached flexible force-sensing resistors (FSRs) [55] onto a thin, transparent thermoplastic polyurethane (TPU) [6] film and affixed them to each fingertip of the individual for binary contact sensing. The TPU material is transparent, flexible, and soft, which absorbs skin deformation during hand-object interaction, prevents sensor slippage, and preserves the natural appearance of the hands. The FSRs are connected to two Arduino MKR Zero boards [2] (one for each hand) mounted on the individual's arms using enameled wire to prevent electrical shorts, and a small amount of silicone adhesive is applied to the FSR pins to ensure stable readings. Our FSRs are extremely sensitive and respond only to vertical forces. Fig. 6 (right) shows the force response curve of our contact sensor, which can detect force as small as 0.2N (the weight of two paperclips) Unlike typical FSRs, they feature a very thin Mylar substrate [5] (3.5 mil) that prevents the conductive layers from coming into contact when the sensor is bent. This design choice greatly reduces false positives, as bending is not necessarily indicative of finger contact between finger movements, and articulation can also cause deformation. We record hand contact data at 200 fps and manually synchronize the contact signal footage with the camera footage to ensure precise alignment with our videos. For detailed information on our ground-truth hand contact sensor, please refer to the supplemental materials.

Foot Contact Sensing To collect the contact signals of the feet, we used the OpenGO [32] insole pressure sensor, a commercially available device, to record foot contact data. Each sensor unit comprises 16 plantar pressure sensors [41] per foot and operates at a rate of 100 frames per second. The sensors are embedded within the subjects' shoes, making them completely invisible on camera. They are used to sense whether the feet have stepped on a surface with decent pressure. Before each data capture session, the device was calibrated using the subject's actual weight to remove signal noise and ensure optimal accuracy by customized pressure thresholding. We manually synchronize the collected foot contact signals with the camera footage to provide accurate multi-modal annotations.

3.4. Implementation Details

During hand localization, we used all hand predictions from the YOLO model to match each projected SMPL hand bounding box. This straightforward approach retains bounding boxes that YOLO correctly predicts, even when they are assigned an incorrect label. We omitted any 2D poses from bounding boxes with a confidence score below 0.3 during triangulation to avoid excessively erroneous hand poses. Our visibility detector leverages the RTMpose framework [19] and utilizes the CSPNext backbone [30]. Its architecture comprises a convolutional layer, a fully connected layer, a Gated Attention Unit, and an additional convolutional layer. The visibility detection of hand keypoints is formulated as a binary classification task for each finger keypoint. We trained the detector exclusively on images manually annotated from COCO-wholebody for 700 epochs. Before our visibility-aware optimization, we initialize the hand 3D keypoints using confidence-weighted triangulation to speed up the data processing.

3.5. Dataset Statistics

With the introduced data collection and processing pipelines, we build the Contact4D dataset. It features

Dataset	Type	Activity	Body Keypoints	Hand Keypoints	Cameras	Images	Source of Contact
Damon [49]	Image	-	✗	✗	-	5k	Manual Annotation
3DIR [54]	Image	-	✗	✗	-	5k	Manual Annotation
Behave [4]	Video	Dexterous Manipulation	✓	✗	4	60k	Mesh
PROX [15]	Video	Human-Scene Interaction	✓	✓	1	100K	Mesh
InterCap [17]	Video	Dexterous Manipulation	✓	✓	6	204k	Mesh
RICH [16]	Video	Human-Scene Interaction	✓	✓	6∼8	580k	Mesh
ARCTIC [12]	Video	Dexterous Operations	✗	✓	9	2.1M	Mesh
Contact4D (Ours)	Video	Dexterous Manipulation	✓	✓	19	2.2M	Pressure Sensor

Table 1. Comparison with existing related datasets. Contact4D stands out as one of the datasets with the most annotation modalities, sensor-based contact annotation, large-scale multi-view videos, and features dexterous interaction. Contact4D thus provides the resources for generalizable study of diverse tasks.

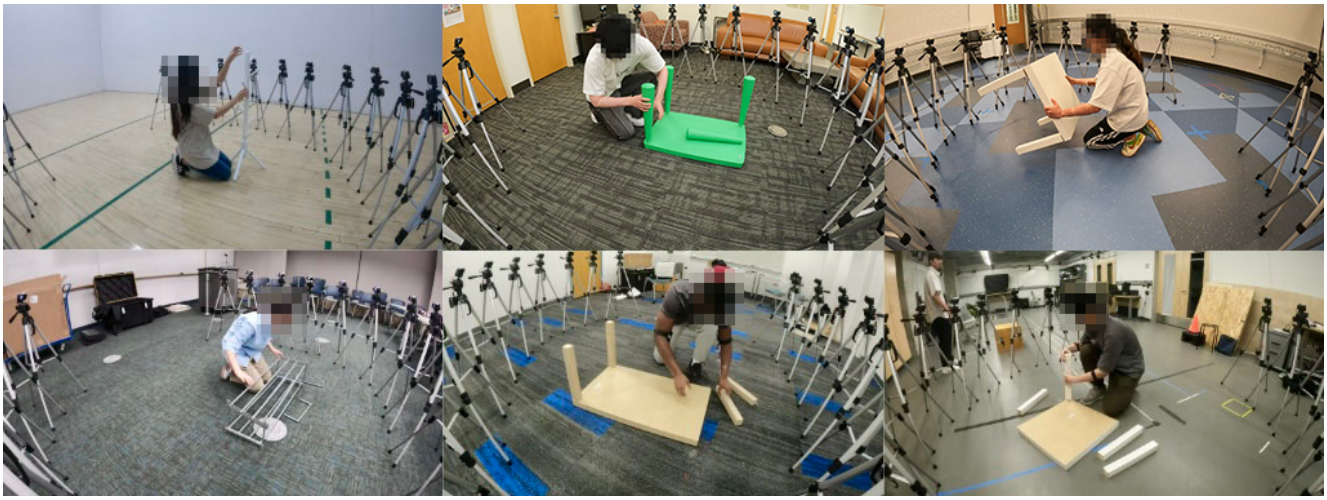

Figure 4. Sample images from Contact4D illustrate its diversity, captured across multiple indoor scenes with participants assembling various pieces of furniture. These varied settings and tasks contribute significantly to the dataset's diversity.

ground-truth finger contact annotations alongside detailed whole-body human keypoints. We collected over 100 minutes of people performing furniture assembly tasks. The videos are temporally synchronized across 19 views (18 exocentric views from GoPro cameras and 1 egocentric view from Aria glasses). Fig. 4 shows sample images captured by our multi-view system. Fig. 5 shows the furnitures present in Contact4D .

Contact4D consists of more than 2M images, divided into 1.6M images for training and 405K images for testing. We manually clipped the videos into 375 sequences across 6 different furniture assembly tasks, ensuring each sequence is at least 15 seconds (300 frames of annotations) long for temporal continuity. The annotation per time step includes camera parameters, bounding boxes, person IDs, 2D/3D human poses, and 3D meshes per subject. All 3D poses at each time step are manually inspected. Fig. 6 (left) shows the motion diversity in Contact4D .

We provide the comparison of Contact4D with existing datasets in Tab. 1. We note that most datasets derive

Figure 5. Furnitures in Contact4D.

contact annotations by computing distances between hand and object mesh vertices and applying a threshold to determine contact. This distance-based estimation are sensitive to mesh reconstruction quality, visual occlusions, and errors in parametric models. Compared to it, our sensor-based contact labels are more reliable.

4. Experiments

In this section, we first evaluate state-of-the-art methods for whole-body and hand pose estimation. We then report results from fine-tuning a pose-estimation model on Contact4D to showcase its usefulness. Next, we benchmark existing contact-estimation methods on Contact4D and pro-

Type	Method	Body Pose				Hand Pose			
		MPJPE↓	PAMPJPE↓	3DPCK↑	AUC↑	MPJPE↓	PAMPJPE↓	3DPCK↑	AUC↑
Hand Pose	HaMeR [37]	–	–	–	–	31.19	6.75	73.36	75.43
Hand Pose	WiLoR [40]	–	–	–	–	24.42	6.13	74.38	76.10
Hand Pose	HandOccNet [35]	–	–	–	–	34.59	10.66	55.14	66.48
Whole-Body	SMPLest-X [56]	95.67	56.15	82.77	52.82	43.71	12.84	31.39	56.55
Whole-Body	Multi-HMR [3]	116.46	78.43	78.16	52.16	47.72	12.21	28.38	55.02
Whole-Body	Multi-HMR-finetuned [3]	50.47	39.02	88.46	54.16	35.11	10.73	53.38	62.47

Table 2. Benchmark of Human Pose Estimation Methods. Contact4D is demonstrated to be challenging for all existing methods.

Method	Finger Contact			Hand Contact			Feet Contact		
	Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑
DECO [49]	0.30	0.19	0.23	0.59	0.26	0.33	0.46	0.99	0.63
BSTRO [16]	0.37	0.14	0.21	0.64	0.26	0.37	0.45	0.89	0.60
Shan [44]	-	-	-	0.53	0.31	0.39	-	-	-
Chen [9]	-	-	-	0.57	0.74	0.64	0.42	0.58	0.49

Table 3. Benchmark of Contact Estimation Methods. All evaluated methods struggle to provide satisfactory performance on Contact4D.

vide in-depth analysis. Finally, we validate our annotation-pipeline design via ablation studies.

4.1. Human Pose Estimation

We evaluate existing human pose estimation methods on Contact4D in two categories: whole-body pose estimation and hand pose estimation (see Tab. 2). We choose Mean Per Joint Position Error (MPJPE) and its Procrustes-aligned version (PA-MPJPE) to measure mean joint errors. In addition, we include 3D Percentage of Correct Keypoints (3DPCK) and Area Under the Curve (AUC) to provide further insight into overall performance. We separate the evaluation on body joints and hand joints.

Whole-body Pose Estimation. We benchmark SMPLest-X [56] and Multi-HMR [3], which are state-of-the-art top-down and bottom-up human pose estimation methods, respectively. Since both SMPLest-X and Multi-HMR predict SMPL-X [36] parameters, we regress body keypoints from the reconstructed meshes and compute keypoint accuracy.

Compared to the pose estimation performance on previous human pose datasets, such as 3DPW [51] and Human3.6M [18], the whole-body pose estimation methods show a significantly worse performance on Contact4D . For example, Multi-HMR [3] achieves an MPJPE of 61.4 mm on 3DPW but 116.46 mm on the Contact4D test set, while SMPLest-X reports 70.5 mm on 3DPW versus 95.67 mm on Contact4D. The severe self- and object-occlusions pose significant challenges for these methods. This suggests a unique value of Contact4D by providing large-scale of human motion which is hardly covered by existing datasets.

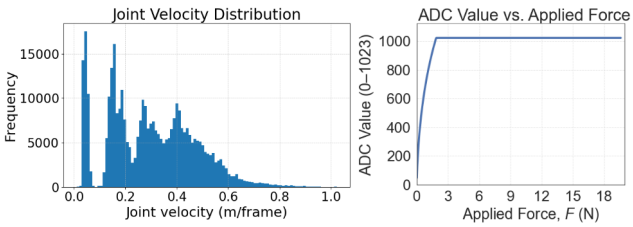

Figure 6. Joints velocity distribution (left) and force response curve for our contact sensor (right).

Hand Pose Estimation. We evaluate WiLoR [40], HaMeR [37], and HandOccNet [35]. All methods receive ground-truth hand bounding boxes. We regress hand joints from the predicted MANO [42] meshes and report joint accuracy.

In our benchmarks, WiLoR and HaMeR perform best on our test set, significantly outperforming the whole-body pose estimation methods. However, the performance is much lower than on existing hand pose datasets, such as HO3D [13] and Freihand [10]. The primary challenge of Contact4D arises from interactions with objects, which cause severe occlusion and truncation. Additionally, the significant performance gap between hand and whole-body pose estimation methods underscores the urgent need for datasets that include whole-body annotations for the scenes of dexterous manipulation to bridge this gap.

Finetuning Multi-HMR. We fine-tune Multi-HMR on the Contact4D training set and observe improvements of **56.6%** for body MPJPE and **26.4%** for hand MPJPE. This highlights Contact4D 's value in helping existing methods adapt to scenes with severe self- and object-occlusion. Notably, fine-tuned Multi-HMR performs on par with HandOccNet on hand-pose metrics, despite HandOccNet being trained specifically for hand-pose estimation.

4.2. Contact Estimation

We evaluate state-of-the-art contact estimation methods on Contact4D in two distinct categories: vertex-level contact estimation and body part-level contact estimation. In Tab. 3, we report the precision, recall, and F1 scores for finger, hand, and foot contacts for selected methods.

Vertex-level Contact Estimation For vertex-level estimation, we evaluate DECO [49] and BSTRO [16]. Given a single image, both methods predict the binary contact label on SMPL vertices. To evaluate their performance, we derive the mesh vertices on fingers and assign the sensor-based contact label to them. We then consider a finger or foot in contact if at least one vertex on its corresponding mesh part is considered in contact.

Body-part Level Contact Estimation For body-part level estimation, we evaluate the methods of Shan [44] and Chen [9]. Chen [9]'s method predicts human-centric contact labels on a 2D image by assigning a body part label to each pixel. In this case, hand or foot contact is considered present if at least one pixel is annotated with the corresponding body part. We do not report a finger contact metric for Chen's method, as fingers are not defined as a separate body part in their method.

Our results reveal challenges for both categories of existing contact estimation methods. For hand contact, DECO and BSTRO achieve F1 scores of 0.33 and 0.37; for finger contact, they drop to 0.23 and 0.21, underscoring the difficulty of estimating fine-grained hand and finger contacts. Both methods exceed 0.6 F1 for foot contact. For body part-level estimation, Shan [44]'s method struggles on hands due to occlusions. In contrast, Chen [9] performs the best for hand contact, likely aided by training on the HOT dataset, which targets human-object contact detection. However, Chen's method performs worse on foot contact. Overall, all evaluated methods struggle to generalize to Contact4D, likely due to the insufficient annotation modality provided by their training data.

4.3. Rationality of annotation pipeline

As introduced in Sec. 3, our dataset processing pipeline separates hand and body pose estimation instead of using a single whole-body pose reconstruction method. Here, we conduct ablated experiments to support this design choice. We select one sequence from Contact4D and employ Sapiens [23] to estimate whole-body poses from all cameras. Then we use the multi-view 2D results for confidence-dependent triangulation (see Sec. 3.2.1) followed by our visibility-aware optimization. As shown in Fig. 7 upper part, Sapiens produces body keypoints that are nearly as accurate as those from our method. However, it fails to extract precise hand keypoints. On the other hand, by combining a body pose estimation method (ViTPose) and a specifically designed hand pose estimation method (WiLoR), the keypoint accuracy is much better than the single whole-body pose estimation method, especially on hand joints. Moreover, we notice that the 2D projection from triangulated 3D joint position is more robust on occluded keypoints than estimation from single-view 2D input. Through this experiment, we verify the design of our current data processing and annotation pipeline.

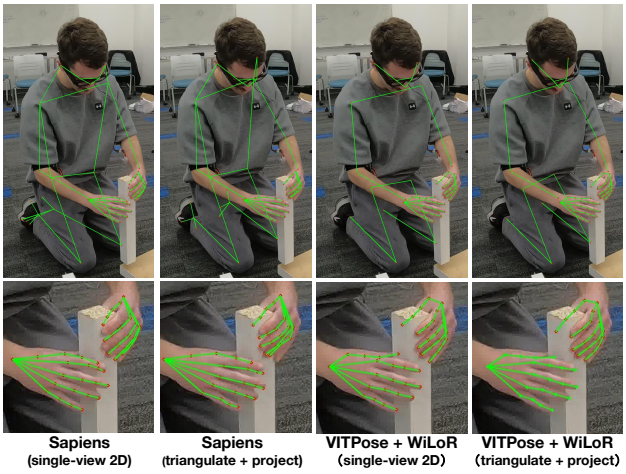

Figure 7. Comparison of keypoint derived from different pipelines. Hand-specific models show a significant advantage over whole-body models on human joint estimation. By triangulating multi-view 2D results and optimizing with visibility labels, the keypoint position can be more robust to occlusions.

5. Conclusion

We present Contact4D, a multi-view video dataset for whole-body human motion with ground-truth fingertip-contact annotations. To obtain accurate labels, we design a visibility-aware multi-view pose-estimation pipeline that reconstructs the body and hands separately. Our evaluations show that state-of-the-art methods perform poorly on Contact4D, underscoring the need for occlusion-rich, whole-body data. Finally, fine-tuning existing models on Contact4D substantially improves both body- and hand-keypoint metrics, mitigating the domain gap.

Acknowledgement. This work was based on results obtained from the Programs for Bridging the Gap between R and D and the IDeal Society (Society 5.0) and Generating Economic and Social Value (BRIDGE) initiative, Practical Global Research in the AI and Robotics Services, implemented by the Cabinet Office, Government of Japan. It was also supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) under the AI Excellence Global Innovative Leader Education Program (RS-2022-00143911).

References

[1] Sally A Applin and Catherine Flick. Facebook's project aria indicates problems for responsible innovation when broadly deploying ar and other pervasive technology in the commons. *Journal of Responsible Technology*, 5:100010, 2021. 1, 3

[2] Store Arduino Arduino. Arduino. *Arduino LLC*, 372, 2015. 5

[3] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 7

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 6

[5] Roger Bollström, Anni Määttänen, Daniel Tobjörk, Petri Ihalainen, Nikolai Kaihovirta, Ronald Österbacka, Jouko Peltonen, and Martti Toivakka. A multilayer coated fiber-based substrate suitable for printed functionality. *Organic Electronics*, 10(5):1020–1023, 2009. 5

[6] A Boubakri, Nader Haddar, K Elleuch, and Yves Bienvenu. Impact of aging conditions on mechanical properties of thermoplastic polyurethane. *Materials & Design*, 31(9):4194–4201, 2010. 5

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

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021. 3

[9] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023. 7, 8

[10] Jimei Yang Bryan Russel Max Argus Christian Zimmermann, Duygu Ceylan and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 7

[11] Konstantinos G Derpanis. Overview of the ransac algorithm. *Image Rochester NY*, 4(1):2–3, 2010. 4

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 1, 3, 6

[13] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 3, 7

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, 2019. 1, 6

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 1, 4, 6, 7, 8

[17] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 4, 6

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 3, 7

[19] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 5

[20] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 4

[21] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[22] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19807–19819, 2023. 3, 4

[23] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models, 2024. 8

[24] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37:107270–107285, 2024. 3

[25] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects

for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021. 3

[26] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 4

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 4

[28] Bin Luo and Edwin R Hancock. Iterative procrustes alignment with the em algorithm. *Image and Vision Computing*, 20(5-6):377–396, 2002. 3

[29] Yiyue Luo, Yunzhu Li, Pratyusha Sharma, Wan Shou, Kui Wu, Michael Foshey, Beichen Li, Tomás Palacios, Antonio Torralba, and Wojciech Matusik. Learning human–environment interactions using conformal tactile textiles. *Nature Electronics*, 4(3):193–201, 2021. 1

[30] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 5

[31] 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)*, 2020. 3

[32] Moticon. Opengo – wireless smart insole system, 2025. Accessed: 2025-03-05. 5

[33] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 33:7841–7851, 2020. 3

[34] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12999–13008, 2023. 2

[35] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 7

[36] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 4, 7

[37] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 7

[38] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2883–2896, 2017. 1

[39] Chandradeep Pokhariya, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2197–2208, 2024. 1, 3

[40] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024. 4, 7

[41] Abdul Hadi Abdul Razak, Aladin Zayegh, Rezaul K Begg, and Yufridin Wahab. Foot plantar pressure measurement system: A review. *Sensors*, 12(7):9884–9912, 2012. 5

[42] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 4, 7

[43] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

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

[45] Gabe Sibley, Larry Matthies, and Gaurav Sukhatme. Sliding window filter with application to planetary landing. *Journal of field robotics*, 27(5):587–608, 2010. 5

[46] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2, 3

[47] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 3

[48] Arvin Tashakori, Zenan Jiang, Amir Servati, Saeid Soltanian, Harishkumar Narayana, Katherine Le, Caroline Nakayama, Chieh-ling Yang, Z Jane Wang, Janice J Eng, et al. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nature Machine Intelligence*, 6(1):106–118, 2024. 1

[49] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 3, 6, 7, 8

[50] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G Narasimhan. Self-supervised multi-view person association and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2794–2808, 2020. 5

[51] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 3, 7

[52] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 2, 3

[53] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 3

[54] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024. 3, 6

[55] SI Yaniger. Force sensing resistors: A review of the technology. *Electro International, 1991*, pages 666–668, 1991. 5

[56] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025. 7

[57] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3. 6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20166–20177, 2023. 1