Interaction Replica: Tracking human–object interaction and scene changes from human motion

Vladimir Guzov [1,2] Julian Chibane [1,2] Riccardo Marin [1] Yannan He [1]
Yunus Saracoglu [1] Torsten Sattler [3] Gerard Pons-Moll[1,2]

[1]University of Tübingen, Tübingen AI Center, Germany, [2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3]CIIRC, Czech Technical University in Prague, Czech Republic

{vladimir.guzov, riccardo.marin, yannan.he, gerard.pons-moll}@uni-tuebingen.de,
jchibane@mpi-inf.mpg.de, yunus.saracoglu@student.uni-tuebingen.de, torsten.sattler@cvut.cz

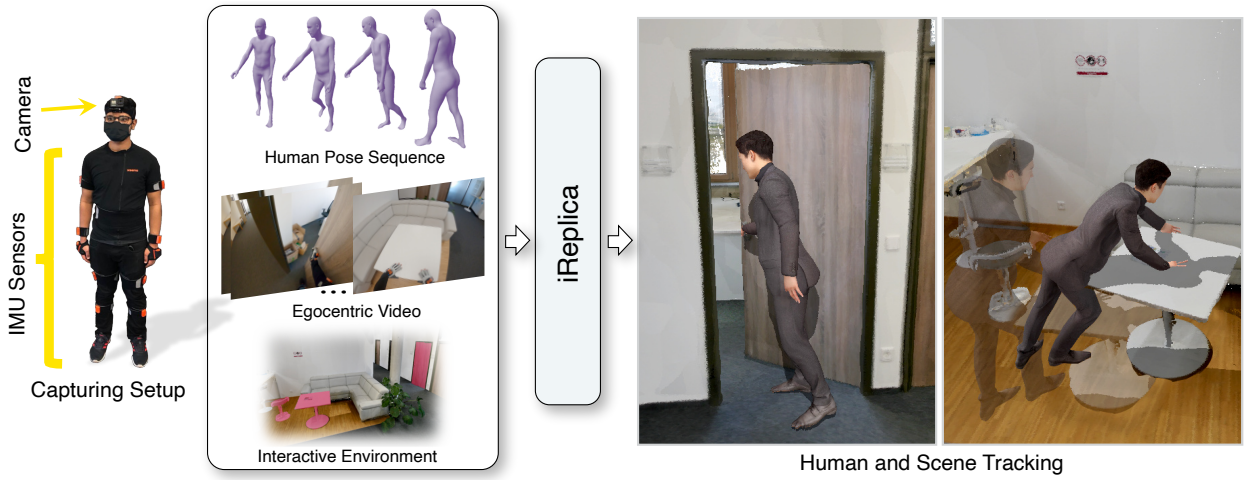

Figure 1. **Interaction Replica (iReplica)**. iReplica estimates location and full 3d pose of a subject within a large 3D scene and dynamically tracks changes made to the scene by the subject - using only wearable sensors (*left*). We obtain an approximate 3D human pose sequence using IMU sensors and use head camera self-localization to localize the subject in the prescanned 3d interactive environment scene. iReplica predicts human-scene contacts and updates the scene in case of interaction.

Abstract

Our world is not static and humans naturally cause changes in their environments through interactions, e.g., opening doors or moving furniture. Modeling changes caused by humans is essential for building digital twins, e.g., in the context of shared physical-virtual spaces (metaverses) and robotics. In order for widespread adoption of such emerging applications, the sensor setup used to capture the interactions needs to be inexpensive and easy-to-use for non-expert users. I.e., interactions should be captured and modeled by simple ego-centric sensors such as a combination of cameras and IMU sensors. Yet, to the best of our knowledge, no work tackling the challenging problem of modeling human-object interactions via such an egocentric sensor setup exists. This paper closes this gap in the literature by developing a novel approach that combines visual localization of humans in the scene with contact-based reasoning about human-object interactions from IMU data. Interestingly, we are able to show that even without visual observations of the interactions, human-object contacts and interactions can be realistically predicted from human motion. Our method, iReplica (Interaction Replica), is an essential first step towards the egocentric capture of human interactions and modeling of dynamic scenes, which is required for future AR/VR applications in immersive virtual universes and for training machines to behave like humans. To encourage the community to work on this challenging and important problem, we will make our new datasets and our code available.

1. Introduction

Current augmented and virtual reality (AR/VR) applications show promising potential: interesting applications include collaborative developments, virtual meeting rooms, and personal assistants that help users navigate the world. While it is clear that for an immersive experience blending real and digital worlds is crucial, the current AR/VR experience is restricted to small spaces, i.e., in general, a few square meters, possibly free from objects. But consider daily actions like moving across rooms, opening and clos-

ing doors, or gathering chairs around a table. Even these simple actions is not easy to capture with present technology, which limits the scope of AR/VR applications.

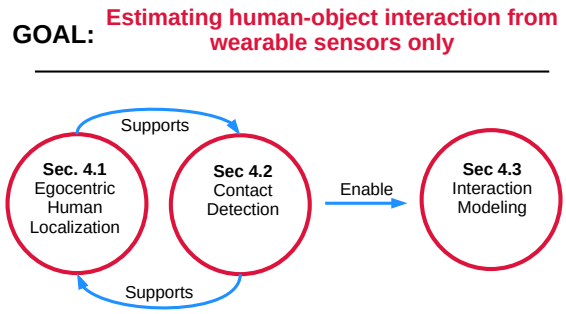

Figure 2. **Problem subdivision.** We demonstrate that joint integration of different sub-research problem improves and support each other. We show this is fundamental to achieve our goal of estimating human-object interaction from wearable sensors only.

The predominant approach for 3D human motion estimation relies on external cameras [14, 21, 26, 27, 29, 34, 40, 67]. Yet, asking non-technical users to mount and calibrate complex multi-camera systems is clearly infeasible. Body-mounted sensors, *e.g.*, cameras and IMU sensors, seem much more ready for mass adoption.

Prior ego-centric trackers such as HPS [15], EgoLocate [62] or HSC4D [9] estimate human movement and position the person by combining head camera visual localization with IMU-based pose estimation. Methods like HPS, however, do not track scene changes. For example, if a person opens and walks through a door, such methods will only localize the person but can not infer the door movement, creating implausible reconstructions; see Figure 7, HPS.

In this work, we address, for the first time, the problem of human-object interaction capture *from wearable sensors only*. Localizing the person with sufficient precision to track scene changes is hard, let alone estimating object motion. A major challenge is that the object is often not visible or is only partially visible in the camera; see Figure 3. In addition, since the head camera is in motion, the object's motion relative to the static world can not be directly inferred.

Since no external sensors can measure the scene changes directly, how can we predict them? Our key observations and findings are that 1) contact poses are distinctive and can be detected without visual clues, 2) knowing contact time stamps can regularize human localization, 3) objects move when the human contacts them [1]. This together has not been fully exploited in the literature. Motivated by this, we propose *iReplica - Interaction Replica*, a novel human-centric method which automatically localizes the human in the scene (1. egocentric human visual localization), detects the time of contact and release with the object (2. contact time detection), and infers object motion based on contacts and human motion (3. interaction modelling). While there exist works in each of these three sub-areas of research, no work integrates them simultaneously.

To successfully integrate the aforementioned three sub-areas of research, we needed several scientific innovations. First, we improve upon human visual localization (HPS [15]) by an optimization which smoothly deforms the human trajectory to match reliable head camera poses and detected spatio-temporal contacts. Second, we train a transformer-based contact time detection approach based solely on the human pose, which achieves a remarkable accuracy of 0.91 and average precision of 0.81. Third, based on the refined human visual localization in the scene and the accurate contact predictions, we infer object motion coherent with the human. Our results demonstrate that joint integration is beneficial (Fig. 2). The contact time information can be used to regularize visual localization by forcing the virtual human to contact the scene. Having precise human localization in the scene, along with contact timestamps, allows us to infer: 1) where contacts occur and 2) the object's motion without seeing the object or contacts in the camera.

During this project, we captured two new datasets. To train a contact detection method, we captured a dataset consisting of 8 subjects and more than 3 hours of human-object interactions annotated with contact time stamps. To validate our proposed method, we captured a dataset with subjects moving and interacting with different objects in large 3D scenes. Our experiments show that iReplica can capture, for the first time, full interactions, including the human motion, its location within the 3D scene and the scene changes, all from wearable sensors alone. We demonstrate that our human-centric approach outperforms baselines which rely on SOTA camera-based contact detection or visual object localization [10, 47].

In summary, our contributions are the following:

1. *Novel Problem & Method*: We are the first to tackle capturing human-object interactions while localizing the human in the scene from wearable sensors alone. We propose a method to address this problem, obtaining, *for the first time*, a digital replica of the human interaction in the scene without any external cameras.
2. *Novel Data & Metrics*: We provide H-contact – a dataset of 2300+ human-object interactions with ground truth annotated contacts. Additionally, we provide EgoHOI – a dataset of human-object interactions in digitally scanned environments. Alongside datasets, we provide metrics to measure the visual plausibility of reconstructed interactions and the accuracy of contact prediction and object localization.

To foster progresses in this new research area, we will release the method code, the evaluation protocol, and the

[1] In this work we only consider static objects moved by the captured human.

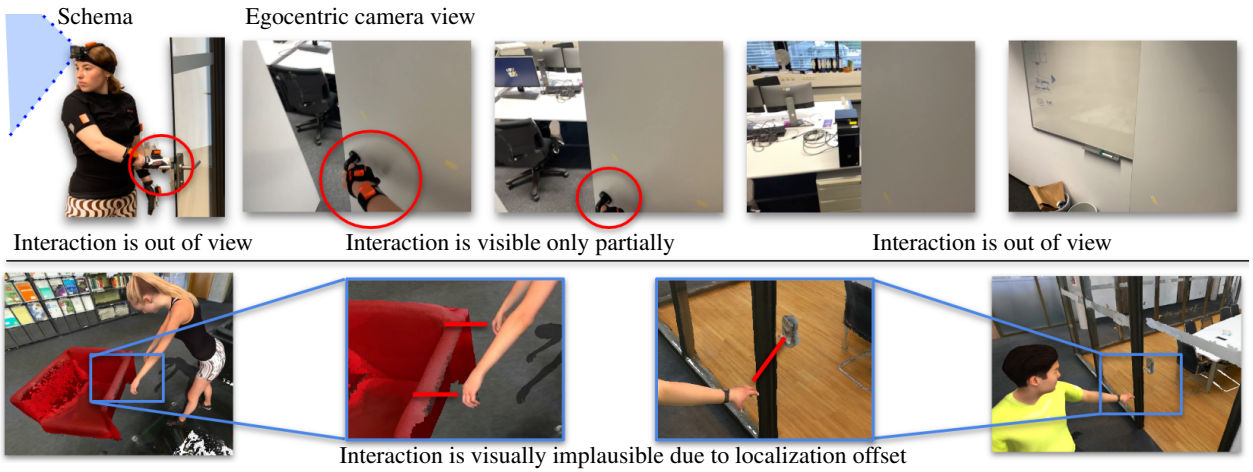

Figure 3. **Challenges.** Top row: We need to predict human-scene contacts (red circles). The prediction is hard because the interactions are frequently not in the camera view. Bottom row: Virtual replica of human pose and localization by prior work HPS [15]. HPS achieved great progress in localizing the human solely by wearable sensors (camera+IMUs). However, for our task, the localization error of 4–16 cm (red lines) lead to visually implausible results for scene interactions.

datasets, including scans of the scenes, human motion capture aligned with them and annotated contact timestamps.

2. Related Work

Human–object interaction. Most current methods that record human–object interactions use external cameras. Methods to capture the full body pose also use external cameras and mostly static scenes [16, 18, 34, 67] or work with a single dynamic object without any scene context [2, 19, 50, 57, 60, 66]; similarly for human–scene and human–object interaction generation methods [17, 49, 51, 69]. RigidFusion [58] tracks objects using an external RGBD sensor. Some methods work with first-person view footage. However, they study the upper body or limited to static cameras, *e.g.*, hand–object pose estimation [12, 30, 33, 38, 52], or do not model dynamic objects [70]. Our method works with body-mounted sensors and a moving camera while capturing the full-body pose and object position.

Embodied research. Body-mounted sensor setups are heavily used to solve various tasks: activity recognition methods [1, 7, 13, 37, 42, 63] use egocentric cameras looking at the body. However, they typically concentrate on capturing the upper body. Many full-body capturing methods [41, 54, 59] work with similar head-mounted setups, but as the cameras are pointed at the person wearing them rather than outwards, these methods ignore the environment around the subject. Some methods work with an outwards-facing camera [23, 64, 65]. However, they do not use any additional sensors to capture the body pose and predict it using motion priors. On the opposite side, methods like [24, 61] use inertial sensors to capture the body pose, but lack of visual cues results in an accumulating position drift, and body poses far from the ground truth. [32] proposes action recognition and localization using a first-person perspective video but does not model scene change. [68] works with a head-mounted camera, but uses it to capture the pose of other subjects, making it an external-camera method. Most related to our method, HPS [15] and following works [9, 62] capture human motion using body-mounted IMUs and a head-mounted camera looking outwards to capture the subject location within a pre-built 3D scan of the environment. We extend HPS by not only tracking the human pose but also an object the person interacts with. Whereas HPS is restricted to static environments and cannot model scene changes caused by human–object interactions, iReplica removes these restrictions.

Visual localization. Visual localization aims to estimate the pose of a camera in a known environment. Current state-of-the-art approaches are based on 2D–3D matches between pixels in the camera image and 3D scene points. These 2D–3D matches are either estimated based on local features [20, 31, 36, 43, 45, 46] or by regressing a 3D point coordinate for each pixel [3, 4, 8, 11, 48][2]. A recent line of works focuses on the robustness of localization algorithms [11, 22, 53, 56], *e.g.*, to illumination, weather, and seasonal changes, as well as to changes caused by human actions (rearranging furniture, *etc.*). These approaches assume that a large enough part of the scene remains static and is observed by the camera to facilitate pose estimation. The second assumption is violated in our scenario and we use IMU-based human pose tracking to bridge gaps where visual localization algorithms will most likely fail. As in [15], for the head-mounted camera localization, we use a state-of-the-art localization pipeline [43, 44]. Similar to the idea of visual-inertial approaches, *e.g.* [6, 25, 35], we use data

[2]We refer the interested reader to [5] for a discussion and comparison of both types of approaches.

from the IMU sensor to stabilize the predicted camera trajectory during periods of low scene visibility or when a lot a scene changes are happening. Note that our main contribution, *i.e.*, jointly reasoning about human and object pose, is not tied to any particular localization algorithm.

3. Problem Setting

Goal. Our goal is to *estimate human-object interaction from wearable sensors only*, without information from external sensors, using only body-mounted IMUs and an egocentric camera. This opens a broad set of interconnected challenges: how to define the interaction? How do we detect the start and the end of it? And how do we track the object's motion without having sensors dedicated?

Assumptions. We assume to have a static 3D scan of the scene, along with a set of marked interactive objects, knowing their initial position and degrees of freedom (*e.g.*, we expect that a sofa can slide on the ground but not be lifted and that a door rotates around a hinge). We will refer to this as *interactive environment* (IE).

Input/Output. We require a set of body-mounted wearable IMUs (we use 17 sensors from XSens [39]) and a video stream from a head-mounted camera. Using only wearable sensors lets us handle large scenes consisting of multiple rooms. Compared to using external cameras, wearable sensors are much more consumer-friendly as they are easier to set up. iReplica outputs a virtual replica of the interaction, *i.e.*, coherent human and object motion in the scene.

4. iReplica

Overview. We obtain initial localization and pose estimation for the person relying on an improved version of HPS [15] (Sec 4.1, Fig 4 A). Our method considers only the human pose at each instant and predicts the probability of contact with an object (Sec 4.2, Fig 4 B). Once the contact is detected, we model the object dynamically as follows: we deform the human trajectory to match the object contact; the object is attached to the human and driven in space according to its degrees of freedom (Sec 4.3, Fig 4 C); when our method infers from the human pose the end of the contact, the object is released (Fig 4 D).

4.1. Egocentric human visual localization.

Problem. Our method is built on a combination of IMUs and head-mounted camera data. Previous methods rely on optimizations to get from these two modalities smooth trajectories estimation [15, 25, 35]. However, no previous approach considers the human's interaction with the scene nor shows extensions to incorporate constraints coming from this. Also, if 10 cm of error (average for HPS [15]) might not seem much for human localization in a building, for human-object interaction (which is our ultimate goal), this can cause dramatic inconsistencies. Instead, we see (and take advantage of) the relation between human localization and contact prediction: solving for contact prediction supports human localization in large volumes; human localization helps detect object contact in time and space.

Trajectory optimization. We start introducing an improvement over the HPS approach [15]. We deploy a simple optimization that is flexible and can be used to incorporate interaction constraints. While we work with 3D trajectories, we consider a 2D optimization since one dimension (gravity axis) is constrained by the ground of the scene. Consider the trajectory described as a 2D curve $\mathbf{l}(\tau) = (x(\tau), y(\tau))$ defined in the time interval $\tau = [\tau_{\text{start}}, \tau_{\text{end}}]$, and a list of K control points $\mathbf{p} = \{\mathbf{p}_i = (x_i, y_i)\}_{i=1}^{K}$ (constraints) at times $\tau_{\text{start}} \leq \tau_1, \ldots, \tau_K \leq \tau_{\text{end}}$. We want to recover a new trajectory $\hat{\mathbf{l}}(\tau) = (\hat{x}(\tau), \hat{y}(\tau))$ that gets close to the control points while not deviating too much from the initial trajectory. We introduce an energy E_{tr} that encodes the trajectory deviation in terms of angles.

$$E_{tr}(\hat{\mathbf{l}}, \mathbf{l}) = \int_{\tau_{\text{start}}}^{\tau_{\text{end}}} \left(\frac{d\hat{\alpha}(\tau)}{d\tau} - \frac{d\alpha(\tau)}{d\tau} \right) d\tau, \qquad (1)$$

where:

$$\hat{\alpha}(\tau) = \text{atan2}\left(\frac{d\hat{y}}{d\tau}, \frac{d\hat{x}}{d\tau}\right), \quad \alpha(\tau) = \text{atan2}\left(\frac{dy}{d\tau}, \frac{dx}{d\tau}\right).$$

Concretely, E_{tr} measures the difference between two trajectories at each instant in terms of direction (angle) variation. We define the difference only in terms of angles since, as pointed out in previous works [15], the total distance tracked by the IMUs is well measured, while the curvature tends to accumulate drift over time.

We then correct the human trajectory by optimizing the following energy:

$$F_{tr}(\mathbf{l}, \mathbf{p}) = \arg\min_{\hat{\alpha}} \left(\sum_{i=1}^{K} (||\hat{\mathbf{l}}(\tau_i) - \mathbf{p}_i||_2) + \lambda E_{tr}(\hat{\mathbf{l}}, \mathbf{l}) \right), \qquad (2)$$

where λ is the global rigidness coefficient, which encodes how much local angles should retain the initial estimation.

Contact-based human trajectory correction. In iReplica, we perform the above optimization two times. We consider the input trajectory recovered by the IMUs, and we optimize it using the control points returned by the camera localization. Then, our method detects the moments of contact along the human motion sequence. For each detection, we select the nearest object in the scene within a reasonable range (*e.g.*, 50 cm). The contact is ignored as a false positive if no object is that close. We select a contact point $\mathbf{p}_c$ as the closest point of the object to the contacting hand. Then, we rerun our optimization again, considering $\mathbf{p}_c$ as the only control point to satisfy. We report details in supplementary.

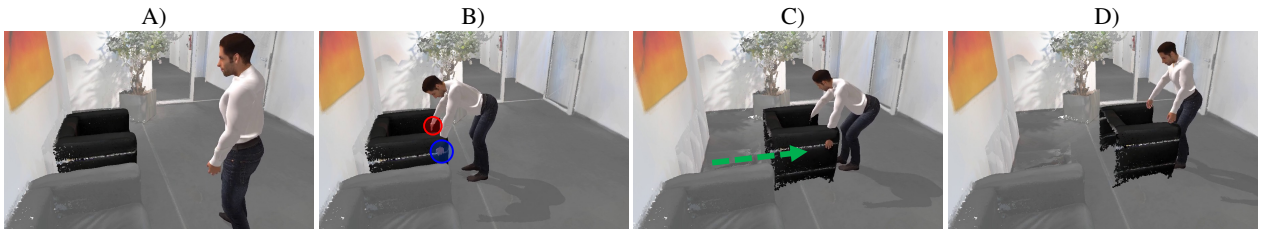

Figure 4. **Overview of iReplica.** iReplica estimates the location and full 3d pose of a subject within a large 3D scene and dynamically tracks changes made to the scene by the subject – using only wearable sensors. We do so in 4 steps: **A)** We obtain an initial localization of the subject in the interactive environment by head camera self-localization. **B)** The start of the interaction is predicted by a neural network. Predictions are provided as contact / no-contact classification of the subject's hands (red and blue areas). The contacts are used to correct inaccurate head camera localization of the subject, snapping the human trajectory smoothly to the object. **C)** The motion of the contacted regions are used to infer the object trajectory (green). **D)** The network predicts the release of the interaction, which is essential to stop object dragging. The algorithm is detailed in Sec. 4.

4.2. Contact detection

Problem. The key ingredient for accurate localization is detecting when and where the user interacts with an object. In this work, we purely focus on human poses (obtained from IMUs) for contacts instead of relying on camera data for multiple reasons: (1) contacts are often not visible, *i.e.*, camera data alone is insufficient for the task. (2) IMUs are cheaper and much more power-efficient than cameras. At the same time, processing their lower-dimensional output requires significantly less compute (and thus power). This makes purely IMU-based contact prediction very attractive for applications running on mobile devices such as AR/VR headsets or robots. Naturally, combining inertial and visual data should improve performance, similar to visual-inertial localization. However, we leave this integration for future work and focus on IMU-only contact detection.

Training data. Existing datasets for human-object contact prediction contain only a limited number of samples per object type [2], or only hand-held objects [50]. In our context, the interaction involves large objects appearing in real scenes. Hence, we collect and annotate a training dataset (H-contact, Sec. 5.1) of ∼680k pose frames ($>$ 3 hours) recorded with 8 subjects wearing IMUs and 12 different objects. Our dataset is noticeably bigger compared to several other human–object and human–scene interaction datasets (BEHAVE [2] contains ∼15k frames, PROX [16] ∼100k).

Transformer. To predict contacts, we train a sequence-to-sequence Transformer [55] to map a sequence of poses to a sequence of per-hand contact probabilities. Specifically, we concatenate 61 SMPL pose vectors in a sequence, forwarding them to an MLP, appending the frame position as positional encoding, and processing them with a Transformer to output a sequence of contact probabilities for each hand. We use a sliding-window approach, and at each instance, we retain only the central (30^{th}) prediction. The contact is considered active once the probability reaches a certain threshold. The architecture is visualized in Fig. 5. To remove false negatives, any gap of $\leq 0.5\,\text{s}$ between two active contacts is filled (*i.e.* marked active). This produced the best results on a validation set (see supplementary).

While focusing on hands is not entirely descriptive of the way in which humans interact with the world, they cover the majority of cases in which humans cause changes in their environments. Our method can easily extend to other body parts; more detailed analyses are left for future works.

Contact intervals. Each group of consecutive frames with active contact is considered as a *contact interval*. If the network predicts the end of the interaction only for one hand, while the other is still considered to be in contact, iReplica splits the contact interval into two consecutive interactions (a two-handed and a one-handed one). Similar cases (*e.g.*, interchanging hands) are treated the same way. Likewise, our method can handle multi-object interaction – please see supplementary for details.

Training details. The network is trained for 100 epochs with a batch size of 100 using the Adam optimizer [28] with a learning rate of 10^{-3} and a standard binary cross-entropy loss. The resulting architecture is lightweight, with 21.9k network parameters in total and an inference time of less than a second per minute of motion (3600 motion windows) on an Nvidia RTX 3090 GPU.

4.3. Interaction modeling

The benefits of iReplica's pose-based contact prediction and human localization are best visualized by dynamically adapting the scene changes as their consequence. Concretely, when contact with an object is predicted, we attach the object to the user; its dynamic is driven by human motion given through IMU pose and the object's degrees of freedom defined in the interactive environment. Please see the supplementary paper for the technical details.

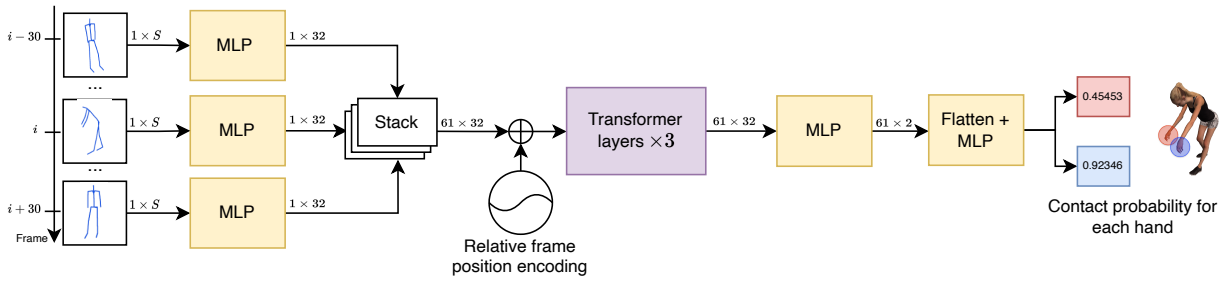

Figure 5. **Contact prediction based on human pose.** Interactions are frequently unobserved in an egocentric view, see Fig. 3 (top row), making contact prediction ill-posed. Instead, we propose to predict from sequences of full 3d human poses. We leverage a transformer-based architecture that takes 61 frames $\{i-30, \ldots, i+30\}$ of SMPL pose vectors of size $S = 69$ and predicts the contact probability for each hand for the middle frame i. See Sec. 4.2 for details.

5. Experiments

5.1. Datasets

Introducing a new task and method, we lack of appropriate datasets, baselines, and evaluation metrics. Thus, we captured and annotated two new datasets: **H-contact** and **EgoHOI**, which we will publicly release, together with our annotation tool.

H-contact is a dataset of human–object interactions, designed to serve as a training set for our contact predictor. We captured and annotated more than 2300 human–object interactions in >3 hours of recordings divided into 30 uninterrupted sequences. A total of around 680k frames, providing interaction for 8 subjects and 12 objects, whose lengths range from 1 to 19 seconds. To obtain ground-truth contact labels, we built a GUI-based annotation tool for this task. Using synchronized video from an external camera, we asked annotators to define the contact classifications.

Egocentric Human–Object Interactions (EgoHOI) is a dataset of humans performing everyday interactions with objects in real scenes recorded with wearable sensors. The sensors are placed directly on the human to allow for large recording volumes not restricted by external camera placement. The wearable setup consists of the IMU-based motion-capturing suit Xsens Awinda [39], allowing us to obtain human pose sequences, and a head-mounted RGB camera for visual localization of the subject in the scene. The dataset also includes the related interactive environments (IE): a 3D scans of the scene, segmented objects and their degrees of freedom. EgoHOI contains interactions with 14 objects (tables, windows, doors, drawers, sofas, chairs, *etc*.) in multiple IE for a total of more than 100k motion frames. We also recorded RGBD data from an external multi-camera setup to measure reconstruction accuracy.

5.2. Baselines

Due to the novelty of the proposed human–object tracking task, no published baselines exist. We introduce novel baselines and briefly describe them, see supp.mat. for details.

HPS. We compare to HPS [15] that localizes the human within the prescanned scene using the images of the head-mounted camera. HPS does not reason about human–object interactions and does not track scene changes.

HPS w/ GT combines HPS with ground-truth data to predict the object motion. It has access to the ground-truth final object pose and ground-truth start and end time of the object interaction. To obtain the object motion estimate, it linearly interpolates the object poses in the time window. Obviously, the required ground-truth data for HPS w/ GT is not available in real-world applications. This baseline is used to show that a simple linear interpolation model is not sufficient for real-world scenes.

HPS w/ RGB Obj. Loc. localizes the object using solely the RGB frame from the head-mounted camera. As HPS uses visual localization algorithm [43] to localize the camera in the scene, we use the same process to localize the interacted object w.r.t. the camera. Knowing the relative pose of the object to the camera localization in world space, we can estimate the object pose in world space. To simplify the matching process, this baseline uses GT object segmentations of the objects of interest.

iReplica w/ HOD [47] and iReplica w/ VISOR [10]. Our approach predicts human-object contacts based on human pose information. Alternatively, the head-mounted RGB camera can be used to make these predictions. As baselines, we use two state-of-the-art, pretrained, RGB-based hand-contact understanding methods: **HOD** [47] and **VISOR** [10]; both predict 2D hand and object locations, and contact probabilities per hand. We use the contact probabilities as a drop-in replacement for the contact predictor in iReplica, while keeping the rest of the method fixed.

5.3. Results

Qualitative results. Fig. 6 shows our results for sample frames from interactions in multiple scenes. Videos are included in the supp. mat. and we urge the reader to look at them as interactions are best judged in motion. Our results show that egocentric motion data alone is sufficient to localize the human in the scene, model the interaction

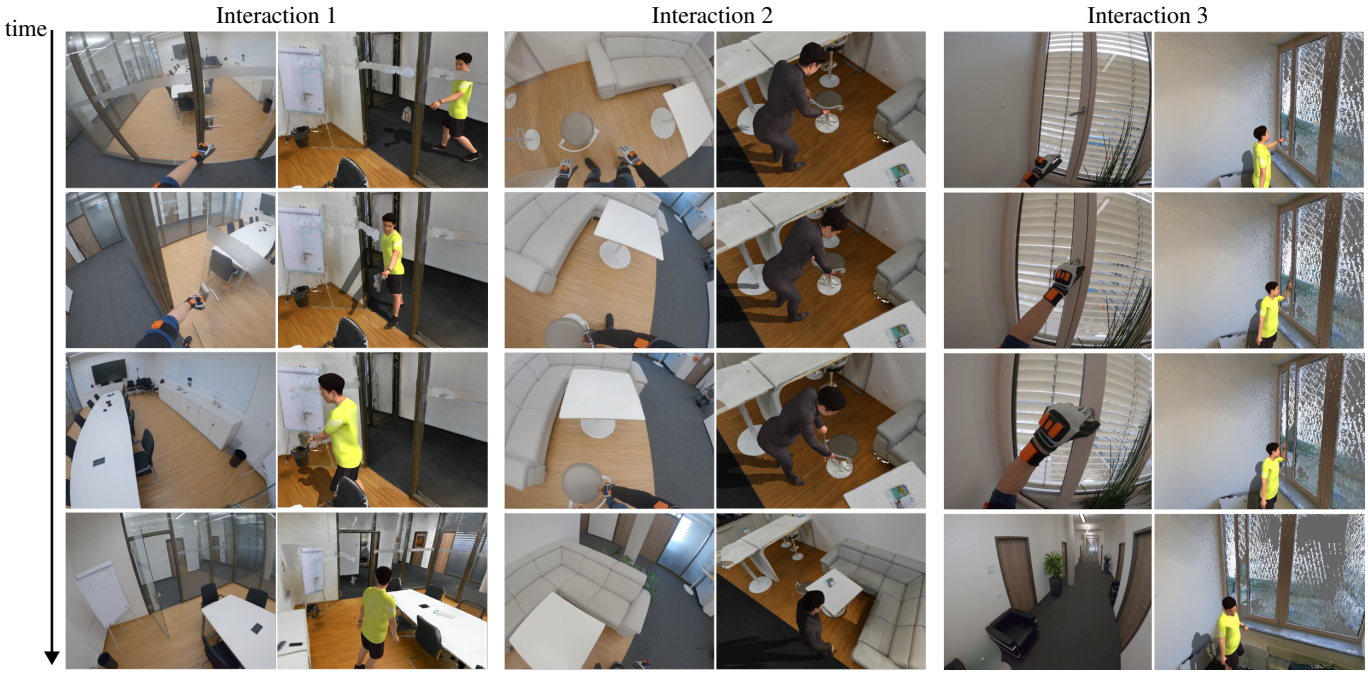

Figure 6. **Qualitative results.** We show three examples of human interaction, pairing the head-mounted camera view with the interaction modeling achieved by iReplica. The object is not always visible during the interaction (Interaction 1), hand grasping can be difficult to understand from the camera (Interaction 2), or object occludes a majority of the first person view (Interaction 3). By relying on human-centric contact detection, iReplica achieves reliable modeling in all these challenging scenarios. Please see our video for more results.

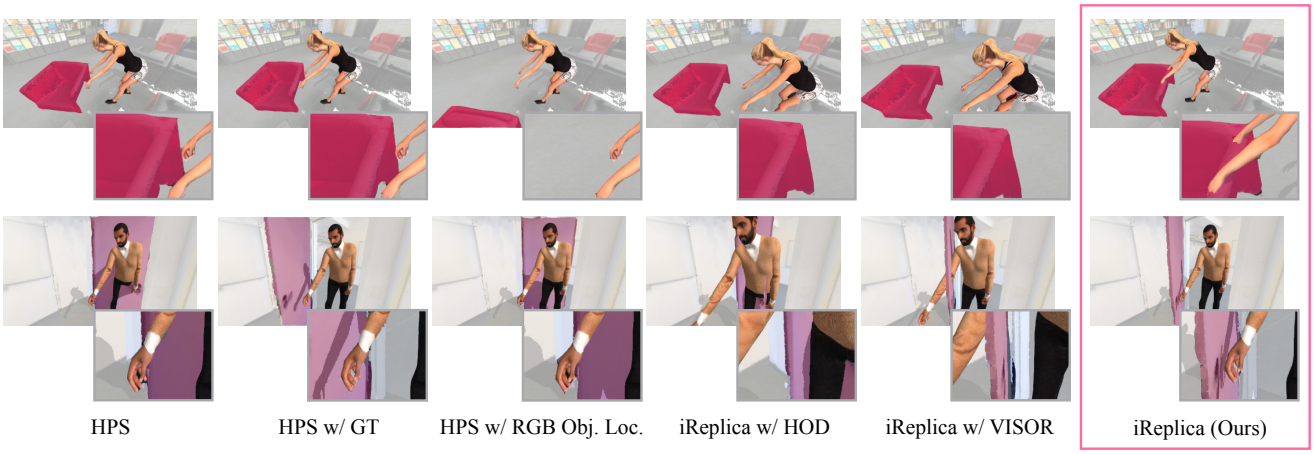

Figure 7. **Qualitative comparison.** We visually compare iReplica (ours) to the baseline methods (interacted object highlighted in red for visual clarity). For the sofa sequence (top row) no baseline can dynamically track the sofa and correctly place the subjects' hands. Similarly, on the door sequence, notice that the door is incorrectly placed by all baselines and that the hand is not in contact with the handle. In contrast, iReplica obtains visually plausible results, by adjusting human and object locations to satisfy contact constraints. Please see comparisons in our supp. video.

between the human and objects, and update the scene accordingly. Our contact predictor allows iReplica to estimate object tracking only based on the human pose; *e.g.*, doors, chairs, and tables can be interacted with in the scene.

Qualitative comparisons to baselines. Fig. 7 visually compares iReplica to our baselines by showing individual frames from some of the interactions. The interactions are best viewed in the supp. video. *HPS* does not track scene changes and thus obtains unrealistic motions. For example, the door opening is not tracked. The subject should have opened the door with the handle, but the door is still closed. (Fig. 7, door). Or the sofa is dragged by the subject, but the object stands still. The sofa, therefore, is visibly not in contact with the subject's hands (Fig. 7, sofa). *HPS w/ GT* linearly interpolates the object motion given ground-truth object start and end pose and contact times. Resulting interaction is not visually plausible, due to a large mismatch between the hands of the subject and the sofa. Human-object

interaction motion is highly non-linear in its nature, so linear approximations seem unrealistic: according to our user study, iReplica results were preferred over this baseline and HPS in 84.2% of the cases (see suppl.). *HPS w/ Obj. Loc.* fails to detect the accurate object position during the interaction, resulting in misaligned results, to the point that it fails to localize the object at all (*e.g.*, Tab. 2, Box). *iReplica w/ HOD* and *iReplica w/ VISOR* both suffer from false negatives, resulting in a sudden contact loss in the middle of the interaction and missed contacts between object and subject. *iReplica* ensures that the subject's hands are close to the object during the whole interaction – a key aspect of visual plausibility not achieved by the baselines. This shows the value of iReplica's correction of the human trajectory based on the human–object interaction.

Reconstruction quality compared to real scenes. We quantitatively validate iReplica's object and human localization results in terms of the reconstruction quality with respect to the original scene. We measure deviations from the virtual replica to the real scene using the EgoHOI dataset. Tab. 1 shows the object localization accuracy at the end of the interaction, where the GT object pose was annotated. On average, iReplica improves the results considerably (col. *All*). All object types are localized with a distance below 10 cm and an orientation error below 13 degrees.

Ablation of Contact-based human trajectory correction. We ablate the proposed contact-based human trajectory correction, by excluding it from iReplica. We report the results in Table 1 and 2 (**iReplica w/o Contact corr.**). We show that method greatly benefits from the proposed correction. Moreover, we measure the error of human localization with (iReplica) and without (HPS [15]) correction on a special sequence that additionally has ground truth point clouds obtained via an external multi-view system of depth cameras. iReplica again improves upon HPS – see the supp. mat. for details.

Visual plausibility. We aim to measure the visual plausibility of iReplica results compared to the baselines. One key factor of the plausibility of interactions is that the human needs to be in contact with the object. To validate this, we measure the mean distance from the object to the interacting hand, see Tab. 2. iReplica keeps this distance below 3 cm. Keeping track of contacts and using them for attaching the object and the human motion creates the lowest human–object distances. This can be seen qualitatively in the videos.

Contact prediction accuracy. We benchmark the accuracy of iReplica contact prediction in isolation in Tab. 3, comparing it to our two RGB contact prediction baselines, HOD [47] and VISOR [10]. We treat the network predictions as probabilities in a binary classification task and compute 4 metrics: Average Precision (AP), Precision, Recall and Accuracy on the binarization threshold of 0.5. Our contact prediction, solely based on the 3D human pose, significantly outperforms the RGB-based reasoning - one cause is that interaction is not always visible in the camera. Once more, we remark on how 3D human poses in isolation is a highly informative indicator of interaction contacts.

Error ↓	Method	Door	Sofa	Table	Box	All
Distance (in cm)	**HPS**	79.27	69.54	25.31	41.92	60.81
	HPS w/ RGB Obj. Loc.	28.66	1684.06	119.59	—	597.83
	iReplica w/ HOD [47]	57.50	55.78	1.62	**3.33**	38.58
	iReplica w/ VISOR [10]	43.40	66.74	5.98	11.31	39.59
	iReplica w/o Contact corr.	18.54	11.70	1.84	7.79	11.68
	iReplica (Ours)	**9.97**	**6.66**	**0.90**	7.09	**6.88**
Angle (in °)	**HPS**	109.19	23.53	12.16	3.76	46.89
	HPS w/ RGB Obj. Loc.	34.36	118.02	60.08	—	61.43
	iReplica w/ HOD [47]	75.74	7.74	0.78	**2.71**	28.41
	iReplica w/ VISOR [10]	56.64	17.36	2.87	12.78	27.27
	iReplica w/o Contact corr.	22.16	**5.83**	0.88	4.81	10.28
	iReplica (Ours)	**12.94**	**5.83**	**0.43**	4.81	**7.13**

Table 1. **Object localization accuracy.** Distance (in cm) and angle (in °) between object center at the end of the interaction in the GT pose and object center in the pose predicted by the algorithm.

Class label	Door	Sofa	Table	Box	All
HPS	46.00	38.32	26.35	6.64	33.61
HPS w/ GT	17.28	6.90	7.55	6.74	10.44
HPS w/ RGB Obj. Loc.	65.26	724.63	136.27	—	287.12
iReplica w/ HOD [47]	48.42	35.96	13.31	5.52	31.26
iReplica w/ VISOR [10]	33.76	51.14	13.39	**3.84**	31.17
iReplica w/o Contact corr.	18.15	9.80	6.89	5.45	11.37
iReplica (Ours)	**2.83**	**1.46**	**3.49**	5.49	**2.93**

Table 2. **Visual plausibility of human-object interaction.** Mean distance between the object and the contacting hand (in cm) over the interaction time interval.

Contact predictor	AP ↑	Precision@0.5 ↑	Recall@0.5 ↑	Accuracy@0.5 ↑
HOD [47]	0.044	0.251	0.818	0.364
VISOR [10]	0.217	0.313	0.098	0.732
Ours	**0.807**	**0.786**	**0.880**	**0.905**

Table 3. **Contact prediction performance.** Metrics obtained on our test set with subjects that are not appearing in training data.

6. Discussion and Conclusion

In this work we proposed the novel problem of capturing human-scene interactions and dynamic 3D scenes solely from wearable sensors - that is, IMUs and a head-mounted camera, and not relying on any external cameras or object trackers. Our results show that egocentric motion data alone can be used to localize the human in the scene, model the interaction between the human and objects, and update the scene accordingly. Our contact predictor allows iReplica to estimate object tracking only based on human pose. *E.g.*, doors in the virtual scene can be opened or objects (sofas, boxes, tables) can be displaced. Inaccuracies in human localization from HPS are corrected plausibly as iReplica ensures that the human and the interacted object are close to

each other.

Future work can investigate including physics simulations for interaction modeling, such as those available in game engines. Instead of prescanning the IE, future work could explore how to build scene reconstructions (e.g. with SLAM-based models) and annotations on the fly (e.g. using ScanNet instance-segmentation models). We believe this work constitutes a first step towards an exciting new research direction of capturing human-scene interactions and dynamic 3D scenes from wearable sensors. We will release the iReplica datasets and code to inspire work on the capture of human-scene interaction in dynamic environments.

Acknowledgments: Special thanks to RVH team members, and reviewers, their feedback helped improve the manuscript. The project was made possible by funding from the Carl Zeiss Foundation. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and the Czech Science Foundation (GA ČR) EXPRO (grant no. 23-07973X). Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 - Project number 390727645. Julian Chibane is a fellow of the Meta Research PhD Fellowship Program - area: AR/VR Human Understanding. Riccardo Marin is supported by an Alexander von Humboldt Foundation Research Fellowship.

References

[1] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, and C.V. Jawahar. Unsupervised learning of deep feature representation for clustering egocentric actions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1447–1453, 2017. 3

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3, 5

[3] Eric Brachmann and Carsten Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *CVPR*, 2018. 3

[4] Eric Brachmann and Carsten Rother. Visual camera relocalization from RGB and RGB-D images using DSAC. *arXiv:2002.12324*, 2020. 3

[5] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *ICCV*, 2021. 3

[6] Carlos Campos, Richard Elvira, Juan J. Gómez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 3

[7] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[8] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Victor A. Prisacariu, Luigi Di Stefano, and Philip H. S. Torr. Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 3

[9] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6792–6802, 2022. 2, 3

[10] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 2, 6, 8

[11] Siyan Dong, Qingnan Fan, He Wang, Ji Shi, Li Yi, Thomas Funkhouser, Baoquan Chen, and Leonidas J Guibas. Robust neural routing through space partitions for camera relocalization in dynamic indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8544–8554, 2021. 3

[12] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020. 3

[13] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *2011 international conference on computer vision*, pages 407–414. IEEE, 2011. 3

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2

[15] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 2, 3, 4, 6, 8

[16] 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. 3, 5

[17] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 3

[18] 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. 3

[19] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 281–299. Springer, 2022. 3

[20] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *CVPR*, 2009. 3

[21] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 2

[22] Ara Jafarzadeh, Manuel López Antequera, Pau Gargallo, Yubin Kuang, Carl Toft, Fredrik Kahl, and Torsten Sattler. Crowddriven: A new challenging dataset for outdoor visual localization. In *ICCV*, 2021. 3

[23] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 3

[24] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[25] Eagle S. Jones and Stefano Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *The International Journal of Robotics Research*, 30(4):407–430, 2011. 3, 4

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018. 2

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[29] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[30] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. *arXiv preprint arXiv:2104.11181*, 2021. 3

[31] Yunpeng Li, Noag Snavely, Dan P. Huttenlocher, and Pascal Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *ECCV*, 2012. 3

[32] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *European Conference on Computer Vision*, pages 621–638. Springer, 2022. 3

[33] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 3

[34] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *arXiv preprint arXiv:2206.09106*, 2022. 2, 3

[35] Simon Lynen, Torsten Sattler, Michael Bosse, Joel A Hesch, Marc Pollefeys, and Roland Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *Robotics: Science and Systems*, page 1, 2015. 3, 4

[36] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39 (9):1061–1084, 2020. 3

[37] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going deeper into first-person activity recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016. 3

[38] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1898–1912, 2019. 3

[39] Monique Paulich, Martin Schepers, Nina Rudigkeit, and G. Bellusci. *Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications*, 2018. 4, 6

[40] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[41] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):162, 2016. 3

[42] Grégory Rogez, James S Supancic, and Deva Ramanan. First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4325–4333, 2015. 3

[43] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 3, 6

[44] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*, 2020. 3

[45] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *PAMI*, 2017. 3

[46] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler. Semantic Visual Localization. In *CVPR*, 2018. 3

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

[48] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[49] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 3

[50] 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, 5

[51] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022. 3

[52] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 4511–4520, 2019. 3

[53] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, Fredrik Kahl, and Torsten Sattler. Long-Term Visual Localization Revisited. *TPAMI*, pages 1–1, 2020. 3

[54] Denis Tome, Thiemo Alldeick, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando de la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[56] Johanna Wald, Torsten Sattler, Stuart Golodetz, Tommaso Cavallari, and Federico Tombari. Beyond controlled environments: 3d camera re-localization in changing indoor scenes. In *European Conference on Computer Vision*, pages 467–487. Springer, 2020. 3

[57] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021. 3

[58] Yu-Shiang Wong, Changjian Li, Matthias Niessner, and Niloy J. Mitra. Rigidfusion: Rgb-d scene reconstruction with rigidly-moving objects. *Computer Graphics Forum*, 40(2), 2021. 3

[59] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^{2}Cap^2 : Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019. 3

[60] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3

[61] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3

[62] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2, 3

[63] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, and Y. Taniguchi. Egocentric articulated pose tracking for action recognition. In *International Conference on Machine Vision Applications (MVA)*, 2015. 3

[64] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 3

[65] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[66] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[67] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11343–11353, 2021. 2, 3

[68] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 180–200. Springer, 2022. 3

[69] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3

[70] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 3