EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices

Siwei Zhang[1] Qianli Ma[1] Yan Zhang[1] Zhiyin Qian[1] Taein Kwon[1]
Marc Pollefeys[1,2] Federica Bogo[2]⋆ Siyu Tang[1]

[1]ETH Zürich [2]Microsoft
{siwei.zhang, qianli.ma, yan.zhang, taein.kwon, marc.pollefeys,
siyu.tang}@inf.ethz.ch zhqian@ethz.ch fbogo@fb.com

Abstract. Understanding social interactions from *egocentric* views is crucial for many applications, ranging from assistive robotics to AR/VR. Key to reasoning about interactions is to understand the body pose and motion of the interaction partner from the egocentric view. However, research in this area is severely hindered by the lack of datasets. Existing datasets are limited in terms of either size, capture/annotation modalities, ground-truth quality, or interaction diversity. We fill this gap by proposing EgoBody, a novel large-scale dataset for human pose, shape and motion estimation from egocentric views, during interactions in complex 3D scenes. We employ Microsoft HoloLens2 headsets to record rich egocentric data streams (including RGB, depth, eye gaze, head and hand tracking). To obtain accurate 3D ground truth, we calibrate the headset with a multi-Kinect rig and fit expressive SMPL-X body meshes to multi-view RGB-D frames, reconstructing 3D human shapes and poses relative to the scene, over time. We collect 125 sequences, spanning diverse interaction scenarios, and propose the first benchmark for 3D full-body pose and shape estimation of the interaction partner from egocentric views. We extensively evaluate state-of-the-art methods, highlight their limitations in the egocentric scenario, and address such limitations leveraging our high-quality annotations. Data and code are available at https://sanweiliti.github.io/egobody/egobody.html.

Keywords: pose estimation, egocentric view, motion capture, dataset

1 Introduction

Humans constantly interact and communicate with each other; understanding our social interaction partners' motions, intentions and emotions is almost instinctive for us. However, the same does not hold for machines. A first step towards automated human interaction understanding is the estimation of the 3D body pose, shape and motion of the social interaction partner ("*interactee*") from egocentric views, *e.g.* from head-mounted devices (HMD). Addressing this challenging problem is crucial for many applications, ranging from assistive robotics to Augmented and Virtual Reality (AR/VR), where sensors typically perceive

⋆ Now at Meta Reality Labs Research.

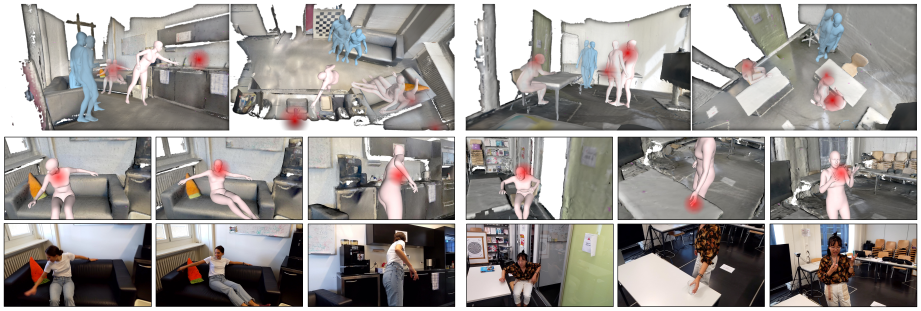

Fig. 1: EgoBody is a large-scale dataset capturing ground-truth 3D human motions during social interactions in 3D scenes. Given two interacting subjects, we leverage a lightweight multi-camera rig to reconstruct their 3D shape and pose over time (top row). One of the subjects (blue) wears a head-mounted device, synchronized with the rig, capturing egocentric multi-modal data like eye gaze tracking (red circles in first two rows) and RGB images (bottom).

the interactee from the egocentric view. Despite its importance, the problem has received little attention in the literature so far. While there are a large number of methods for full-body pose (and sometimes also shape) estimation from RGB(D) frames [14, 20, 30, 40, 41, 45, 48, 60, 73, 82, 91, 94, 99–102], they tend to perform poorly on data captured with an HMD (see Sec. 5). Indeed, this setup brings its own unique challenges, which most methods have not explicitly addressed so far. Any method aiming at understanding the pose and shape of the interactee must deal with severe body truncations, motion blur (exacerbated by the embodied movement of the HMD), people entering/exiting the field of view, to name a few.

A reason for such limited attention is the lack of data. On one hand, most human motion datasets are captured by *third-person-view* cameras without egocentric frames [23,30,34,37,38,65,86,103], which do not faithfully replicate AR/VR scenarios; most capture only one subject at a time, without interactions [30,34]. On the other hand, existing *egocentric* datasets are limited in terms of annotation modalities, scale and interaction diversity. They either focus on coarse-level interaction/action labels [21,56,67,75,80], or provide only the camera wearer's pose without considering [84,85,93,98], or with very limited data involving [29], the interactee. You2Me [68] collects egocentric RGB frames of two-people interactions, annotated with 3D skeletons, without 3D scene context, nor the body shape. Recently, Ego4D [27] collects a large amount of egocentric videos for various tasks including action and social interaction understanding, but without 3D ground truth for human pose, shape and motions.

To fill this gap, we propose EgoBody, a unique, large-scale egocentric dataset capturing high-quality 3D human motions during social interactions. We focus on 2-people interaction cases, and define interaction scenarios based on the social interaction categories studied in sociology [8]. Unlike most existing datasets that only provide RGB streams, EgoBody collects egocentric multi-modal data, with

Table 1: Comparison with existing image-based datasets with 3D human pose annotations. "Fr.#" denotes frame numbers. "3rd-PV" and "Ego" refers to the third-person-view and egocentric view, respectively. "Mesh" refers to the body mesh. "Interact" refers to social interactions. "Global-Cfg." refers to global translation and rotation.

Dataset	Fr.#	Sub.#	3rd-PV	Ego	Mesh	Gaze	3D-Scene	Interact	Global-Cfg.
TNT15 [64]	13k	4	✓		✓				✓
3DPW [63]	51k	7	✓		✓*			✓	
PROX [30]	100k	20	✓		✓*		✓		✓
Panoptic [38]	297k	180+	✓					✓	✓
HUMBI [97]	380k	772	✓		✓*	✓			✓
TotalCapture [86]	1,900k	5	✓						✓
Human3.6M [34]	3,600k	11	✓		✓				✓
Mo2Cap2 [93]	15k	5		✓					
You2Me [68]	150k	10		✓				✓	
HPS [29]	300k	7		✓	✓*		✓		✓
Ours	220k	36	✓	✓	✓*	✓	✓	✓	✓

* Body Mesh defined by parametric body models.

accurate 3D human shape, pose and motion ground-truth for both interacting subjects, accompanied by eye gaze tracking for the camera wearer. Furthermore, EgoBody includes accurate 3D scene reconstructions, providing a holistic and consistent 3D understanding of the physical world around the camera wearer.

The egocentric data is captured with a Microsoft HoloLens2 headset [3], which provides rich multi-modal streams: RGB, depth, head, hand and eye gaze tracking, correlated in space and time. In particular, eye gaze carries vital information about human attention during interactions. By providing eye gaze tracking synchronized with other modalities, EgoBody opens the door to study relationships between human attention, interactions and motions. We obtain high-quality 3D human shape and motion annotations in an automated way, by leveraging a marker-less motion capture approach. Namely, we utilize a multi-camera rig consisting of multiple Azure Kinects [1] as our motion capture system.

However, combining raw data streams from the egocentric- and the third-person-view remains highly challenging due to hardware limitations. Specifically, the Kinect-HoloLens2 calibration exhibit inaccuracies due to not perfectly accurate factory calibration and tracking drift. We address this by proposing a refinement scheme based on body keypoints. With carefully calibrated data, we further build an efficient motion capture pipeline based on [102] to fit the SMPL-X body model [73] to multi-view and egocentric RGB-D data, reconstructing accurate 3D full-body meshes for both the camera wearer and the interactee. In this way, we get accurate and well calibrated ground truth across all sensor coordinates, as well as the world coordinate, which is not available in most existing datasets. The setup is lightweight and easy to deploy in various environments.

With EgoBody we propose the first benchmark for 3D human pose and shape estimation (3DHPS) of the interactee, in interactions captured by the HMD. By evaluating state-of-the-art 3DHPS methods on the EgoBody's test set, we carefully analyze and highlight the limitations of existing methods in this egocentric setup. We show the usefulness of EgoBody by fine-tuning three recent methods [36,48,57] on its training set, obtaining significantly improved performance on our *test set*. Finally, in a cross-dataset evaluation we show how models fine-tuned on EgoBody also achieve a better performance on the *You2Me* dataset [68].
Contributions. In summary, we: **(1)** provide the first large-scale egocentric dataset, EgoBody, comprising both egocentric- and third-person-view multi-modal data, annotated with high-quality 3D ground-truth motions for *both* interacting people and 3D scene reconstructions; **(2)** extensively evaluate state-of-the-art 3DHPS methods on our test set, showing their shortcomings in this egocentric setup and providing insights for future methods in this direction; **(3)** show the usefulness of our training set: a simple fine-tuning on it significantly improves existing methods' performance and robustness on both our test set *and a different egocentric dataset*; **(4)** provide the first benchmark for 3DHPS estimation of the interactee in the egocentric view during social interactions.

2 Related Work

Datasets for 3D human pose, motion and interactions. A large number of datasets focus on 3D human pose and motion from *third-person-views* [23,30,34,37,38,52,63–65,72,86,97,103]. For example, Human3.6M [34] and AMASS [62] use optical marker-based motion capture to collect large amounts of high-quality 3D motion sequences; they are limited to constrained studio setups and images – when available – are polluted by markers. PROX [30] performs marker-less capture of people moving in 3D scenes from monocular RGB-D, without human-human interactions. The quality of the reconstructed motion is further improved by LEMO [102]. The Panoptic Studio datasets [37–39,92] capture interactions between people using a multi-view camera system, annotated with body and hand 3D joints plus facial landmarks. CHI3D [23] focuses on close human-human contacts, using a motion capture system to extract ground-truth 3D skeletons. 3DPW [63] reconstructs the 3D shape and motion of people by fitting SMPL [59] to IMU data and RGB images captured with a hand-held camera, without 3D environment reconstruction. None of these datasets provides egocentric data.

Among datasets for *egocentric* vision, a lot of attention has been put on hand-object interactions and action recognition, often without 3D ground-truth [10,16,17,22,42–44,51,56,67,70,75,77,80,96,104]. Mo2Cap2 [93] and xR-EgoPose [84,85] provide image-3D skeleton pairs for egocentric body pose prediction of the camera wearer, without the interactee involved. HPS [29] reconstructs the body pose and shape of the camera wearer moving in large 3D scenes; only a few frames include interactions with an interactee. You2Me [68] provides 3D skeletons for both interacting people paired with images captured with a chest-mounted camera plus external cameras; there are no body shape or 3D scene annotations.

EgoMoCap [58] analyzes the interactee body shape and pose in outdoor social scenarios capturing only the egocentric RGB stream.

Table 1 compares EgoBody with the most related human motion datasets. EgoBody is the first motion capture dataset that collects calibrated egocentric- and third-person-view images, with various interaction scenarios, multi-modal data and rich 3D ground-truth. Additionally, EgoBody provides the camera wearer's eye gaze to facilitate potential social interaction studies which jointly analyze human attention and motion.

3D human pose estimation. The problem of estimating 3D human pose from *third-person-view* RGB(D) images has been extensively studied in the literature – either from single frames [5, 9, 12, 15, 20, 26, 28, 30, 40, 46–50, 55, 57, 66, 71, 73, 81, 83, 87, 89, 91, 94, 101, 105], monocular videos [14, 41, 45, 60, 82, 99, 100, 102] or multi-view camera sequences [19,24,33,39,78,90]. SPIN [48] estimates SMPL [59] parameters from single RGB images by combining deep learning with optimization frameworks. METRO [57] reconstructs human meshes without relying on parametric body models. Most methods require "full-body" images and therefore lack robustness when parts of the body are occluded or truncated, as it is the case with the interactee in egocentric videos. EFT [36] injects crop augmentations at training time to better reconstruct highly truncated people. PARE [46] explicitly learns to predict body-part-guided attention masks. However, these methods exhibit a significant performance drop when applied to egocentric data. Our dataset helps fill this performance gap, as we show in Sec. 5.

The problem of *egocentric* pose estimation is receiving growing attention. Most methods estimate the *camera wearer*'s 3D skeleton, based on images, IMU data, scene cues or body-object interactions [29,35,61,79,84,85,98]. You2Me [68] estimates the camera wearer's pose given the interactee's pose as an additional cue. Liu et al. [58] estimate 3D human pose and shape of the interactee given egocentric videos in outdoor scenes, with limited interaction diversity.

Egocentric social interaction learning. Egocentric videos provide a unique way to study social interactions. Most methods focus on social interaction recognition [6,7,18,21,54,67,77,95,96]. Lee et al. [53] produce a storyboard summary of the camera wearer's day given egocentric videos. Northcutt et al. [69] collect an egocentric communication dataset focusing on conversations. Recently Ego4D [27] dataset collects massive egocentric videos for various tasks including hand-object and social interaction understanding, making significant advances in stimulating future research in the egocentric domain. EgoBody is unique in that we are the first egocentric dataset that provides rich 3D annotations including accurate 3D human pose and shape for all interacting subjects.

3 Building the EgoBody Dataset

EgoBody collects sequences capturing subjects performing diverse social interactions in various indoor scenes. For each sequence, two subjects are involved in one or more interaction scenarios (Sec. 3.1). Their performance is captured from both egocentric- and third-person-views. One subject (the camera wearer) wears

Table 2: EgoBody interaction scenarios.

Category	**Interaction Scenarios**
Cooperation	Guess by Action game, catching and tossing, searching for items, etc.
Social exchange	Teaching to dance/workout, giving a presentation, etc.
Conflict	Arguing about a specific topic
Conformity	One subject instructs the other to perform a task
Others	Haggling, negotiation, promotion, self-introduction, casual chat, etc.
Action Types	Sitting, standing, walking, dancing, exercising, bending, lying, grasping, squatting, drinking, passing objects, catching, throwing, etc.

a HoloLens2 headset [3], capturing multi-modal egocentric data (RGB, depth, head, hand and eye gaze tracking streams). Their interaction partner, *i.e.* interactee, does not wear any device. The camera wearer's HoloLens2 is calibrated and synchronized with three to five Azure Kinect cameras [1] which capture the interaction from different viewpoints (Sec. 3.2). Based on this multi-view data, we acquire rich ground-truth annotations for all frames, including 3D full-body pose and shape for both interacting subjects and the reconstructed 3D scene (Sec. 3.3). Statistics for EgoBody are reported in Sec. 4.

3.1 Interaction Scenarios

To guide the subjects and obtain rich, diverse body motions, we define multiple interaction scenarios within five major interaction categories in sociology studies [8]: *cooperation*, *social exchange*, *conflict*, *conformity* and *others*, spanning diverse action types (Tab. 2) and body poses (Fig. 5). For each sequence, we pre-define one or more interaction scenarios and ask the two participants to interact accordingly. We allow the subjects to improvise within each interaction scenario to ensure intra-class variation. The motion diversity is further increased with various human-scene interactions by capturing in 3D scenes.

3.2 Data Acquisition Setup

As mentioned above, EgoBody collects egocentric- and third-person-view multi-modal data, plus 3D scene reconstructions. Fig. 2 illustrates our system setup.

Egocentric-view capture. We use a Microsoft Hololens2 [3] headset to record egocentric data. Using the Research Mode API [88], we capture RGB videos (1920×1080) at 30 FPS, long-throw depth frames (512×512) at 1-5 FPS, as well as eye gaze, hand and head tracking at 60 FPS. Note that we do not record depth at a higher framerate (AHAT) due to the "depth aliasing" described in [88]. We observe that captures exhibit typical challenges for limited power-devices, like frame drops and blurry images.

Third-person multi-view capture. We use three to five Azure Kinect cameras [1] (denoted by *Cam1*∼*Cam5*) to capture multi-view, synchronized RGB-

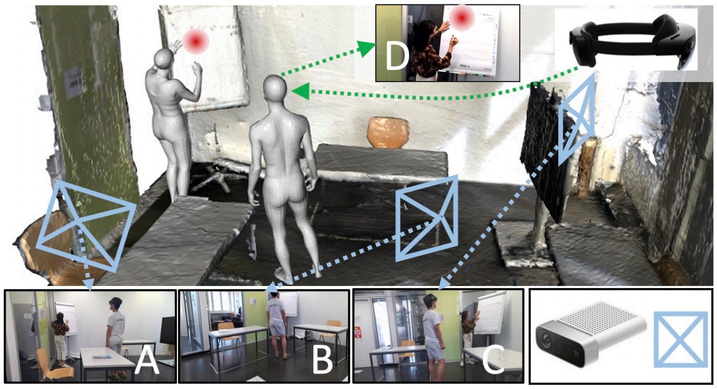

Fig. 2: Capture setup. Multiple Azure Kinects capture the interactions from different views (A, B, C), and a synchronized HoloLens2 worn by one subject captures the egocentric view image (D), as well as the eye gaze (red circle) of the camera wearer.

D videos of interacting subjects. Having multi-view data helps our motion reconstruction pipeline for ground-truth acquisition (Sec. 3.3). The cameras are fixed during recording. They capture synchronized RGB frames (1920×1080) and depth frames (640×576) at 30 FPS.

3D scene representation. We pre-scan the environment using an iPhone12 Pro Max running the 3D Scanner app [2]. Scene reconstructions are stored as 3D triangulated meshes, each with $10^5 \sim 10^6$ vertices. We choose this procedure for its efficiency and reconstruction quality.

Calibration and Synchronization. For each Kinect, we extract its camera parameters via the Azure Kinect DK [1]. For the HoloLens2, we get its camera parameters as exposed by Research Mode [88]. We synchronize the Kinects via hardware, using audio cables. Since it is not possible to synchronize HoloLens2 and Kinect in a similar way, we use a flashlight visible to all devices as signal for the first frame. Kinect-Kinect and Kinect-HoloLens2 cameras are spatially calibrated using a checkerboard and refined by rigid alignment steps (ICP [11]).

The Kinect-HoloLens2 calibration is further optimized based on body keypoints (Sec. 3.3). We use *Cam1* to define our world coordinate frame origin. Once we calibrate the HoloLens2 coordinate frame with *Cam1*'s world origin, we can track the headset position, and therefore its cameras, by relying on its built-in head tracker [88]. We also register the 3D scene into the coordinate frame of *Cam1*, and reconstruct the human body in this space (see details in Supp. Mat.).

3.3 Ground-truth Acquisition

Given the RGB-D frames captured with the multi-Kinect rig and the egocentric frames, our motion reconstruction pipeline estimates, for each frame and each subject, the corresponding SMPL-X body parameters [73], including the

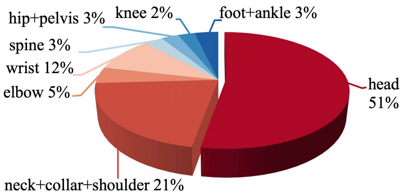

Fig. 3: Which body part attracts more attention? For each joint group, % of the occurrences it is the closest to the 2D gaze point in the image.

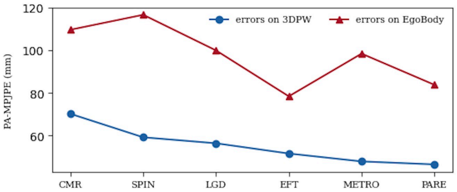

Fig. 4: Accuracy of SoTAs on 3DPW and EgoBody with the advance of the 3DHPS field.

global translation $\boldsymbol{\gamma} \in \mathbb{R}^3$, body shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{96}$ (body and hand) and facial expression $\boldsymbol{\phi} \in \mathbb{R}^{10}$. To address the challenges posed by not perfectly accurate factory calibration of HoloLens2 depth, which further leads to inaccurate Kinect-HoloLens2 calibration, we propose a keypoint-based refinement scheme to better leverage observations from the HoloLens2. We introduce the first solution to reconstruct accurate 3D human pose, shape and motions with multi-view Kinect cameras and an embodied HMD. Thanks to the refined Kinect-HoloLens2 calibration, this provides accurate per-frame pose, natural human motion dynamics and realistic human-scene interactions for both egocentric- and third-person-view frames. Note that we estimate the body in the coordinate frame of *Cam1*.

Data preprocessing. We use OpenPose [13] to detect 2D body joints in all (Kinect and HoloLens2) RGB frames. OpenPose identifies people in the same image by assigning a body index to each detected person. In general, this works well, but gives false positives which we process afterwards. To extract human body point clouds from Kinect depth frames, we use Mask-RCNN [31] and DeepLabv3 [31]. We manually inspect the data to remove spurious detections (*e.g.* irrelevant people in the background, and scene objects misdetected as people). We also ensure consistent subject identification across frames and views, and manually fix inaccurate 2D joint detections, mostly due to body-body and body-scene occlusions. See Supp. Mat. for more details.

Per-frame fitting. As in [102], given Kinect depth and 2D joints, we first optimize the SMPL-X parameters for each subject/frame separately, minimizing an objective function similar to that defined in [30]:

$$E(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}) = E_J + \lambda_D E_D + E_{prior} + \lambda_{contact} E_{contact} + \lambda_{coll} E_{coll}, \tag{1}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}$ are optimized SMPL-X parameters.

Given the preprocessed OpenPose 2D joints J_{OP}^v from n views ($v \in \{1, ..., n\}$), the multi-view joint error term E_J minimizes the sum of 2D distances between J_{OP}^v and the 2D projection of SMPL-X joints onto camera view v for all views:

$$E_J(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{view\ v} E_{J_v}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}, J_{OP}^v, K_v, T_v), \tag{2}$$

where K_v denotes the intrinsics parameters of camera v, and T_v denotes the extrinsics between *Cam v* and *Cam 1*. The depth term E_D penalizes discrepancies between the estimated body surface and body depth point clouds for all views; E_{prior} represents body pose, shape and expression priors; $E_{contact}$ encourages scene-body contacts; and E_{coll} penalizes scene-body collisions. The λ_is weight the contribution of each term. We refer the reader to [30, 102] for more details.

Kinect-HoloLens2 calibration refinement. The Kinect-Hololens2 calibration is represented by the extrinsics T between Kinect *Cam1*, and the HoloLens2 coordinate system's origin. For each capture session, this origin is fixed in the world [88]; as the HMD moves, its head tracker provides the transformation between this origin and each egocentric frame t, denoted by T_t^{ego}. To address the inaccurate initial Kinect-HoloLens2 calibration T_{init} caused by imperfect HoloLens2 depth factory calibration, we propose a keypoint-based scheme to refine it. For each frame t, we project the 3D SMPL-X joints $J_{3D,t}$ (obtained from per-frame fitting, in *Cam1*'s coordinate) onto the egocentric image. We minimize the 2D error between the projected 2D joints and the OpenPose joint detections $J_{OP,t}^{ego}$ of the egocentric frame t, and optimize the transformation T:

$$E_T(T) = \sum_t ||K^{ego} T_t^{ego} T J_{3D,t} - J_{OP,t}^{ego}||_2^2 + \lambda ||T - T_{init}||_2^2, \tag{3}$$

where K^{ego} denotes the HoloLens2 RGB camera intrinsic parameters, and λ weights the regularizer.

Temporally consistent fitting. Per-frame fitting gives us a set of reasonable, initial pose estimates, which however are jittery and inconsistent over time. We therefore run a second optimization stage based on LEMO priors [102] to obtain smooth, realistic human motions. Furthermore, to improve consistency between egocentric- and third-person-view estimates, we consider also egocentric data given the refined Kinect-HoloLens2 calibration. We take OpenPose 2D joint estimations from HoloLens2 RGB frames and use them as further constraints. Still, we optimize for each subject separately. The resulting objective function minimized in the temporal fitting stage is:

$$E(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}) = E_J + E_{J_{ego}} + E_{prior} + \lambda_{smooth} E_{smooth} + \lambda_{fric} E_{fric}, \tag{4}$$

where E_{fric} is the contact friction term defined in [102] to prevent body sliding, E_{smooth} and E_{prior} denote temporal and static priors as in [102]. $E_{J_{ego}}$ is the 2D projection term which minimizes the error between OpenPose detections on egocentric view frames and the 2D projections of SMPL-X joints onto the egocentric view; $E_{J_{ego}}$ is only enabled for the interactee when they are visible in the egocentric frames. The λ_is weight balance the contribution of each term.

4 EgoBody Dataset

EgoBody collects 125 sequences from 36 subjects (18 male and 18 female) performing diverse social interactions in 15 indoor scenes. In total, there are 219,731

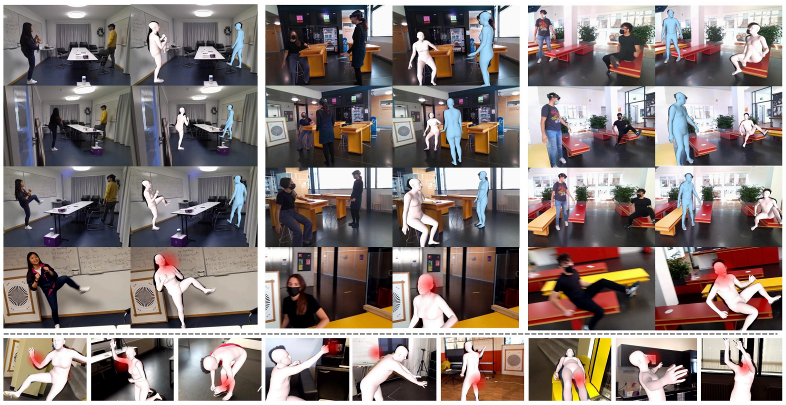

Fig. 5: Reconstructed ground-truth bodies overlaid on third-person-view images from 3 Kinects (row 1-3), and the corresponding egocentric view image (row 4). Left/middle/right shows three different frames. Row 5 shows more examples from the egocentric view. Blue denotes the camera wearer, and pink denotes the interactee. Eye gaze of the camera wearer are in red circles.

synchronized frames captured from Azure Kinects, from multiple third-person-views. We refer to this as the "Multi-view (MV)Set". For each MV frame, we provide 3D human full-body pose and shape annotations (as SMPL-X parameters) for both interacting subjects together with the 3D scene mesh. Furthermore, we have 199,111 egocentric RGB frames (the "EgoSet"), captured from HoloLens2, calibrated and synchronized with Kinect frames. Given the camera wearer's head motion, the interactee is not visible in every egocentric frame; in total, we have 175,611 frames with the interactee visible in the egocentric view ("EgoSet-interactee"). Fig. 5 shows example images. For EgoSet, we also collect the head, hand and eye tracking data, plus the depth frames from the HoloLens2. We also provide SMPL [59] body annotations via the official transfer tool [4]. Below we provide dataset statistics; for more detailed analysis and ground truth annotation quality please refer to the Supp. Mat.

Training/validation/test splits. We split data into training, validation and test sets such that they have no overlapping subjects. The EgoBody training set contains 116,630 MVSet frames, 105,388 EgoSet frames and 90,124 EgoSet-interactee frames. The EgoBody validation set contains 29,140 MVSet frames, 25,416 EgoSet frames and 23,332 EgoSet-interactee frames. The test set contains 73,961 MV frames, 68,307 EgoSet frames and 62,155 Ego-interactee frames.

Joint visibility. The camera wearer's motion, the headset's field of view and the close distance between the interacting subjects cause the interactee to be often truncated in the egocentric view. To quantify the occurrences of truncations, we project the fitted 3D body joints onto the HoloLens2 images, and deem a projected 2D joint as "visible" if it lies inside the image. As shown in Fig. 6 (2nd row, right), the lower body parts are more frequently truncated in the images.

Please refer to Sec. 5 for the impact of joint visibility on 3DHPS estimation performance.

Eye gaze and attention. We can combine the HoloLens2 eye gaze tracking with our 3D reconstruction of the scene/people to estimate the 3D location the user looks at, and project it on the egocentric images (interpreted as where the user's "attention" is focused), thereby obtaining valuable data to understand interactions. We observe that the camera wearer's attention is highly focused on the interactee during interactions. and tends to be closer to the upper body joints (Fig. 3), which in turn results in lower visibility for the lower body parts.

5 Experiments

We leverage EgoBody to introduce the first benchmark for 3D human pose and shape (3DHPS) estimation from egocentric images. Given a single RGB image of a target subject, the goal of a 3DHPS method is to estimate a human body mesh and a set of camera parameters, which best explain the image data. State-of-the-art (SoTA) 3DHPS methods are mostly trained and evaluated on third-person-view data, and their performance is starting to saturate on common third-person-view datasets [34,63] (see Fig. 4); yet, their capabilities to generalize to real-world scenarios (*e.g.* cropped or blurry images) are still limited [72]. With EgoBody, we can test their capabilities on egocentric images.

We define a benchmark for 3DHPS methods on our EgoSet-interactee test set. Within the social interaction scenarios, the input will be an egocentric view image of the interactee. We evaluate SoTA methods and show that their performance significantly drops on our data. We expose limitations of existing methods by in-depth analysis (Sec. 5.2), given that the egocentric view brings considerable challenges that are rarely present in existing third-person-view datasets.

We also provide valuable insights to boost their performance for egocentric scenarios. In particular, we show that our EgoSet-interactee training set can help address the challenges brought by egocentric view data: using it, we fine-tune three recent methods, SPIN [48], METRO [57] and EFT [36], achieving significantly improved accuracy and robustness on both our test set (Sec. 5.3) and over a cross-dataset evaluation on the You2Me [68] dataset (Sec. 5.4).

5.1 Benchmark Evaluation Metrics

We employ two common metrics: **Mean Per-Joint Position Error (MPJPE)** and **Vertex-to-Vertex (V2V)** errors. We use two types of alignments before computing the accuracy for each metric: (1) translation-only alignment (aligns the bodies at the pelvis joint [72]) and (2) Procrustes Alignment [25] ("PA", solves for scale, translation and rotation). Results are by default reported with translation-only alignment unless specified with the "PA-" prefix. **MPJPE** is the mean Euclidean distance between predicted and ground-truth 3D joints, evaluated on 24 SMPL body joints. **V2V** error is the mean Euclidean distance over all body vertices, computed between two meshes.

Table 3: Evaluation of SoTA 3DHPS estimation methods on our test set. All metrics are in *mm*. "PA-" stands for Procrustes alignment. "SPIN-ft", "METRO-ft" and "EFT-ft" denote results of fine-tuning SPIN, METRO and EFT on our training set.

Method	MPJPE ↓	PA-MPJPE ↓	V2V ↓	PA-V2V ↓
CMR [49]	200.7	109.6	218.7	136.8
SPIN [48]	182.8	116.6	187.3	123.7
LGD [81]	158.0	99.9	168.3	106.0
METRO [57]	153.1	98.4	164.6	106.4
PARE [46]	123.0	83.8	131.5	89.7
EFT [36]	123.9	78.4	134.9	86.0
SPIN-ft (Ours)	106.5	67.1	120.9	78.3
METRO-ft (Ours)	**98.5**	70.0	**110.5**	76.8
EFT-ft (Ours)	102.1	**64.8**	116.1	**74.8**

5.2 Baseline Evaluation

Tab. 3 summarizes the evaluation of SoTA 3DHPS methods from different categories: (1) fitting-based method [81]; and regression-based methods that (2) predict parameters of a parametric body model [36,46,48,49] or (3) predict non-parametric body meshes [57]. For each baseline method, we use the best performing model provided by the authors (trained with the optimal training data).

In Fig. 4 we plot the PA-MPJPE error of these methods on our dataset and on an existing major third-person-view benchmark[1], On average, the methods yield a 77% higher 3D joint error on EgoBody than on 3DPW. More importantly, while the accuracy curve drives towards saturation on 3DPW, different SoTA methods still show largely varying performance on our dataset. This suggests that current datasets are not sufficient to train models that can handle egocentric view images well. Below we discuss two key challenging factors that impact performance.

Motion blur. Motion blur is common in the egocentric view images due to the motion of the camera wearer. To study how motion blur influences 3DHPS estimation accuracy, we plot in Fig. 6 (1st row, left) the MPJPE of all methods vs. the image sharpness score. The sharpness score is defined as the variance of the Laplacian of an image [74], upper-thresholded at 60; higher scores mean sharper images. We observe that, surprisingly, most methods are insensitive to blurriness, except for heavily blurred cases (score <10). However, our fine-tuned models (SPIN-ft / METRO-ft / EFT-ft) are more robust against motion blur: among all methods, they achieve the lowest standard deviation over the seven image sharpness levels; see the number next to each method in the legend of Fig. 6 (1st row, left).

Joint visibility. While most 3DHPS methods assume that the target body is (almost) fully visible in the image as in existing third-person-view datasets such as 3DPW [63], this is seldom the case in egocentric view images. To assess the importance of this issue, we analyze the performance of each baseline with

[1] The results on 3DPW are taken from the respective original papers.

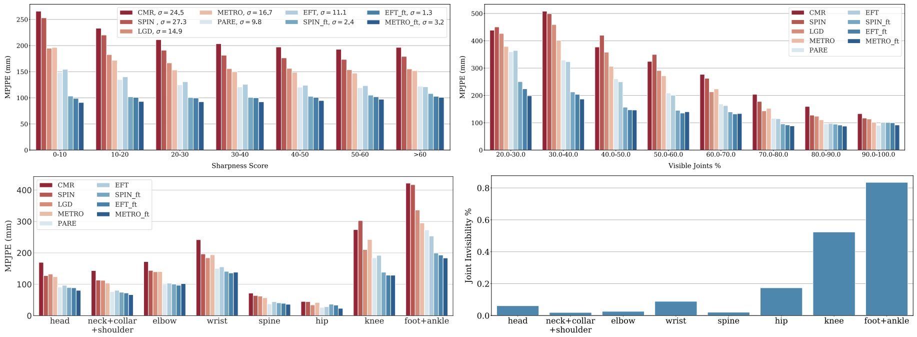

Fig. 6: First Row: Impact of motion blur (left) / joint visibility (right) on SoTA method accuracies. Second Row: 3D joint error analysis by body parts (left) / ratio of each joint group being *in*visible (truncated) from the images in our test set (right).

respect to the portion of visible body joints ("visibility", see Sec. 4) in the images from our test set. The result is summarized in Fig. 6 (row 1, right). Note that our definition of a joint's visibility is related to, but differs from, the concept of *occlusion*: both measure how much pixel information is missing for a body part, but visibility focuses on how much of the body is *truncated* from the image. A joint that is occluded by an object can still be considered visible by our definition.

Overall, all methods yield a lower error when there is less body truncation. Two recent methods, PARE [46] and EFT [36], achieve the best results. PARE is designed to be robust against occlusions by explicitly employing a body part attention mechanism, whereas EFT handles body truncation "implicitly" by aggressively cropping images as training data augmentation.

We further plot the MPJPE and the *in*visibility ratio of each joint group in Fig. 6 (2nd row). Overall the two are in accordance: the lesser a joint is visible, the higher the error it exhibits. An exception is on the wrist joints: despite good visibility, their error remains relatively high. As observed also in [46], high errors on the extremities are a common problem with existing 3DHPS models, possibly because most current models only use a single, global feature from the input image for regression. This points to potential future work that deploys local image features, which has been shown effective in recent 3DHPS models [28,46].

5.3 Baseline Improvement

To evaluate the effectiveness of the EgoBody training set, we use it to fine-tune three of the baseline methods: two model-based methods, SPIN [48] and EFT [36], as they both use the same architecture (HMR [40] network) that is the backbone for many other recent models [40,45,76]; a model-free method METRO [57] which directly predicts the body mesh. The pre-trained EFT differs from SPIN majorly in that it is trained with extended 3D pseudo ground-truth data (from the EFT-dataset) and uses aggressive image cropping as data augmentation. We use the same hyperparameters provided by the authors and select the fine-tuned model with the best validation score.

As shown in Tab. 3, after fine-tuning, the error is largely reduced for all three methods on all metrics: SPIN-ft/METRO-ft/EFT-ft has 42%/36%/18% lower MPJPE, and 35%/33%/14% V2V than their corresponding original models.

The improvement can also be seen for all blurriness/visibility categories in Figs. 6. For the motion blur specifically, the fine-tuned models not only achieve a lower error at every image sharpness level, but also show increased robustness. This is shown by the standard deviations of each method across the sharpness levels, dropping from 27.3 to 2.4 for SPIN, from 16.7 to 3.2 for METRO, and from 11.1 to 1.3 for EFT, respectively, after fine-tuning. The results show that our training set can serve as an effective source to adapt existing 3DHPS methods to the egocentric setting.

5.4 Cross-dataset Evaluation on You2Me

Is the effect of our training set only specific to our capture scenario, or does it generalize to other egocentric pose estimation datasets? To verify this, we evaluate SPIN, EFT and METRO against their fine-tuned counterparts on the You2Me [68] dataset. Here we report the PA-MPJPE for pose errors (in *mm*): SPIN (152.8) vs. SPIN-ft (87.9); EFT (95.8) vs. EFT-ft (85.6), and METRO (117.7) vs. METRO-ft (88.2). Again, fine-tuning on our training set improves all models' performance; see Supp. Mat. for more details. These results suggest that our data empowers existing models with the ability to address challenges faced in the generic egocentric view setup.

6 Conclusion

We presented EgoBody, a dataset capturing human pose, shape and motions of interacting people in diverse environments. EgoBody collects multi-modal egocentric- and third-person-view data, accompanied by ground-truth 3D human pose and shape for all interacting subjects. With this dataset, we introduced a benchmark on egocentric-view 3D human body pose and shape (3DHPS) estimation, systematically evaluated and analyzed limitations of state-of-the-art methods on the egocentric setting, and demonstrated a significant, generalizable performance gain in them with the help of our annotations. This paper has shown EgoBody's unique value for the 3DHPS estimation task, and we see its great potential in moving the fields towards a better understanding of egocentric human motions, behaviors, and social interactions. In the future, adding more participants and even richer data modalities (*e.g.* audio recordings and motion annotations by natural language descriptions) could further enrich the dataset.

Acknowledgements. This work was supported by the SNF grant 200021 204840 and Microsoft Mixed Reality & AI Zurich Lab PhD scholarship. Qianli Ma is partially funded by the Max Planck ETH Center for Learning Systems.

References

1. Azure Kinect. https://docs.microsoft.com/en-us/azure/kinect-dk/ 3, 6, 7
2. Laan Labs 3D Scanner app. https://apps.apple.com/us/app/3d-scanner-app/id1419913995 7
3. Microsoft Hololens2. https://www.microsoft.com/en-us/hololens 3, 6
4. SMPL model transfer. https://github.com/vchoutas/smplx/tree/master/transfer_mode 10
5. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. IEEE transactions on pattern analysis and machine intelligence **28**(1), 44–58 (2005) 5
6. Aghaei, M., Dimiccoli, M., Ferrer, C.C., Radeva, P.: Towards social pattern characterization in egocentric photo-streams. Computer Vision and Image Understanding **171**, 104–117 (2018) 5
7. Aghaei, M., Dimiccoli, M., Radeva, P.: With whom do i interact? detecting social interactions in egocentric photo-streams. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 2959–2964. IEEE (2016) 5
8. A.Nisbet, R.: The Social Bond: An Introduction to the Study of Society (1970) 2, 6
9. Bălan, A.O., Black, M.J.: The naked truth: Estimating body shape under clothing. In: European Conference on Computer Vision. pp. 15–29. Springer (2008) 5
10. Bambach, S., Lee, S., Crandall, D.J., Yu, C.: Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: Proceedings of the IEEE international conference on computer vision. pp. 1949–1957 (2015) 4
11. Besl, P., McKay, N.D.: A method for registration of 3-d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2), 239–256 (1992) 7, 1
12. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision. pp. 561–578 (2016) 5
13. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 8, 1
14. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1964–1973 (2021) 2, 5
15. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: European Conference on Computer Vision (ECCV) (2020) 5
16. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 720–736 (2018) 4
17. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision **130**(1), 33–55 (2022) 4
18. Dhand, A., Dalton, A.E., Luke, D.A., Gage, B.F., Lee, J.M.: Accuracy of wearable cameras to track social interactions in stroke survivors. Journal of Stroke and Cerebrovascular Diseases **25**(12), 2907–2910 (2016) 5

19. Dong, J., Shuai, Q., Zhang, Y., Liu, X., Zhou, X., Bao, H.: Motion capture from internet videos. In: European Conference on Computer Vision. pp. 210–227. Springer (2020) 5
20. Fang, Q., Shuai, Q., Dong, J., Bao, H., Zhou, X.: Reconstructing 3d human pose by watching humans in the mirror. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12814–12823 (2021) 2, 5
21. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1226–1233. IEEE (2012) 2, 5
22. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: 2011 international conference on computer vision. pp. 407–414. IEEE (2011) 4
23. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7214–7223 (2020) 2, 4
24. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. International journal of computer vision **87**(1-2), 75 (2010) 5
25. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975) 11
26. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3d structure with a statistical image-based shape model. In: ICCV. vol. 3, p. 641 (2003) 5
27. Grauman, K., et al.: Ego4D: Around the world in 3000 hours of egocentric video. arXiv preprint arXiv:2110.07058 (2021) 2, 5
28. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019) 5, 13
29. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: 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. pp. 4318–4329 (2021) 2, 3, 4, 5
30. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2282–2292 (2019) 2, 3, 4, 5, 8, 9
31. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 8, 1
32. Hernandez, A., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7134–7143 (2019) 3
33. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 international conference on 3D vision (3DV). pp. 421–430. IEEE (2017) 5
34. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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) 2, 3, 4, 11
35. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3509. IEEE (2017) 5

36. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation (2021) 4, 5, 11, 12, 13, 7
37. Joo, H., Simon, T., Cikara, M., Sheikh, Y.: Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10873–10883 (2019) 2, 4
38. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., et al.: Panoptic studio: A massively multiview system for social interaction capture. IEEE transactions on pattern analysis and machine intelligence **41**(1), 190–204 (2017) 2, 3, 4
39. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8320–8329 (2018) 4, 5
40. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (2018) 2, 5, 13
41. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5614–5623 (2019) 2, 5
42. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 4
43. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019) 4
44. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR 2011. pp. 3241–3248. IEEE (2011) 4
45. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5253–5263 (2020) 2, 5, 13
46. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: Proceedings International Conference on Computer Vision (ICCV). pp. 11127–11137. IEEE (Oct 2021) 5, 12, 13, 7
47. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: Proc. International Conference on Computer Vision (ICCV). pp. 11035–11045 (Oct 2021) 5
48. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2252–2261 (2019) 2, 4, 5, 11, 12, 13, 7
49. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019) 5, 12, 7
50. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021) 5
51. Kwon, T., Tekin, B., Stuhmer, J., Bogo, F., Pollefeys, M.: H2O: Two hands manipulating objects for first person interaction recognition. In: International Conference on Computer Vision (ICCV) (2021) 4
52. Lab, C.G.: CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/ (2000) 4

53. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1346–1353. IEEE (2012) 5
54. Li, H., Cai, Y., Zheng, W.S.: Deep dual relation modeling for egocentric interaction recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7932–7941 (2019) 5
55. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3383–3393 (2021) 5
56. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 619–635 (2018) 2, 4
57. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021) 4, 5, 11, 12, 13, 7
58. Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., Tang, S.: 4D human body capture from egocentric video via 3D scene grounding. 2021 international conference on 3D vision (3DV) (2021) 5
59. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 4, 5, 10
60. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: Proceedings of the Asian Conference on Computer Vision (2020) 2, 5
61. Luo, Z., Hachiuma, R., Yuan, Y., Iwase, S., Kitani, K.M.: Kinematics-guided reinforcement learning for object-aware 3d ego-pose estimation. arXiv preprint arXiv:2011.04837 (2020) 5
62. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5442–5451 (2019) 4, 3
63. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018) 3, 4, 11, 12
64. von Marcard, T., Pons-Moll, G., Rosenhahn, B.: Human pose estimation from video and imus. Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1533–1547 (Jan 2016) 3, 4
65. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017) 2, 4
66. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: European Conference on Computer Vision (ECCV) (2020) 5
67. Narayan, S., Kankanhalli, M.S., Ramakrishnan, K.R.: Action and interaction recognition in first-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 512–518 (2014) 2, 4, 5
68. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9890–9900 (2020) 2, 3, 4, 5, 11, 14, 6, 9

69. Northcutt, C., Zha, S., Lovegrove, S., Newcombe, R.: Egocom: A multi-person multi-modal egocentric communications dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 5
70. Ogaki, K., Kitani, K.M., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–7. IEEE (2012) 4
71. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018) 5
72. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021) 4, 11, 5
73. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019) 2, 3, 5, 7, 1
74. Pech-Pacheco, J.L., Cristóbal, G., Chamorro-Martinez, J., Fernández-Valdivia, J.: Diatom autofocusing in brightfield microscopy: a comparative study. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. vol. 3, pp. 314–317. IEEE (2000) 12, 4
75. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2847–2854. IEEE (2012) 2, 4
76. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: IEEE International Conference on Computer Vision Workshops (2021) 13
77. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2730–2737 (2013) 4, 5
78. Saini, N., Price, E., Tallamraju, R., Enficiaud, R., Ludwig, R., Martinovic, I., Ahmad, A., Black, M.J.: Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 823–832 (2019) 5
79. Shiratori, T., Park, H.S., Sigal, L., Sheikh, Y., Hodgins, J.K.: Motion capture from body-mounted cameras. In: ACM SIGGRAPH 2011 papers, pp. 1–10 (2011) 5
80. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7396–7404 (2018) 2, 4
81. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent (2020) 5, 12, 7
82. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5349–5358 (2019) 2, 5

83. Tan, J.K.V., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3d human body shape and pose prediction (2017) 5
84. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. arXiv preprint arXiv:2011.01519 (2020) 2, 4, 5
85. Tome, D., Peluse, P., Agapito, L., Badino, H.: xr-egopose: Egocentric 3d human pose from an hmd camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7728–7738 (2019) 2, 4, 5
86. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: 2017 British Machine Vision Conference (BMVC) (2017) 2, 3, 4
87. Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: Advances in Neural Information Processing Systems. pp. 5236–5246 (2017) 5
88. Ungureanu, D., Bogo, F., Galliani, S., Sama, P., Duan, X., Meekhof, C., Stuhmer, J., Cashman, T.J., Tekin, B., Schonberger, J.L., Tekin, B., Olszta, P., Pollefeys, M.: HoloLens 2 Research Mode as a Tool for Computer Vision Research. arXiv:2008.11239 (2020) 6, 7, 9
89. Wandt, B., Rudolph, M., Zell, P., Rhodin, H., Rosenhahn, B.: Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2021) 5
90. Wang, Y., Liu, Y., Tong, X., Dai, Q., Tan, P.: Outdoor markerless motion capture with sparse handheld video cameras. IEEE transactions on visualization and computer graphics **24**(5), 1856–1866 (2017) 5
91. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 334–343 (2021) 2, 5
92. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019) 4
93. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo2Cap2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera. IEEE transactions on visualization and computer graphics **25**(5), 2093–2101 (2019) 2, 3, 4
94. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7760–7770 (2019) 2, 5
95. Yang, J.A., Lee, C.H., Yang, S.W., Somayazulu, V.S., Chen, Y.K., Chien, S.Y.: Wearable social camera: Egocentric video summarization for social interaction. In: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2016) 5
96. Yonetani, R., Kitani, K.M., Sato, Y.: Recognizing micro-actions and reactions from paired egocentric videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2629–2638 (2016) 4, 5
97. Yu, Z., Yoon, J.S., Lee, I.K., Venkatesh, P., Park, J., Yu, J., Park, H.S.: Humbi: A large multiview dataset of human body expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2990–3000 (2020) 3, 4

98. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10082–10092 (2019) 2, 5
99. Yuan, Y., Wei, S.E., Simon, T., Kitani, K., Saragih, J.: Simpoe: Simulated character control for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7159–7169 (2021) 2, 5
100. Zanfir, A., Bazavan, E.G., Xu, H., Freeman, B., Sukthankar, R., Sminchisescu, C.: Weakly supervised 3d human pose and shape reconstruction with normalizing flows. arXiv preprint arXiv:2003.10350 (2020) 2, 5
101. Zhang, J., Yu, D., Liew, J.H., Nie, X., Feng, J.: Body meshes as points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 546–556 (2021) 2, 5
102. Zhang, S., Zhang, Y., Bogo, F., Marc, P., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: International Conference on Computer Vision (ICCV) (Oct 2021) 2, 3, 4, 5, 8, 9
103. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 4d association graph for realtime multi-person motion capture using multiple video cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1324–1333 (2020) 2, 4
104. Zhang, Z., Crandall, D., Proulx, M., Talathi, S., Sharma, A.: Can gaze inform egocentric action recognition? In: 2022 Symposium on Eye Tracking Research and Applications. pp. 1–7 (2022) 4
105. Zhou, Y., Habermann, M., Habibie, I., Tewari, A., Theobalt, C., Xu, F.: Monocular real-time full body capture with inter-part correlations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4811–4822 (2021) 5