Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors

Vladimir Guzov * [1,2] Aymen Mir * [1,2] Torsten Sattler [3] Gerard Pons-Moll[1,2]

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

{vguzov, amir, gpons}@mpi-inf.mpg.de torsten.sattler@cvut.cz

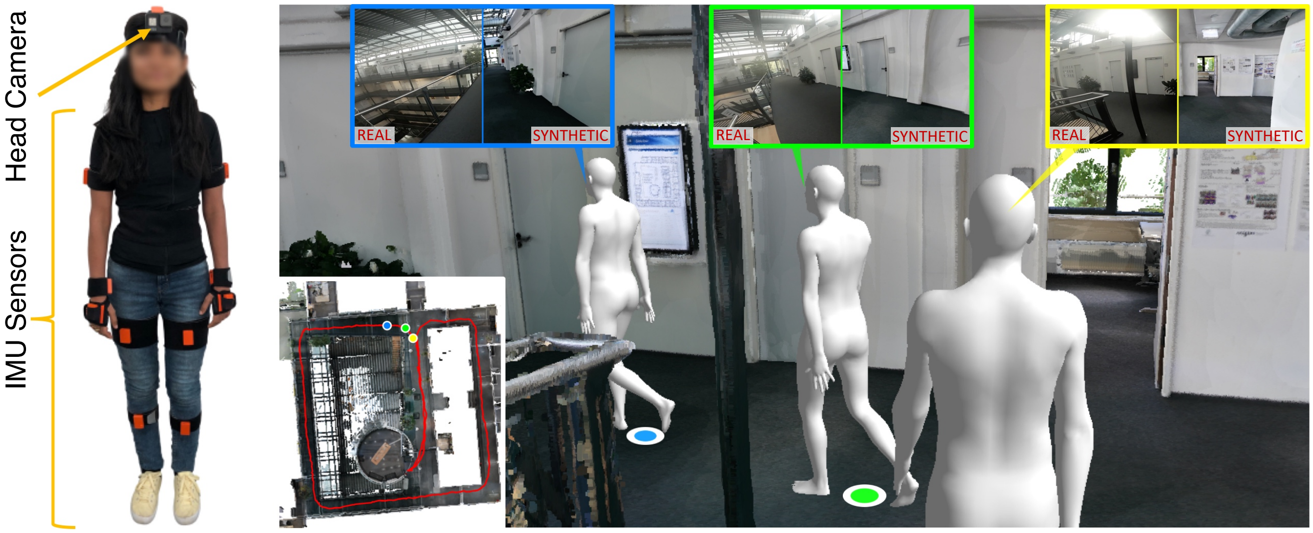

Figure 1. HPS jointly estimates the full 3D human pose and location of a subject within large 3D scenes, using only wearable sensors. Left: subject wearing IMUs and a head mounted camera. Right: using the camera, HPS localizes the human in a pre-built map of the scene (bottom left). The top row shows the split images of the real and estimated virtual camera.

Abstract

We introduce (HPS) Human POSEitioning System, a method to recover the full 3D pose of a human registered with a 3D scan of the surrounding environment using wearable sensors. Using IMUs attached at the body limbs and a head mounted camera looking outwards, HPS fuses camera based self-localization with IMU-based human body tracking. The former provides drift-free but noisy position and orientation estimates while the latter is accurate in the short-term but subject to drift over longer periods of time.

We show that our optimization-based integration exploits the benefits of the two, resulting in pose accuracy free of drift. Furthermore, we integrate 3D scene constraints into our optimization, such as foot contact with the ground, resulting in physically plausible motion. HPS complements more common third-person-based 3D pose estimation methods. It allows capturing larger recording volumes and longer periods of motion, and could be used for VR/AR applications where humans interact with the scene without requiring direct line of sight with an external camera, or to train agents that navigate and interact with the environment based on first-person visual input, like real humans.

With HPS, we recorded a dataset of humans interacting with large 3D scenes (300-1000 m^2) consisting of 7 subjects and more than 3 hours of diverse motion. The dataset, code and video will be available on the project page: http://virtualhumans.mpi-inf.mpg.de/hps/.

1. Introduction

Capturing the full 3D pose of a human, while localizing and registering it with a 3D reconstruction of the environment, using *only wearable sensors*, opens the door to many applications and new research directions. For example, it will allow Augmented / Mixed / Virtual Reality users to move freely and interact with virtual objects in the scene,

* Joint first authors with equal contribution.

without the need for external cameras. From the captured data, we could train digital humans that plan and move like real humans, based on visual data arriving at their eyes. Moreover, by relying only on ego-centric data, we could capture a wider variety of human motion, outside of a restricted recording volume imposed by external cameras.

The dominant approach in vision has been to analyze humans from an *external third-person camera*, often without considering scene context [4, 30, 39, 45, 51, 55]. A few recent methods capture 3D scenes and humans [24], but again using a third-person camera. Capturing with external cameras is undoubtedly a central problem in vision, but it has its limitations – occlusions are a problem, and interactions across multiple rooms or beyond the viewing area cannot be captured; consequently recordings are typically short.

We propose *Human POSEitioning System* (HPS), the first method to recover the full body 3D pose of a human registered with a large 3D scan of the surrounding environment relying *only on wearable sensors* – body-mounted IMUs and a head mounted camera, approximating the visual field of view of the human. Inspired by visual-inertial odometry and localization [29, 40], as well as IMU-based human pose estimation [50, 71, 73], HPS fuses information coming from body-mounted IMUs with camera pose obtained from camera self-localization [57,59,64] (see Fig. 1). Instead of placing the camera towards the body [52, 67], we place it towards the scene, which allows us to capture what the human observes together with their 3D pose. In comparison to third-person pose methods, the body is not seen by the camera, which poses new challenges.

Pure IMU-based tracking is known to drift over time and camera localization produces many outliers. By jointly integrating IMU tracking with camera self-localization, we are able to remove drift [29, 40], and recover the human trajectory when self-localization fails. Furthermore, since we can approximately locate the person in the 3D scene, we incorporate scene constraints when foot contact is detected. Overall, with HPS we recover natural human motions, registered with the 3D scene and free of drift, during *long periods* of time, and over *large areas*.

To demonstrate the capabilities of HPS, we capture a dataset of real people moving in large scenes. Our HPS dataset consists of 8 types of environments - some being larger than $1000m^2$, and 7 subjects performing a variety of activities such as walking, excercising, reading, eating, or simply working in the office. The dataset can be used as a testbed for ego-centric tracking with scene constraints, to learn how humans interact and move within large scenes over long periods of time, and to learn how humans process visual input arriving at their eyes.

We make the following contributions: **1**) to the best of our knowledge, HPS is the first approach to estimate the full 3D human pose while localizing the person within a prescanned large 3D scene using wearable sensors. **2**) we introduce a joint optimization which integrates camera localization, IMU-based tracking and scene constraints, resulting in smooth and accurate human motion estimates. **3**) we provide the *HPS dataset*, a new dataset consisting of 3D scans of large scenes (some larger than 1000 m^2), ego-centric video, IMU data, and our 3D reconstructed humans moving and interacting with the scene. In contrast to existing 3D pose datasets, which are captured from a third-person view, ours is captured from an egocentric view. We believe both HPS and HPS dataset will provide a step towards developing future algorithms to understand and model 3D human motion and behavior within the 3D environment from an egocentric (or third-person) perspective.

2. Related Work

IMU-based 3D Human Pose Estimation: Although commercial solutions for IMU-based pose estimation have improved the stability of earlier solutions [53], they still suffer from severe drift, especially in the global orientation and location of the body. Early work [70] developed a custom suit to capture 3D human pose during daily activities. One line of work has focused on reducing the amount of IMUs necessary to capture motion via space-time optimization [73] or with deep learning [26]. In order to reduce drift and improve accuracy, visual-inertial approaches combine IMUs with multiple external cameras [42, 49, 50, 69, 72], a depth-camera [25,85] or even a single hand-held RGB camera [71]–which allowed collecting the 3DPW [71] dataset with accurate 3D poses outdoors. However, they all require an external camera, which limits the field of view to be captured, or requires someone to follow the person being tracked. Instead, we mount the camera (approximating the person's field of view) on the head and use it to self-localize the person in the scene.

Ego-centric capture and prediction: In contrast to our method, most ego-centric body-capture approaches mount the camera on the head looking towards the body. While ego-centric capture has received considerable attention for activity recognition [6, 12, 17, 41, 54, 80], methods at most detect the upper body. For full body capture, a pioneering method [52] relied on a helmet with sticks holding a camera away from the body. More recent methods [67, 78] work reasonably well even when the camera is close to the head. However, the accuracy is still far from desired.

Another group of methods place the camera looking outwards (like humans), and aim at estimating 3D pose from the ego-centric view alone, but 3D poses are inaccurate and have high uncertainty [28, 81, 82]. These methods to infer 3D pose from an ego-centric view [28, 81, 82] would benefit from our captured data, which contains ego-centric video with corresponding accurate 3D pose registered with the environment. An alternative approach places many cam-

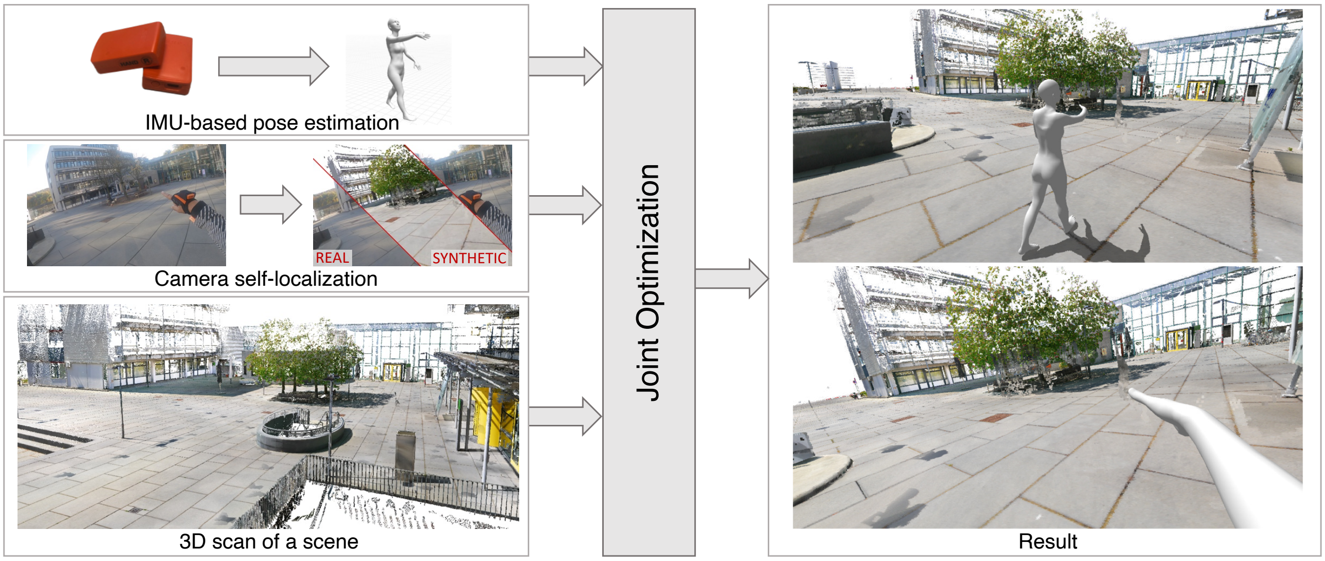

Figure 2. **Overview.** We use IMU data, RGB video from a head mounted camera, and a pre-scanned scene as input. We obtain an approximate 3D body pose using IMU data, and use head camera self-localization to localize the subject in the 3D scene. We then integrate the approximate body pose, the camera position and orientation, along with the 3D scene in a joint optimization to obtain the final location and pose estimates. We urge readers to see the video at http://virtualhumans.mpi-inf.mpg.de/hps/.

eras on the body looking out and use multi-camera structure from motion [61], but it can only recover slow motions.

Camera Localization: Most 6-DoF camera localization algorithms can be split into three groups. The first group is *structure-based* [11,14,37,59,62,63,65,66], which matches 2D points in the query image with 3D scene keypoints to estimate the camera pose by minimizing the reprojection error. While they provide precise position in small scenes, they do not scale to large scenes as matching becomes ambiguous and computationally expensive.

The second group of methods is referred to as *image-based*. The idea is to retrieve nearest neighbors in an image database based on a global descriptor [5,68,76]. The camera pose can then be approximated by the known poses of the retrieved images. They are more robust and scalable compared to structure based methods, but less precise, and the quality depends on the size of the image database.

In the third group are *hybrid approaches* [10, 56, 57] which combine the benefits of the last two. First, a set of relevant database images are found using an image-based method, and then the precise camera pose is recovered using structure-based methods. Another set of methods directly regress the camera pose using a CNN [60,74], but their accuracy leaves a lot to be desired. *Hybrid approaches* have been shown to be precise and to scale to large scenes, and hence the self-localization part of HPS builds upon them.

Humans and Scenes: The relationship between humans, scenes, and objects is a recurrent subject of study in vision. Examples are methods for 2D pose and object detection [15, 21, 27, 31, 48, 79], 3D object detection using human poses [20, 22], learning to insert people in scenes [19,35,75,84], constraining pose [24,83], estimating forces [36], or predicting long term motion [13] conditioned on the scene. *Most approaches predict only static poses in a single room*, and reasoning is done from a third-person perspective. In contrast, our analysis is from a first-person perspective, and uses the scene to self-localize the human in it. Furthermore, our method enables to capture humans in motion in multiple-room and outdoor environments. All aforementioned methods would benefit from the HPS dataset.

3. Method

Our goal is to recover the 3D body pose and location of a subject in a known scene from egocentric measurements. To this end, our method requires as input: 1) a head-mounted camera, 2) body-mounted IMUs, and 3) a pre-built 3D scan of a scene, along with a database of RGB scene images with known camera parameters. Using camera data, our method localizes the person within a pre-scanned 3D scene (Sec. 3.2), estimates their 3D pose using IMUs (Sec. 3.3), and in a joint optimization step (Sec. 3.4) integrates camera localization, IMU pose estimates and scene constraints, resulting in smooth and accurate human motion estimates. For an overview of our method, see Fig. 2. For more details on the 3D scene reconstruction, image database collection, camera and IMU setup, we refer to the supplementary.

3.1. SMPL Body Model

We use the Skinned Multi-Person Linear (SMPL) body model [38] to represent the human subject. SMPL is a differentiable function $M(\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}) : \mathbb{R}^{72\times3\times10} \mapsto \mathbb{R}^{6890\times3}$ that maps pose $\boldsymbol{\theta}$, translation $\boldsymbol{t}$ and shape $\boldsymbol{\beta}$ parameters

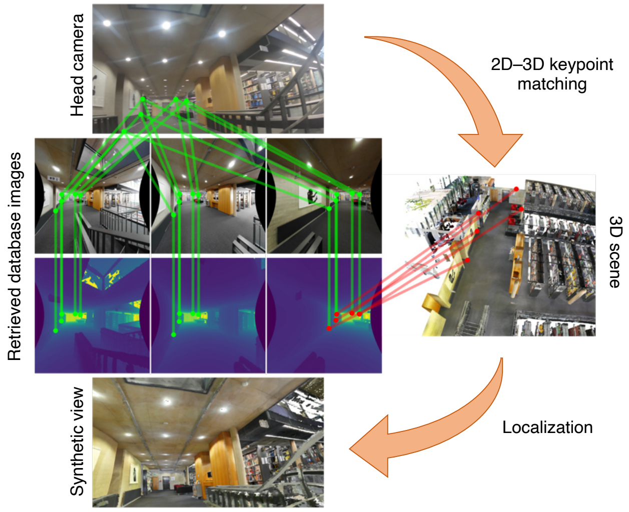

Figure 3. **Camera self-localization.** We match the head camera image keypoints with the keypoints from the prefiltered database with known 2D-3D scene correspondences. We then localize the camera in the scene by minimizing a reprojection error of the keypoints. *From top to bottom:* head camera image (query), top-3 retrieved images from a dataset, depthmaps rendered from the same position to map 2D database keypoints to 3D, synthetic view of the scene from the inferred camera position.

to the vertices of a watertight human mesh. The underlying skeleton of SMPL has 24 joints. The pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ correspond to the relative orientation of each joint in the SMPL skeleton, expressed in axis-angles. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the PCA coefficients of a shape space learnt from a corpus of registered scans. We use the notation $M_n(\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{\beta}) \in \mathbb{R}^3$ to indicate the n^{th} vertex of SMPL. We obtain approximate shape parameters $\boldsymbol{\beta}$ of a person from body measurements. We assume that $\boldsymbol{\beta}$ remains constant during a sequence and aim to recover $\boldsymbol{\theta}$ and $\boldsymbol{t}$ of the subject registered with the 3D environment. Henceforth we drop $\boldsymbol{\beta}$ for notational convenience.

3.2. Camera Self-localization

The camera self-localization stage aims to estimate the position and orientation of the human head from a head-mounted camera. To scale to large scenes, we use a hierarchical structure-based localization algorithm [57, 58] (Fig. 3). It first identifies a set of potentially relevant database images, *i.e.*, images used to build the 3D scene map, through image retrieval via NetVLAD [5] descriptors. 2D-3D matches are established between local SuperPoint [16] features extracted in the query image and 3D points visible in the top-40 retrieved images. These matches are then used to estimate the camera pose by applying a P3P solver [23, 32, 33] inside a RANSAC loop [18] with local optimization [34]. Rather than building a separate sparse Structure-from-Motion point cloud for localization, as originally used in [57], we obtain 3D point positions from our dense scene 3D model [64]. For each pixel in a database image, we obtain the corresponding 3D point by rendering the 3D model from the known pose of the image. 2D-2D matches between the query and the top-40 retrieved database images thus yield the required 2D-3D matches. From the camera self-localization step, we obtain estimates for camera orientation $\mathbf{R}^C$ and position $\boldsymbol{t}^C$.

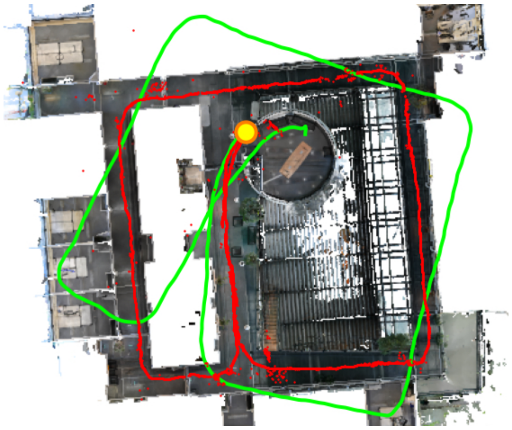

Figure 4. **Comparison of the trajectories** of IMUs (in green) with camera self-localization (in red). The yellow dot marks the start. Notice the red trajectory is free of drift but has outliers.

3.3. IMU based Pose Estimation

We use a commercial inertial mo-cap system provided by XSens [47], which uses 17 IMUs attached to the body with velcro-straps or a suit. XSens IMUs provide 3D pose estimates, denoted as $\boldsymbol{\theta}^I$ and location estimates relative to the starting position of a recording - denoted as $\boldsymbol{t}^I$, using a proprietary algorithm based on a Kalman filter and a kinematic model of the human body to reduce drift. While it provides accurate articulation, our experiments show that the global orientation and position drift significantly over time, and consequently scene constraints are not satisfied (Fig. 4, 6). Using acceleration information, IMUs also detect feet contacts with the ground, which we integrate in our joint optimization algorithm.

3.4. Joint Optimization

Our joint optimization algorithm finds the pose parameters of the SMPL body model in order to satisfy i) the head camera self-localization, ii) scene, and iii) smoothness constraints while remaining as close as possible to the IMU pose estimate $\boldsymbol{\theta}^I$ (excluding global orientation and position) – while we could optimize SMPL to match the raw IMU data directly [71, 73], we chose not to, because it contains a lot of drift. Mathematically, we minimize the following objective over a batch of T frames (T is fixed for all scenes)

$$E(\boldsymbol{\theta}_{1:T}, \boldsymbol{t}_{1:T}) = w_s E_{\text{self}} + w_{sc} E_{\text{scene}} + w_{\text{sm}} E_{\text{sm}} + w_p E_{\text{IMU}}, \tag{1}$$

with respect to pose $\boldsymbol{\theta}_{1:T}$ and translation $\boldsymbol{t}_{1:T}$ parameters. $\boldsymbol{\theta}_{1:T} \in \mathbb{R}^{72T}$ and $\boldsymbol{t}_{1:T} \in \mathbb{R}^{3T}$ are stacked model poses and translations for each time step $j = 1 \ldots T$. In the following, we explain each of the terms in more detail.

Self-localization Term E_{self}: We use the estimated orientation of the camera to constrain the orientation of SMPL. Specifically, we minimize the geodesic distance [73] from the head camera orientation as inferred from SMPL, $\overline{R}^C(\boldsymbol{\theta})$, to the self-localization estimate $\mathbf{R}^C$ over a batch of frames T:

$$E_{\text{self}} = \frac{1}{T} \sum_{j=1}^{T} ||(\log((\overline{R}^C(\boldsymbol{\theta}_j))^\top \mathbf{R}_j^C))^\vee||_2 \ , \tag{2}$$

where the $\log$ operation recovers the skew-symmetric matrix from the relative rotation matrix, and the $^\vee$ operator converts it to its axis-angle representation. The mapping $\overline{R}^C(\boldsymbol{\theta}_j)$ can be derived as follows. First, we obtain the head bone orientation by traversing the kinematic chain of SMPL

$$R^H(\boldsymbol{\theta}) = \prod_{i \in \mathcal{P}_{\text{Head}}} \exp(\widehat{\boldsymbol{\theta}^i}) \ , \tag{3}$$

where $\mathcal{P}_{\text{Head}}$ is an ordered list of all the parents to the head joint. The $^\wedge$ operator maps an axis-angle to its corresponding skew-symmetric matrix and $\exp(\widehat{\boldsymbol{\theta}^i})$ are the relative joint rotation matrices obtained from $\boldsymbol{\theta}^i \in so(3)$ using the Rodrigues formula. While $R^H : \mathbb{R}^{72} \mapsto \text{SO}(3)$ maps from pose to head rotation, we need a mapping to camera orientation. Since the camera is rigidly attached to the head, there is a constant camera to head offset that can be estimated at frame 0 [50, 73]:

$$\mathbf{R}_{HC} = (R^H(\boldsymbol{\theta}_0^I))^\top \mathbf{R}_0^C \ . \tag{4}$$

We find the desired mapping from pose to camera at a subsequent frame j as $\overline{R}^C(\boldsymbol{\theta}_j) = R^H(\boldsymbol{\theta}_j)\, \mathbf{R}_{HC}$.

Scene Contact Term E_{scene}: When the IMUs detect a foot contact, we force it to be in contact with the ground by using an energy term consisting of two subterms $E_{\text{scene}} = w_c E_{\text{contact}} + w_v E_{\text{slide}}$. Let $\mathcal{B}_k$ with $k \in [1, 2, 3, 4]$ denote 4 sets of manually defined vertex indices in the SMPL corresponding to the toe and heel regions for the left and right foot (more details in supplementary), and let $c_j^k \in [0, 1]$ be a binary variable indicating if part k is in contact with the ground at frame j. We define the following contact term, which snaps the foot vertices to the closest scene vertices

$$E_{\text{contact}} = \frac{1}{4T} \sum_{j=1}^{T} \sum_{k=1}^{4} \sum_{n \in \mathcal{B}_k} \frac{1}{|\mathcal{B}_k|} c_j^k ||M_n(\boldsymbol{\theta}_j, \boldsymbol{t}_j) - v(n)||_2 \ , \tag{5}$$

where $M_n(\boldsymbol{\theta}_j, \boldsymbol{t}_j)$ is the n^{th} vertex of the SMPL mesh at frame j, and $v(n) = \underset{\boldsymbol{v}_s \in \mathbf{V_s}}{\operatorname{argmin}}(||M_n(\boldsymbol{\theta}_j, \boldsymbol{t}_j) - \boldsymbol{v}_s||_2)$ returns the closest scene point $\boldsymbol{v}_s \in \mathbf{V_s}$ to $M_n(\boldsymbol{\theta}_j, \boldsymbol{t}_j)$. To prevent the foot from sliding when in contact with the scene, we also constrain the distance between foot parts in contact with the scene in two successive frames to be zero.

$$E_{\text{slide}} = \frac{1}{4(T-1)} \sum_{j=1}^{T-1} \sum_{k=1}^{4} \sum_{n \in \mathcal{B}_k} \frac{1}{|\mathcal{B}_k|} c_j^k c_{j+1}^k ||M_n(\boldsymbol{\theta}_j, \boldsymbol{t}_j) - M_n(\boldsymbol{\theta}_{j+1}, \boldsymbol{t}_{j+1})||_2 \ . \tag{6}$$

Smoothness Term E_{sm}: This term ensures smooth changing of the global translation and orientation, as well as head orientation

$$E_{\text{sm}} = w_T E_T + w_G E_G + w_H E_H, \tag{7}$$

where the translation term equals:

$$E_T = \frac{1}{T-1} \sum_{j=1}^{T-1} ||(\boldsymbol{t}_j - \boldsymbol{t}_{j+1})||_2 \ . \tag{8}$$

Defining $R^G : \mathbb{R}^{72} \mapsto \text{SO}(3)$ as $R^G(\boldsymbol{\theta}) = \exp(\widehat{\boldsymbol{\theta}^G})$ where $\boldsymbol{\theta}^G$ is the axis-angle representation of the root (global) joint, the global orientation smoothness term is

$$E_G = \frac{1}{T-1} \sum_{j=1}^{T-1} ||(\log((R^G(\boldsymbol{\theta}_j))^\top R^G(\boldsymbol{\theta}_{j+1})))^\vee||_2 \tag{9}$$

Using Eq (9), the head orientation smoothness term is enforced with an equivalent term replacing R^G by R^H.

Pose Term E_{IMU}: The pose recovered by IMUs captures the articulation of the body well, but is inaccurate for global orientation and translation. Hence, we constrain the pose parameters corresponding to the body to remain close to the IMUs estimate. Let $\mathbf{B}$ be an identity matrix with zeros at the diagonal entries corresponding to the root joint. With this, the pose is regularized with the following equation:

$$E_{\text{IMU}} = \frac{1}{T} \sum_{j=1}^{T} \sqrt{(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^I)^\top \mathbf{B} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^I)} \ . \tag{10}$$

For implementation details of our joint optimization algorithm, please see the supplementary.

3.5. Initialization

Since the objective function in Eq. (1) is highly nonconvex, convergence to a good minimum hinges on good initialization. We initialize translation parameters $\boldsymbol{t}_j$ using camera localization estimates $\boldsymbol{t}_j^C$. Camera localization results are typically noisy, so instead of using raw results we first detect outliers by computing the velocity of translation between each result and its inlier neighbours. We mark a result as an outlier if its velocity exceeds the threshold $\epsilon = 3m/s$. We repeat this process until convergence and replace all outliers by interpolation.

For poses $\boldsymbol{\theta}$, the simplest choice is to initialize with the IMU pose estimate $\boldsymbol{\theta}_j = \boldsymbol{\theta}_j^I$. However, the global body orientation often deviates from the more accurate self-localization trajectory (see Fig. 4, 6). Observing that the body orientation is often perpendicular to the trajectory, our idea is to rotate the IMU pose to align it to the self-localization trajectory. To this end, we first estimate the tangent direction of the self-localization and IMU trajectories

$$\boldsymbol{v}_j^C = \frac{\boldsymbol{t}_{j+\gamma}^C - \boldsymbol{t}_j^C}{||\boldsymbol{t}_{j+\gamma}^C - \boldsymbol{t}_j^C||_2} \quad , \quad \boldsymbol{v}_j^I = \frac{\boldsymbol{t}_{j+\gamma}^I - \boldsymbol{t}_j^I}{||\boldsymbol{t}_{j+\gamma}^I - \boldsymbol{t}_j^I||_2} \quad ,$$

($\gamma = 10$ in our case) and correct the root orientation $\exp(\widehat{\boldsymbol{\theta}_j^{I,G}})$ of the IMU pose with the following formula

$$\boldsymbol{\theta}_j^{I,G*} = (\log(\exp(\widehat{\boldsymbol{v}_j^I \times \boldsymbol{v}_j^C})\exp(\widehat{\boldsymbol{\theta}_j^{I,G}})))^\vee \quad , \tag{11}$$

where $\exp(\widehat{\boldsymbol{v}_j^I \times \boldsymbol{v}_j^C})$ is the planar rotation that aligns $\boldsymbol{v}_j^I$ with $\boldsymbol{v}_j^C$. For stationary frames, we use the correction matrix of the last frame with non-zero velocity. We find that in practise, for stationary frames, this a good approximation.

3.6. Coordinate Frame Alignment

While the camera estimates $\mathbf{R}^C$ and $\boldsymbol{t}^C$ are in the 3D scene coordinates, IMU estimates $\boldsymbol{\theta}^I$ and $\boldsymbol{t}^I$ are not. Before the initialization step (Sec. 3.5) of our joint optimization algorithm (Sec. 3.4), we align the IMU coordinate frame with the 3D scene frame by finding a planar rotation $\mathbf{R}_A^*$ that orients the SMPL head at frame zero $R^H(\boldsymbol{\theta}_0^I)$ to match the camera orientation $\mathbf{R}_0^C$ at the same frame. Mathematically, this entails minimizing the following objective

$$\mathbf{R}_A^* = \underset{\mathbf{R}_A \in \mathcal{R}}{\text{argmin}} ||(\log(\mathbf{R}_A R^H(\boldsymbol{\theta}_0^I))^\top \mathbf{R}_0^C))^\vee||_2 \quad . \tag{12}$$

We use the axis-angle parameterization to define the set of rotation matrices $\mathcal{R} = \{\exp(\widehat{x\boldsymbol{\alpha}}) : x \in \mathbb{R}\}$. where $\boldsymbol{\alpha} = [0, 0, 1]^\top$ is the z-axis unit vector. The IMU pose $\boldsymbol{\theta}_j^I$ and position $\boldsymbol{t}_j^I$ estimate of each subsequent frame are aligned to the 3D scene reference frame by

$$\boldsymbol{\theta}_j^{I,G} = (\log(\mathbf{R}_A^* \exp(\widehat{\boldsymbol{\theta}_j^{I,G}})))^\vee \quad , \boldsymbol{t}_j^I = \mathbf{R}_A^* \boldsymbol{t}_j^I \quad . \tag{13}$$

4. Dataset

HPS allows us to collect the *HPS dataset* - a dataset of 3D humans interacting with large 3D scenes (300-1000 m^2, up to 2500 m^2). Our dataset contains images captured from a head-mounted camera coupled with the reference 3D pose and location of the person in a pre-scanned 3D scene. We capture 7 people in 8 large scenes performing activities such as exercising, reading, eating, lecturing, using a computer, making coffee, dancing. *All subjects have agreed to release their data for research purposes.* In total, the dataset provides more than 300K synchronized RGB images coupled with the reference 3D pose and location. We plan to keep updating the dataset by adding more long-term motion recordings with a variety of scene interactions. Figure 7 shows qualitative results from our dataset. For more examples, please see the video [1].

Distance traveled	IMU	IMU + Cam	IMU + Cam (filtered)	HPS w\o scene	HPS
At start	6.85	9.24	10.48	7.21	**5.20**
70 m	54.49	742.32	6.93	6.48	**4.60**
200 m	69.02	136.81	5.93	5.80	**4.26**
380 m	108.44	32.17	6.15	5.69	**4.53**

Table 1. **Drift and cam. outliers:** 3D error (in cm) for the subject standing in A-pose after moving freely around the scene.

Distance traveled	IMU	IMU + Cam	IMU + Cam (filtered)	HPS w\o scene	HPS
At start	6.77	2189.75	10.05	9.19	**6.44**
70 m	51.57	569.71	21.75	20.68	**15.96**
200 m	61.11	719.44	7.34	6.67	**4.76**
380 m	100.44	261.72	12.59	11.96	**10.07**

Table 2. **Drift and cam. outliers (dynamic):** 3D error (in cm) for the subject walking, standing and leaning on the table, after moving around the scene. Error is measured from the dynamic ground truth point cloud to the result (3D mesh in motion). Rows indicate distance traveled before evaluation.

5. Experiments

This section shows that HPS does not drift with time and distance traveled, is robust to non-persistent camera localization outliers, and satisfies scene constraints (feet stay on the ground during contact, and do not slide).

Since this is the first method to track humans in large scenes, there exist no published baselines to compare to, and ground truth 3D human pose and localization cannot be obtained for unbounded areas like ours. Hence, we use depth cameras to obtain ground truth dynamic point clouds of the human in a small sub-area of the scene. Subjects are then asked to move freely in the large scene, and return to the sub-area, where we can evaluate accuracy and drift.

5.1. Quantitative Evaluation

We evaluate the accuracy of our method by comparing our output SMPL mesh (including translation) with a dynamic *ground-truth point cloud* of the person obtained from three synchronized and calibrated external depth cameras (Azure Kinect [2]). We register the point cloud to the scene in three steps involving camera self-localization, ICP, and manual correction. For an explanation of the Kinect setup and point cloud registration we refer to the supplementary. We report the bidirectional Chamfer distance between the

Metric	IMU	IMU + Cam	IMU + Cam (filtered)	HPS w\o scene	HPS
Dist. to Surf.	188.38	39.8	0.95	0.32	**0.056**
Foot Sliding	0.92	52.09	1.75	2.00	**0.90**

Table 3. **Foot contact:** For frames when foot contact is detected, we report (in cm) **Distance to surface**: Average distance between foot vertices and the scene, and **Foot Sliding**: Average distance on the surface plane between foot vertices in two successive frames. Numbers are computed for a 3 minute long walking sequence.

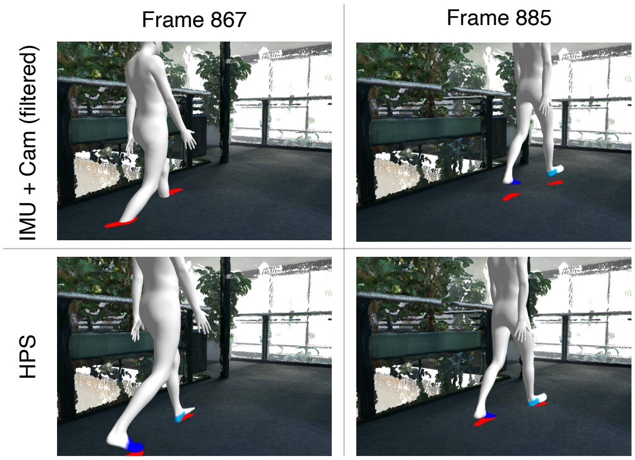

Figure 5. **Effect of integrating predicted 3D scene contacts.** As a baseline we used camera localization results for localizing SMPL model. Red regions mark closest surface to feet, heels and toes are colored with light blue and blue when IMUs detect ground contact.

SMPL model (result) and ground truth point cloud from depth sensors *without* Procrustes alignment.

Movements: For quantitative evaluation, we record using the following protocol: a subject starts within the recording volume of the three RGB-D sensors and performs different actions including standing in A-pose, leaning on a table and walking. The subject then leaves the recording volume and moves within the scene, returns back and repeats the same actions inside that volume again. This is repeated several times, each time choosing a different path.

Baselines: There are no established baselines to compare to, as no other method tackles the same problem. Hence, to understand the influence of each component, we use the following baselines: 1) **IMU:** pure IMU tracker, 2) **IMU+Cam**: pose from IMU, and translation from camera self-localization, 3) **IMU+Cam (filtered)**: Like IMU+Cam but with filtered camera outliers (same as in Sec. 3.5), 4) **HPS w\o scene**: Optimization without 3D scene contact constraints.

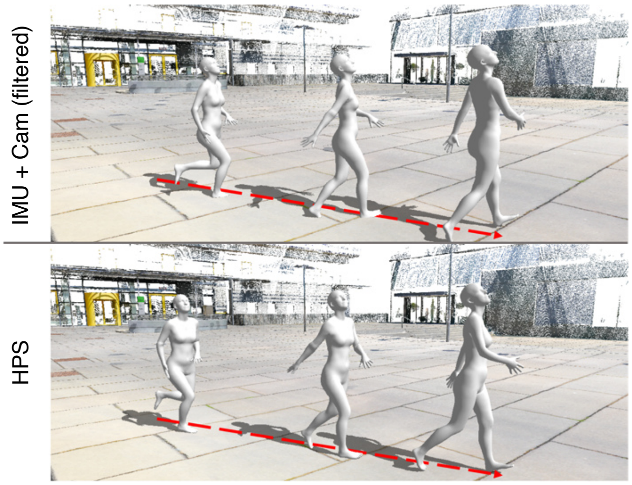

Figure 6. **Global body orientation improvement.** Combining the IMU pose with position from camera localization (IMU+Cam (filtered)) results in unnatural motion–the global body orientation does not face the direction of movement. By contrast, HPS correctly estimates the the global orientation. We refer to the video at project page [1] for more visual examples.

Drift and Outliers: In Tables 1 and 2, we compare HPS to the baselines. We observe that the IMU-only method drifts over time, particularly the global body translation and orientation. IMU+Cam corrects drift with camera localization, but produces translation noise and severe jitter. IMU+Cam (filtered) mitigates this, but lacks precision and suffers from global orientation errors (Fig. 6). HPS w\o scene further improves results, but without knowledge about foot-scene contacts, it is easily misled by incorrect camera localization, and the subject penetrates or flies over the ground. HPS results satisfy these scene constraints, and consistently achieve the best accuracy. HPS is inaccurate when filtered camera localization fails for a long period (see 2nd and 4th rows of Table 2), but it can recover once the camera can be well localized in nearby frames (see 3rd row of Table 2). Overall, the analysis reveals that HPS does not drift (error does not increase with distance traveled or time), and is robust to non-persistent camera localization outliers.

For scenes with with persistent camera localization failures (outdoor scenes, indoor scenes with repetitive patterns), we implemented a slightly modified version of HPS, described in the supplementary.

Foot contacts: We also report in Table 3 the average foot-to-scene distance and foot-sliding-along-the-surface distance during contacts detected with the IMUs. HPS better preserves foot contact with the surface than the baselines, and has slightly lower foot-sliding compared to the raw IMU tracker, which also integrates constraints with a *virtual* imaginary ground. Foot contacts in HPS result in stable and natural motion, see Fig. 5, and the video [1].

5.2. Qualitative evaluation

In Fig. 5 we show the effect of foot contact constraints. As we encourage contact with the scene surface each time a contact is detected, the human mesh does not fly in the air or penetrate the ground like the baseline. The motion is more stable and physically correct. In Fig. 7 we show examples of

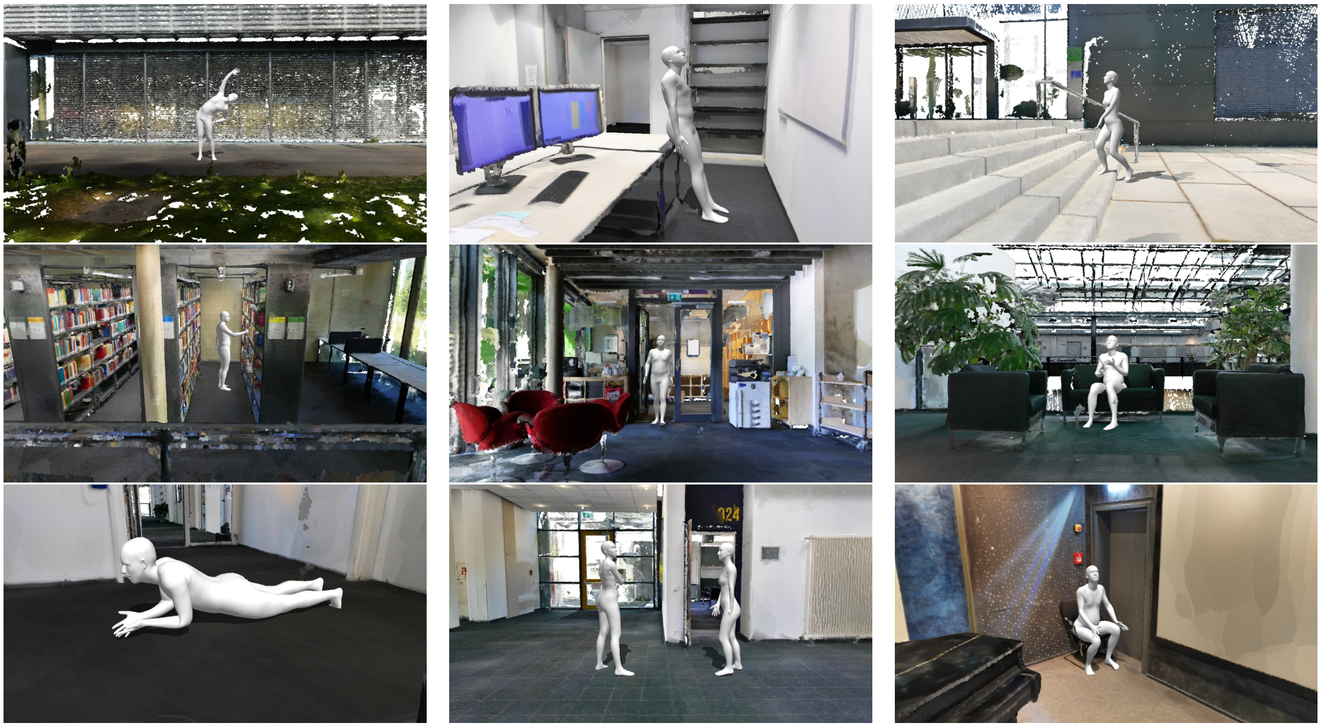

Figure 7. **We show qualitative results of our method.** Our method can localize and estimate the 3D pose of people performing activities as diverse as exercising, dancing, reading, sitting, eating, talking in a range of indoor and outdoor scenes, all *without* external cameras.

humans performing different actions including sitting, leaning on a table, dancing or performing push-ups. For more examples, please see the video at our project page [1].

6. Conclusions and Future Work

We introduced HPS, to the best of our knowledge, the first method to estimate full body pose registered with a pre-scanned 3D environment from *only wearable sensors*. We demonstrate that HPS produces natural human motion, removes the typical drift of pure IMU based systems, and is robust to non-persistent camera localization outliers. HPS is able to continuously track humans in large scenes ($300 - 1000m^2$) including multiple rooms and outdoors.

The error of HPS does not accumulate with time or distance traveled. However, if camera localization is inaccurate for long periods of time, HPS performance deteriorates. This can be seen in the errors, which range from $4cm$ to $15cm$. Two factors influence localization accuracy: 1) Lack of features, 2) scene changes between the static 3D scan and the real images, captured from the head camera.

While HPS achieves a remarkable accuracy and stability, many applications will require errors in localization and pose of less than $1cm$. We envision many exciting research directions to improve HPS. First, a local map could be built on the fly to update the large static scene with objects that move, and adding new objects. This would improve localization and allow interaction with dynamic objects. It is not inconceivable that, in the future, a dynamic 3D reconstruction of the world will be stored on the cloud, and will be continuously updated from cameras worn by people [3]. Second, camera localization could incorporate semantics [9, 86], e.g. detecting static and reliable objects. Third, while HPS integrates foot contacts, scene constraints with other body parts can further improve results. More powerful would be to learn a model to *anticipate human intent* to improve tracking. For example, we could detect when the person is about to sit on a chair, or about to grab an object. Conversely, HPS can be used to build models of environment interaction and navigation [43, 77] from human captures consisting of several hours, as we believe natural behavior arises only during long recordings. Fourth, we want to combine HPS with virtual humans of appearance [7, 8, 44, 46] to generate realistic data for training and evaluation of 3D human analysis methods.

HPS is the first step in a new exciting research direction. We will *release the HPS dataset and code* for research use [1], and hope it will foster new methods to perceive and model scenes and humans from an ego-centric perspective.

Acknowledgments: We thank Bharat Bhatnagar, Verica Lazova, Anna Kukleva and Garvita Tiwari for their feedback. This work is partly funded by the DFG - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), the EU Horizon 2020 project RICAIP (grant agreeement No.857306), and the European Regional Development Fund under project IMPACT (No. CZ.02.1.01/0.0/0.0/15 003/0000468).

References

[1] http://virtualhumans.mpi-inf.mpg.de/hps/. 6, 7, 8

[2] *Microsoft Azure Kinect*, accessed November 15, 2020. https://en.wikipedia.org/wiki/Azure_Kinect. 6

[3] *Project Aria*, accessed November 15, 2020. https://about.fb.com/realitylabs/projectaria/. 8

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2293–2303. IEEE, Oct 2019. 2

[5] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 3, 4

[6] 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. 2

[7] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 8

[8] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 8

[9] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam—learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018. 8

[10] Eric Brachmann and Carsten Rother. Expert Sample Consensus Applied to Camera Re-Localization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

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

[12] 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. 2

[13] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*. 2020. 3

[14] 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

[15] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *European Conference on Computer Vision*, pages 158–172. Springer, 2012. 3

[16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 4

[17] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding egocentric activities. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2

[18] M. Fischler and R. Bolles. Random Sampling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. *Communications of the ACM (CACM)*, 24:381–395, 1981. 4

[19] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. *International journal of computer vision*, 110(3):259–274, 2014. 3

[20] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR 2011*, pages 1529–1536. IEEE, 2011. 3

[21] Abhinav Gupta and Larry S Davis. Objects in action: An approach for combining action understanding and object perception. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[22] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, pages 1961–1968. IEEE, 2011. 3

[23] R.M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal of Computer Vision (IJCV)*, 13(3):331–356, 1994. 4

[24] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings International Conference on Computer Vision*, pages 2282–2292. IEEE, Oct. 2019. 2, 3

[25] Thomas Helten, Andreas Baak, Gaurav Bharaj, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *International Conf. on 3D Vision*, pages 279–286, 2013. 2

[26] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):185:1–185:15, nov 2018. 2

[27] Umar Iqbal, Martin Garbade, and Juergen Gall. Pose for action-action for pose. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 438–445. IEEE, 2017. 3

[28] 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. 2

[29] 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. 2

[30] 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

[31] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011. 3

[32] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 4

[33] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Closed-Form Solutions to Minimal Absolute Pose Problems with Known Vertical Direction. In *Asian Conference on Computer Vision (ACCV)*, 2010. 4

[34] Karel Lebeda, Juan E. Sala Matas, and Ondřej Chum. Fixing the Locally Optimized RANSAC. In *British Machine Vision Conference (BMVC)*, 2012. 4

[35] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12368–12376, 2019. 3

[36] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8640–8649, 2019. 3

[37] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2372–2381, 2017. 3

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015. 3

[39] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[40] 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*, volume 1, page 1, 2015. 2

[41] 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. 2

[42] Charles Malleson, Marco Volino, Andrew Gilbert, Matthew Trumble, John Collomosse, and Adrian Hilton. Real-time full-body motion capture from video and imus. In *2017 Fifth International Conference on 3D Vision (3DV)*, 2017. 2

[43] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 8

[44] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 8

[45] Mohamed Omran, Christop Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conf. on 3D Vision*, 2018. 2

[46] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 8

[47] Monique Paulich, Martin Schepers, Nina Rudigkeit, and G. Bellusci. *Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications*, 05 2018. 4

[48] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013. 3

[49] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixé, Meinard Muller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *Proceedings of the 2011 International Conference on Computer Vision (ICCV)*, pages 1243–1250, 2011. 2

[50] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 663–670, 2010. 2, 5

[51] Gerard Pons-Moll and Bodo Rosenhahn. *Model-Based Pose Estimation*, chapter 9, pages 139–170. Springer, 2011. 2

[52] 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. 2

[53] Daniel Roetenberg, Henk Luinge, and Per Slycke. Moven: Full 6dof human motion tracking using miniature inertial sensors. *Xsen Technologies, December*, 2007. 2

[54] 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. 2

[55] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020. 2

[56] P.E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena. Leveraging deep visual descriptors for hierarchical efficient localization. In *Conference on Robot Learning*, Zurich, Switzerland, October 2018. 3

[57] 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. 2, 3, 4

[58] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*, 2020. 4

[59] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 2, 3

[60] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[61] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. Motion capture from body-mounted cameras. In *ACM Transactions on Graphics (TOG)*, volume 30, page 31. ACM, 2011. 3

[62] 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

[63] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1455–1461, 2016. 3

[64] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[65] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 3

[66] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic match consistency for long-term visual localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018. 3

[67] 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*, Oct 2020. 2

[68] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015. 3

[69] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017. 2

[70] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM Transactions on Graphics (TOG)*, 26(3):35, 2007. 2

[71] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conf. on Computer Vision*, sep 2018. 2, 4

[72] T von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and IMUs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1533–1547, 2016. 2

[73] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, pages 349–360, 2017. 2, 4, 5

[74] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-Based Localization Using LSTMs for Structured Feature Correlation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[75] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2596–2605, 2017. 3

[76] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016. 3

[77] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 8

[78] 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. 2

[79] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2010. 3

[80] 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. 2

[81] 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. 2

[82] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[83] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 3

[84] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020. 3

[85] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Quionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[86] Shuaifeng Zhi, Michael Bloesch, Stefan Leutenegger, and Andrew J Davison. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2019. 8