EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild

Manuel Kaufmann[1] Jie Song[1] Chen Guo[1] Kaiyue Shen[1] Tianjian Jiang[1]
Chengcheng Tang[2] Juan José Zárate[1] Otmar Hilliges[1]

[1]ETH Zürich, Department of Computer Science [2]Meta Reality Labs

Figure 1: EMDB is a novel dataset that provides accurate SMPL pose and shape parameters including global camera and body trajectories for in-the-wild videos. (left and middle) Challenging example poses taken from EMDB. (right, 1-4) Visualization of global body center trajectory and the corresponding 3D poses projected into the camera view.

Abstract

We present EMDB, the Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. EMDB is a novel dataset that contains high-quality 3D SMPL pose and shape parameters with global body and camera trajectories for in-the-wild videos. We use body-worn, wireless electromagnetic (EM) sensors and a hand-held iPhone to record a total of 58 minutes of motion data, distributed over 81 indoor and outdoor sequences and 10 participants. Together with accurate body poses and shapes, we also provide global camera poses and body root trajectories. To construct EMDB, we propose a multi-stage optimization procedure, which first fits SMPL to the 6-DoF EM measurements and then refines the poses via image observations. To achieve high-quality results, we leverage a neural implicit avatar model to reconstruct detailed human surface geometry and appearance, which allows for improved alignment and smoothness via a dense pixel-level objective. Our evaluations, conducted with a multi-view volumetric capture system, indicate that EMDB has an expected accuracy of 2.3 cm positional and 10.6 degrees angular error, surpassing the accuracy of previous in-the-wild datasets. We evaluate existing state-of-the-art monocular RGB methods for camera-relative and global pose estimation on EMDB. EMDB is publicly available under https://ait.ethz.ch/emdb.

1. Introduction

3D human pose and shape estimation from monocular RGB images is a long-standing computer vision problem with many applications in AR/VR, robotics, assisted living, rehabilitation, or sports analysis. Much progress has been made in estimating camera-relative poses, typically assuming a weak-perspective camera model, *e.g.*, [7, 25, 29, 30, 51]. However, this setting is too restrictive for many applications that involve a moving camera. Such applications must estimate a) human poses in-the-wild, under occlusion and encountering uncommon poses; and b) global locations of humans and the camera. Compared to the camera-relative setting, there is relatively little work on global pose estimation [65, 69]. This is in part due to the lack of comprehensive datasets that contain accurate 3D human pose and shape with global trajectories in a fully in-the-wild setting.

To overcome this bottleneck, in this paper we propose a novel dataset, called EMDB, short for the **E**lectro**M**agnetic **D**ata**B**ase of Global 3D Human Pose and Shape in the Wild. EMDB consists of 58 minutes (105k frames) of challenging 3D human motion recorded in diverse scenes. We provide high-quality pose and shape annotations, as well as global body root and camera trajectories. The dataset contains 81 sequences distributed over 10 participants that were recorded with a hand-held mobile phone.

Recording such data requires a motion capture system that is both mobile and accurate – a notoriously difficult problem. Systems that provide world-anchored 3D body keypoints often require multiple well-calibrated RGB or IR cameras within a static environment, which restricts outdoor use [16, 18, 20, 37]. While body-worn sensors such as head-mounted cameras [46, 63, 72] are promising for mobile use, such egocentric approaches introduce either heavy self-occlusions [46, 63] or are restricted to indoor settings with a fixed capture volume [72]. The 3DPW dataset [59] uses IMU sensors for outdoor recordings, yet the dataset is relatively small and lacks global trajectories. Moreover, IMU drift and the lack of direct positional sensor measurements imposes constraints in terms of pose diversity and accuracy. Instead, following [23], we leverage drift-free electromagnetic (EM) sensors that directly measure their position and orientation. Yet, any sensor-based capture system requires handling of measurement noise, accurate calibration of the sensors to the body's coordinate system and temporal and spatial alignment of the data streams.

Addressing these challenges, we propose a method, Electromagnetic Poser (EMP), that allows for the construction of EMDB. EMP is a multi-stage optimization formulation that fuses up to 12 body-worn EM sensor measurements, monocular RGB-D images and camera poses, and produces accurate SMPL [34] pose and shape parameters alongside global trajectory estimates for the body's root and the camera. EMP works in the following 3 stages.

Calibration and EM Pose: As an initial calibration step, we scan participants in minimal clothing using an indoor multi-view volumetric capture system (MVS, [8]) to obtain ground-truth shape and skin-to-sensor offsets. We subsequently record in-the-wild sequences of the same subject and fit SMPL to the drift-free EM measurements of the sensors' positions and orientations. This provides an accurate SMPL fit, albeit in a EM-local coordinate system.

World Alignment: In the second stage, we align the EM-local pose estimates with a global world space, defined by the tracking space of a hand-held iPhone 13 that films the participants. We model this stage as a joint optimization that fuses the input EM measurements, 2D keypoints, depth, and camera poses. In our experiments we have found that the self-localized 6D poses of the iPhone are accurate to around 2 cm positional and < 1 degree angular error. The fixed body shape and accurate camera poses thus enable EMP to provide global SMPL root trajectories.

Pixel-Level Refinement: In the third stage, we refine the initial global poses via dense pixel-level information to ensure high-quality and temporally smooth image alignment. To this end we leverage recent advancements in neural body modelling for in-the-wild videos and fit a neural body model with detailed geometry and appearance to the RGB images. Following [13], we model the human as a deformable implicit signed distance field and the background as a neural radiance field. This allows us to formulate a pixel-level RGB loss that compares color values obtained via composited neural rendering with the observed pixel value. We jointly optimize the neural body model and the SMPL poses, initialized with the output of the second stage. We experimentally show that this final stage results in temporally smooth results and accurate pose-to-image alignment.

We evaluate EMP on 21 sequences recorded with our MVS [8], the same system we use to register ground-truth SMPL shape parameters. With a pose accuracy of 2.3 cm positional and 10.6° angular error, our evaluations reveal that EMP is more accurate than what has been reported for 3DPW (2.6 cm, 12.1°) [59]. Also, our global SMPL root trajectories are accurate with an estimated error of 5.1 cm compared to our indoor MVS. Finally, we evaluate the performance of recent state-of-the-art camera-relative and global RGB-based pose estimators on EMDB. Our results show that EMDB is a new challenging dataset that will enable future local *and* global pose estimation research.

In summary, we contribute: 1. EMDB, to the best of our knowledge the first comprehensive dataset to provide accurate SMPL poses, shapes, and trajectories in an unrestricted, mobile, in-the-wild scenario. 2. EMP, the first method to fuse EM measurements with image data and camera poses. 3. Extensive evaluations of the accuracy of EMP as well as baseline results of state-of-the-art work when evaluating on EMDB. Data is available under https://ait.ethz.ch/emdb.

2. Related Work

Sensor-based Pose Estimation Modern inertial measurement units (IMUs) are an appealing sensor modality for human pose estimation because they are small and do not require line-of-sight. However, they only measure orientation directly. This lack of reliable positional information can be mitigated by using a large number of sensors [47] or by fusing IMU data with other modalities such as external cameras [3, 11, 35, 43, 44, 54, 59, 73], head-mounted cameras [14], LiDAR [9], or acoustic sensors [33, 58]. Research has attempted to reduce the required number of sensors, *e.g.* [5, 17, 59, 60], which requires costly optimizations [60], external cameras [59], or data-driven priors to establish the sensor-to-pose mapping [17, 19, 62, 66, 67] and deal with the under-constrained pose space. While such methods yield accurate local poses, IMUs are intrinsically limited in that their position estimates drift over time.

Addressing this challenge, EM-POSE [23] puts forth a novel method for body-worn pose estimation that relies on wireless electromagnetic (EM) field sensing to directly measure positional values. A learned optimization [49] formulation estimates accurate body pose and shape from EM inputs. However, [23] is limited to a small indoor capture space, requires external tracking of the root pose and is not aligned with image observations. In this work, we move beyond these limitations and present an EM-based capture system that is mobile, deployed to capture in-the-wild data, and produces high-quality pose-to-image alignment.

RGB-based Pose Estimation The 3D pose of a human is either represented as a skeleton of 3D joints [36, 38, 50, 75] or via parametric body models like SCAPE [1] and SMPL [34] for a more fine-grained representation. We note that almost the entire body of research estimates local (*i.e.*, camera-local) poses. In recent years, deep neural networks have driven significant advancements in estimating body model parameters directly from images or videos [12, 21, 22, 24, 25, 29, 30, 40, 51–53, 55, 57, 64, 70, 74]. In addition, researchers have combined the advantages of both optimization and regression to fit the SMPL body [26, 49]. Others have leveraged graph convolutional neural networks to effectively learn local vertex relations by building a graph structure based on the mesh topology of the parametric body models, *e.g*. [7, 31]. These methods propose transformer encoder architectures to learn the non-local relations between human body joints and mesh vertices via attention mechanisms. Recently, a few approaches have set out to estimate realistic global trajectories of humans and cameras from local human poses [28, 65, 68, 69]. We evaluate several of the above methods on our proposed dataset on the tasks of camera-relative and global human pose estimation.

Human Pose Datasets Commonly used datasets to evaluate 3D human pose estimation are H3.6M [18], MPI-INF-3DHP [37], HumanEva [48], and TotalCapture [20]. Although these datasets offer synchronized video and MoCap data, they are restricted to indoor settings with static backgrounds and limited variation in clothing and activities.

To address these limitations, [59] proposed a method that combines a single hand-held camera and a set of body-worn IMUs to estimate relatively accurate 3D poses, resulting in an in-the-wild dataset called 3DPW. Following this work, HPS [14] estimates 3D human pose with IMUs while localizing the person via a head-mounted camera within a prescanned 3D scene. To further address the issue of IMU drift, HSC4D [9] leverages LiDAR sensors for global localization. However, both HPS and HSC4D assume static scene scans and do not register global body pose in a third-person view. Moreover, they lack an evaluation of how accurate their pose estimates are. Another approach to outdoor performance capture with reduced equipment is to utilize one or multiple RGB-D cameras [2, 15, 16]. In these approaches, the quality of body pose registrations is limited by the cameras' line-of-sight, noisy depth measurements and the capture space is fixed. None of these works provide an estimate of their datasets' accuracy either. EgoBody [72] provides egocentric views and registered SMPL poses but is restricted to a fixed indoor space, requires up to 5 external RGB-D cameras and lacks evaluation of the data accuracy. Synthetic data has been suggested as a means to provide high-quality annotations [41, 57]. However, due to the reliance on static human scans and artificial backgrounds there is a distributional shift compared to real images.

With EMDB we provide the first dataset of 3D human pose and shape that is recorded in an unrestricted, mobile, in-the-wild setting and provides global camera and SMPL root trajectories. To gauge the expected accuracy of EMDB, we rigorously evaluate our method against ground-truth obtained on a multi-view volumetric capture system [8]. These evaluations reveal that EMDB is not only two times larger than 3DPW, but its annotations are also more accurate.

3. Overview

Our goal is to provide a dataset with i) accurate 3D body poses and ii) shapes alongside global trajectories of the iii) body's root and iv) the moving camera. This data is obtained from electromagnetic (EM) sensor measurements and RGB-D data streamed from a single hand-held iPhone. We first describe the capture setup and protocol in Sec. 4. Sec. 5 discusses our method, EMP, for the estimation of global SMPL parameters, summarized in Fig. 2. To gauge the accuracy of EMP, we evaluate it against ground-truth data recorded with a multi-view volumetric system (MVS, [8]). These evaluations are provided in Sec. 6. Finally, using EMP on newly captured in-the-wild sequences, we introduce the Electromagnetic Database of Global 3D Human Pose and Shape in the Wild, EMDB, in Sec. 7, where we also evaluate existing state-of-the-art methods on EMDB.

4. Capture Setup

4.1. Sensing Hardware

EM sensors measure their position $\mathbf{p}_s$ and orientation $\mathbf{R}_s$ w.r.t. a source that emits an electromagnetic field. We use the same wireless EM sensors as [23], which have an estimated accuracy of 1 cm positional and 2-3 degrees angular error. We mount the EM source on the lower back of a participant and arrange the sensors on the lower and upper extremities and the head and torso. For the detailed sensor placement we refer to the Supp. Mat. All sensor data is streamed wirelessly to a laptop for recording.

We record the subjects with a hand-held iPhone 13 Pro Max. The record3d app [45] is used to retrieve depth and the iPhone's 6D pose is estimated by Apple's ARKit. We synchronize the data streams via a hand clap which is easy to detect in the phone's audio and in the EM accelerations.

4.2. Body Calibration

Before we start recording, we first scan each participant in minimal clothing to obtain their ground-truth shape. To

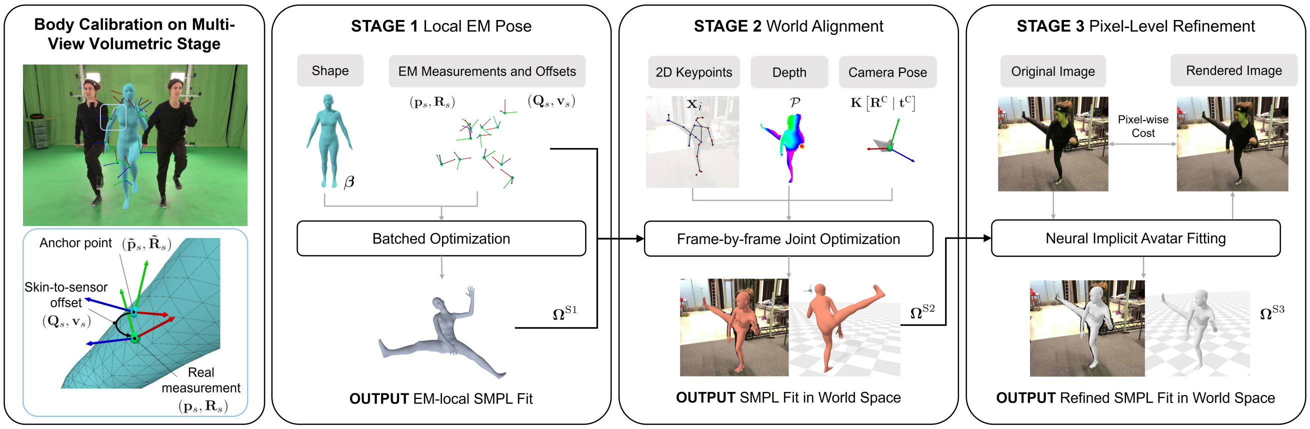

Figure 2: Method overview. We first scan a subject in minimal clothing with a multi-view volumetric capture system to obtain their reference shape parameters $\boldsymbol{\beta}$ and calibrate subject-specific skin-to-sensor offsets in regular clothing (left). We subsequently fit SMPL to in-the-wild data with a multi-stage optimization pipeline. Stage 1 fits SMPL to the EM measurements in EM-local space leveraging the calibrated body shape and skin-to-sensor offsets. Stage 2 aligns the local fit with the world, by jointly optimizing over 2D keypoints, depth, camera poses, EM measurements, and the output of stage 1. Stage 3 then refines the output of stage 2 by fitting a neural implicit body model with detailed geometry and appearance to the RGB images via a pixel-level supervision signal to boost smoothness and image-to-pose alignment.

this end, we leverage our MVS [8] and use the resulting surface scans and 53 RGB views to register the SMPL shape parameters $\boldsymbol{\beta}$. Details on the registration pipeline can be found in the Supp. Mat.

Subsequently, we mount the sensors and EM-source onto the participant under regular clothing (see Fig. 2, left). We then record a 3-second calibration sequence to determine subject-specific skin-to-sensor offsets. We first register SMPL to the calibration sequence and follow [23] to manually select anchor points on the SMPL mesh for every sensor s. An anchor point is parameterized via a position $\tilde{\mathbf{p}}_s$ and orientation $\tilde{\mathbf{R}}_s$. We then compute per-sensor offsets $\mathbf{o}_s = (\mathbf{Q}_s, \mathbf{v}_s)$ by minimizing an objective that equates the measured orientation $\mathbf{R}_s = \tilde{\mathbf{R}}_s \mathbf{Q}_s$ and the measured position $\mathbf{p}_s = \tilde{\mathbf{p}}_s + \tilde{\mathbf{R}}_s \mathbf{v}_s$ (see Fig. 2, left). For this to work, the sensor measurements must be spatially and temporally aligned with the MVS. We thus track the EM source with an Apriltag [27, 39, 61] and use an Atomos Ultrasync One timecode generator [56] for temporal alignment. More details are shown in the Supp. Mat. Note that this procedure must only be done once per sensor placement.

5. Method (EMP)

5.1. Notations and Preliminaries

The inputs to our method are EM sensor measurements $\mathbf{p}_s \in \mathbb{R}^3$ and $\mathbf{R}_s \in SO(3)$, skin-to-sensor offsets $\mathbf{o}_s = (\mathbf{Q}_s, \mathbf{v}_s)$, SMPL shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, RGB images $\mathbf{I} \in \mathbb{R}^{1920\times1440\times3}$, depth point clouds $\mathcal{P} = \{\mathbf{p}_i \mid \mathbf{p}_i \in \mathbb{R}^3\}$, camera extrinsics $\mathbf{C} = \left[\mathbf{R}^{C} \mid \mathbf{t}^{C}\right] \in \mathbb{R}^{3\times4}$ and intrinsics $\mathbf{K} \in \mathbb{R}^{3\times3}$. Note that the EM measurements are in EM-local space, *i.e.*, relative to the source worn on the lower back. From these input measurements, we aim to estimate the SMPL body pose parameters $\boldsymbol{\theta}_b \in \mathbb{R}^{69}$, the SMPL root orientation $\boldsymbol{\theta}_r \in \mathbb{R}^3$ and translation $\mathbf{t} \in \mathbb{R}^3$ in world coordinates such that they align with sensor measurements, images, and camera poses. We fix the world space to be the iPhone's coordinate frame. We summarize SMPL parameters as $\boldsymbol{\Omega} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_b, \mathbf{t}, \boldsymbol{\beta})$. Note that $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is not an optimization variable and is obtained a-priori (see Sec. 4.2). All quantities usually refer to a time step t, but we omit the time subscript for clarity unless necessary.

5.2. Multi-stage Optimization

As shown in Fig. 2, our method employs a multi-stage optimization procedure, which we detail in the following.

Stage 1: Local EM Pose For a given sequence, we start our optimization procedure by first finding SMPL parameters $\boldsymbol{\Omega}$ that best explain the EM measurements in EM-local space. We follow EM-POSE [23] and define a reconstruction cost function E_{rec} that measures how well the current SMPL fit matches the sensor measurements:

$$E_{\text{rec}} = \sum_{s=1}^{S} \lambda_{\text{p}} ||\mathbf{p}_s - \mathbf{p}_s^v(\mathcal{M}(\boldsymbol{\Omega}), \mathbf{o}_s)||_2^2 + \sum_{s=1}^{S} \lambda_{\text{r}} ||\mathbf{R}_s - \mathbf{R}_s^v(\mathcal{M}(\boldsymbol{\Omega}), \mathbf{o}_s)||_2^2 \, , \tag{1}$$

where we use the current SMPL mesh $\mathcal{M}(\mathbf{\Omega})$ and skin-to-sensor offsets $\mathbf{o}_s$ to compute virtual sensor positions $\mathbf{p}_s^v$ and orientations $\mathbf{R}_s^v$. In addition, we penalize impossible joint angles with a simple regularizer E_{bp}. The final optimization objective of the first stage is then $E_{\text{S1}} = \lambda_{\text{rec}} E_{\text{rec}} + \lambda_{\text{bp}} E_{\text{bp}}$. We use a batched optimization to minimize it over all T frames of the sequence. The output of stage 1 are the SMPL parameters in local EM space, $\mathbf{\Omega}^{\text{S1}}$ (see also Fig. 2).

Stage 2: World Alignment Due to accurate sensor data and our body calibration procedure, the $\mathbf{\Omega}^{\text{S1}}$ parameters are already of high quality (see Sec. 6.1). However, the EM space is not aligned with the world space. We align $\mathbf{\Omega}^{\text{S1}}$ with the world in a second optimization stage such that it fits the RGB-D observations and camera pose data. An overview of this stage is provided in Fig. 2.

This stage is guided by a 2D keypoint reprojection loss. Importantly, both 2D keypoints and depth are noisy and fitting to them naïvely can corrupt the initial estimates $\mathbf{\Omega}^{\text{S1}}$. Hence, we must trade-off accurate alignment of human and camera poses in world coordinates with the accuracy of the local pose. Although our trust in the EM fit $\mathbf{\Omega}^{\text{S1}}$ is high, we can still achieve improvements by fitting to RGB-D data for frames in which errors arise from sensor calibration or occasional measurement noise. Furthermore, the temporal alignment of EM and RGB-D data streams can be improved by fitting to the images. We model this trade-off as a joint optimization over all the input modalities.

We first define a 2D keypoint reprojection loss. We extract $N = 25$ 2D keypoints from Openpose [6] denoted by $\mathbf{x}_i \in \mathbb{R}^2$. The 3D keypoints $X(\mathbf{\Omega})$ are obtained via a linear regressor from the SMPL vertices. We then use the camera parameters to perspectively project the 3D keypoints (in homogenous coordinates), $\hat{\mathbf{x}}_i = \mathbf{K}\left[\mathbf{R}^{\text{C}} \mid \mathbf{t}^{\text{C}}\right] X(\mathbf{\Omega})_i$. The reprojection cost is then defined as

$$E_{\text{2D}} = \sum_{i=1}^{N} \mathbb{I}\left[c_i \geqslant \tau\right] \cdot \rho(\hat{\mathbf{x}}_i - \mathbf{x}_i) \tag{2}$$

where ρ is the Geman-McClure function [10], c_i is the confidence of the i-th keypoint as estimated by Openpose and $\mathbb{I}$ the indicator function. We set a high confidence threshold $\tau = 0.5$ in Eq. (2) to account for keypoint noise. Yet, even high confidence keypoints can be wrong. To ensure high quality of the ground-truth annotations provided in EMDB, we carefully review the keypoint predictions by Openpose and manually correct them for challenging samples.

We add two EM-related cost terms to this stage's optimization to further constrain the 3D pose. The first term is the EM reconstruction cost E_{rec} from Eq. (1). Note that here we only optimize the SMPL body pose $\boldsymbol{\theta}_b$ when computing the cost, denoted as E^*_{rec}. The second term is an additional prior on the body pose $\boldsymbol{\theta}_b^{\text{S1}}$ found in the first stage:

$$E_{\text{prior}} = ||\boldsymbol{\theta}_b^{\text{S1}} - \boldsymbol{\theta}_b||_2^2. \tag{3}$$

This E_{prior} formulation is similar to the one of HPS [14]. However, we found that the addition of E_{prior} alone is not sufficient and E^*_{rec} plays a crucial role (see Sec. 6.1).

Finally, we incorporate the iPhone's point clouds $\mathcal{P}$. Since the point clouds are noisy, they mostly serve as a regularizer for the translation $\mathbf{t}$ with the following term:

$$E_{\text{pcl}} = \frac{1}{|\mathcal{P}^h|} \sum_{\mathbf{p}_i \in \mathcal{P}_h} d(\mathbf{p}_i, \mathcal{M}(\mathbf{\Omega})). \tag{4}$$

Here, $d(\cdot)$ finds the closest triangle on the SMPL mesh $\mathcal{M}(\mathbf{\Omega})$ and then returns the squared distance to either the triangle's plane, edge, or vertex. $\mathcal{P}_h$ is a crop of $\mathcal{P}$, where the human is isolated via masks provided by RVM [32]. The final second stage objective is thus:

$$E_{\text{S2}} = \lambda_{\text{2D}} E_{\text{2D}} + \lambda_{\text{rec}} E^*_{\text{rec}} + \lambda_{\text{prior}} E_{\text{prior}} + \lambda_{\text{pcl}} E_{\text{pcl}} \tag{5}$$

We optimize this objective frame-by-frame and use the previous output as the initialization for the next frame. The output of this stage is $\mathbf{\Omega}^{\text{S2}}$ (see also Fig. 2). For the very first frame, we initialize $\mathbf{t}^{\text{S2}}$ as the mean of $\mathcal{P}_h$. All sequences start with a T-pose where the subject is facing the camera, so that it is easy to find an initial estimate of $\boldsymbol{\theta}_r^{\text{S2}}$.

Stage 3: Pixel-Level Refinement Stage 2 finds a good trade-off between accurate poses and global alignment (see Sec. 6.1). However, the jitter in the 2D keypoints causes temporally non-smooth estimates. Reducing the jitter by manually cleaning 2D keypoints is not viable. Instead, we add a third stage to EMP (see also Fig. 2) in which we follow recent developments in neural body modelling for in-the-wild videos. For every sequence, we fit a neural implicit model of clothed human shape and appearance to the RGB images by minimizing a dense pixel-level objective.

More specifically, we leverage Vid2Avatar (V2A [13]) to model the human in the scene as an implicit signed-distance field (SDF) representing surface geometry and a texture field, while the background is treated as a separate neural radiance field (NeRF++) [71]. The SDF is modelled in canonical space and deformed via SMPL parameters $\mathbf{\Omega}$ to pose the human. Then, given a ray $\mathbf{r} = (\mathbf{o}, \mathbf{v})$ whose origin $\mathbf{o}$ is the camera center and $\mathbf{v}$ its viewing direction, a color value $C(\mathbf{r})$ can be computed via differentiable neural rendering and is compared to the actual RGB value $\hat{C}(\mathbf{r})$ to formulate a self-supervised objective:

$$E_{\text{rgb}} = \frac{1}{|\mathcal{R}_t|} \sum_{\mathbf{r} \in \mathcal{R}_t} |C(\mathbf{r}) - \hat{C}(\mathbf{r})| \tag{6}$$

where $\mathcal{R}_t$ is the set of all rays that we shoot into the scene at frame t. Importantly, $C(\mathbf{r})$ depends on the SMPL poses

Method	MPJPE-PA [mm]	MPJAE-PA [deg]	Jitter [10m s^{-3}]
ROMP [51]	57.9 ± 23.6	19.8 ± 6.3	49.0 ± 10.6
HybrIK [29]	50.4 ± 22.3	19.0 ± 5.8	33.3 ± 7.1
Vid2Avatar [13]	50.2 ± 22.8	18.1 ± 6.2	38.7 ± 8.0
LGD [49]	61.1 ± 31.9	20.1 ± 8.0	68.9 ± 10.2
Stage 1	26.0 ± 8.6	10.9 ± 3.1	6.0 ± 2.9
Stage 2 (no E^*_{rec})	31.6 ± 14.1	12.7 ± 4.5	26.8 ± 3.7
Stage 2 (no E_{prior})	35.4 ± 14.2	11.6 ± 3.9	23.0 ± 3.3
Stage 2	$23.7 \pm \mathbf{7.5}$	$\mathbf{10.5} \pm \mathbf{3.0}$	21.7 ± 3.7
Stage 3 (after E_{S3})	23.5 ± 7.6	10.6 ± 3.1	12.7 ± 2.5
Stage 3 (EMP)	$\mathbf{23.4} \pm \mathbf{7.5}$	10.6 ± 3.1	$\mathbf{3.5} \pm \mathbf{1.0}$

Table 1: Comparison of EMP to existing RGB-based methods (top) and self-ablations (middle/bottom) on ground-truth data obtained with our multi-view capture system.

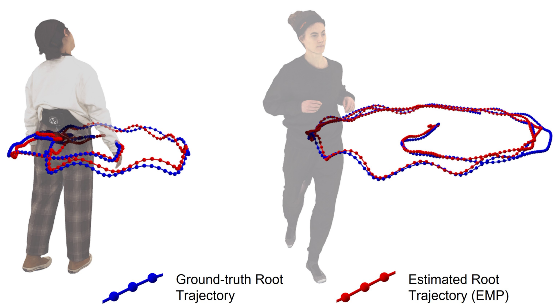

Figure 3: Evaluation of global trajectories on our MVS.

$\mathbf{\Omega}$ that are optimized jointly together with the parameters for the human and background fields. Along with E_{rgb}, the original formulation of V2A minimizes two other objectives: the Eikonal loss E_{eik} and a scene decomposition loss E_{dec} to disentangle the human from the background. For more details we refer the reader to [13]. We initialize the SMPL parameters $\mathbf{\Omega}$ with the outputs of the second stage $\mathbf{\Omega}^{\text{S2}}$ and add a pose regularization term $E_{\text{reg}} = ||\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{S2}}||_2^2$ (where $\boldsymbol{\theta} := [\boldsymbol{\theta}_r, \boldsymbol{\theta}_b]$) to encourage solutions to stay close to the initializations. The final third stage objective for a single time step is thus (omitting weights λ for brevity):

$$E_{\text{S3}} = E_{\text{rgb}}(\boldsymbol{\omega}_h, \boldsymbol{\omega}_b) + E_{\text{eik}}(\boldsymbol{\omega}_h) + E_{\text{dec}}(\boldsymbol{\omega}_h) + E_{\text{reg}}(\boldsymbol{\theta}), \quad (7)$$

where $\boldsymbol{\omega}_h$ summarizes the parameters for the human field, including SMPL pose parameters $\mathbf{\Omega}$, and $\boldsymbol{\omega}_b$ summarizes the weights of the background field. This objective is minimized over all T frames of the given sequence and produces outputs $\mathbf{\Omega}^{\text{S3}}$, which are noticeably less jittery (see Sec. 6.1).

6. Evaluation

6.1. Pose Accuracy

To estimate the accuracy of EMP we recorded a number of sequences with the same capture setup as we use for the in-the-wild sequences, but the motions are performed on our MVS [8] that is synchronized with the EM sensors and the iPhone. We use the surface scans and 53 high-resolution RGB views from this stage to procure SMPL ground-truth registrations (see Supp. Mat. for details), which we can then compare to the outputs of EMP to estimate its accuracy. We have recorded a total of 21 sequences (approx. 13k frames) distributed over all 10 participants for this evaluation. The respective ablation studies and comparisons to other methods are listed in Tab. 1.

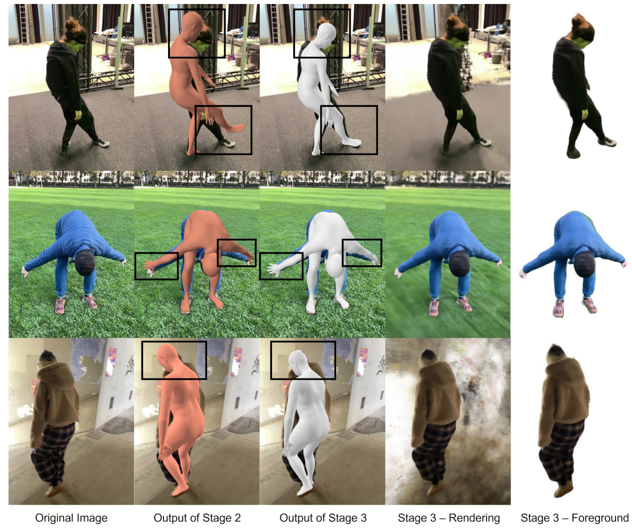

Figure 4: Effect of Stage 3. We visualize the output of stage 2 (second column) and the refined output of stage 3 (third column) showing improved pose-to-image alignment. The two right-most columns show the rendering of the entire scene and the separated human (foreground).

The closest related in-the-wild dataset to ours is 3DPW [59]. It is also the only other dataset that provides ground-truth evaluations of their method. As different sensor technologies are used, a direct comparison to their method is not feasible. Still, to allow for a comparison of the estimated accuracy, we compute and report the same metrics as [59], *i.e.*, the Procrustes-aligned mean per-joint positional and angular errors (MPJPE-PA, MPJAE-PA). To measure smoothness, we follow TransPose [67] and report their jitter metric. In addition we show qualitative comparisons to 3DPW with similar motions in the Supp. Mat.

Results: Tab. 1 allows to draw several conclusions. First, recent monocular methods - whether they use ground-truth bounding boxes (HybrIK [29]) or not (ROMP [51]) - are far below EMP's accuracy. Also V2A [13] suffers without good initial poses. LGD [49], which uses 2D keypoints in a hybrid optimization and outperforms SPIN [26] and Simplify [4] on 3DPW, underperforms compared to EMP. This highlights a clear need for sensor-based methods to procure high-quality 3D poses.

Second, Tab. 1 ablates the contributions of the multi-stage design of EMP. We observe that the first stage, which only fits to the EM measurements, already produces good results. Further, the joint optimization in our second stage finds a good trade-off and even improves the initial poses from the first stage via the addition of E^*_{rec} and E_{prior}. Lastly, the third stage only improves the pose marginally, but helps with smoothness and image alignment ("after E_{S3}" in Tab. 1). We perform a light smoothing pass as a post-processing step on the outputs of E_{S3}. We found that this further reduces jitter without breaking pose-to-image alignment. For a visualization of the effect of stage 3, as well as renderings of the neural implicit human model and the scene, please refer to Fig. 4. Note that naïvely smoothing the outputs of the second stage impacts the alignment negatively, which we show in the Supp. Mat.

6.2. Global Trajectories

iPhone Pose Accuracy We first compare the iPhone's self-localized poses using optical tracking with our MVS. To do so we rigidly attach an Apriltag [27, 39, 61] to the iPhone and move the pair around. An Apriltag of roughly 5 cm side length can be tracked with millimeter accuracy. To compare its pose to the iPhone's pose, we must compute an alignment, the details of which are reported in the Supp. Mat. After alignment, the difference between the iPhone and Apriltag trajectories on a 15 second sequence is 1.8 ± 0.9 cm and 0.4 ± 0.2 deg respectively.

Global SMPL Trajectories To evaluate the accuracy of the global trajectories, we asked half of our participants to move freely in the capture space while we track the iPhone with an Apriltag as above. This enables us to align the iPhone's and the MVS' tracking frames. For details, please refer to the Supp. Mat. After alignment, we compute the Euclidean distance between EMP's predicted trajectory and the ground-truth trajectory obtained on the stage. Over 5 sequences (approx. 3.9k frames) we found that EMP's trajectories are on average 5.1 ± 3.2 cm close to the ground-truth, which is low considering a capture space diameter of 2.5 meters (see Fig. 3 for a visualization).

To gauge the accuracy of the global trajectories in-the-wild, where we cannot track the iPhone, we asked some participants to return to the starting point at the end of the sequence. This allows us to compute a measure of drift for the in-the-wild sequences. For an indoor sequence of 81 meters, this error is 23.4 cm (or 0.3% of the total path length) and for an outdoor sequence of 112 meters length it is 73.0 cm (0.7%) respectively (see also Fig. 9 for a visualization).

Dataset	# number of:		Size	PA Accuracy		Global
	subj.	seqs.	[min.]	MPJPE	MPJAE	Traj.
3DPW [59]	7	60	29.3	2.6 cm	12.1 °	✗
EMDB (Ours)	**10**	**81**	**58.3**	**2.3** cm	**10.6** °	✓

Table 2: Comparison to in-the-wild datasets that provide evaluations of their accuracy. PA: Procrustes-aligned.

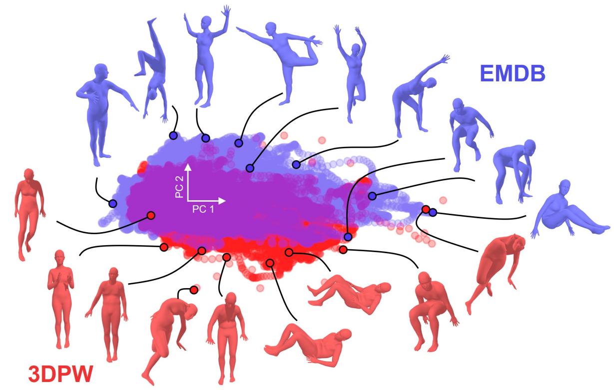

Figure 5: Scatter plot of first two principal components computed on 3DPW and EMDB in VPoser's [42] latent space and associated 3D poses for selected data points.

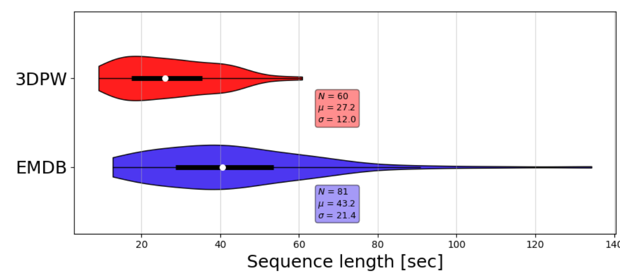

Figure 6: Distribution of sequence lengths in seconds in EMDB and 3DPW (thicker line from 1^{st} to 3^{rd} quartile).

7. EMDB

7.1. Dataset Overview

EMDB contains 10 participants (5 female, 5 male), who were recorded in a total of 81 sequences at 30 fps, resulting in 104,963 frames or 58.3 minutes of motion data. We plot the distribution of sequence lengths in Fig. 6. The ethnic distribution of participants in EMDB is: Middle Eastern (1), Asian (3), Caucasian (6). For a summary of statistics and comparison to other in-the-wild datasets that provide evaluations, please refer to Tab. 2. Of the 105k frames contained in EMDB, approx. 85% are recorded in-the-wild (indoors or outdoors) and the rest were recorded on our MVS. Please refer to the Supp. Mat. for detailed descriptions of every sequence as well as the distribution of body shapes.

Method	MPJPE ↓ [mm]	MPJPE-PA ↓ [mm]	MVE ↓ [mm]	MVE-PA ↓ [mm]	MPJAE ↓ [deg]	MPJAE-PA ↓ [deg]	Jitter ↓ [10m s^{-3}]
PyMAF [70]	131.1 ± 54.9	82.9 ± 38.2	160.0 ± 64.5	98.1 ± 44.4	28.5 ± 12.5	25.7 ± 10.1	81.8 ± 25.6
LGD [49]	115.8 ± 64.5	81.1 ± 51.1	140.6 ± 75.8	95.7 ± 56.8	25.2 ± 13.3	25.6 ± 15.3	73.0 ± 38.5
ROMP [51]	112.7 ± 48.0	75.2 ± **33.0**	134.9 ± 56.1	90.6 ± <u>38.4</u>	26.6 ± 10.4	24.0 ± <u>8.7</u>	71.3 ± 25.3
PARE [25]	113.9 ± 49.5	72.2 ± 33.9	133.2 ± 57.4	85.4 ± 39.1	24.7 ± **9.8**	<u>22.4</u> ± 8.8	75.1 ± 22.5
GLAMR [69]	107.8 ± 50.1	71.0 ± 36.6	128.2 ± 58.5	85.5 ± 40.9	25.5 ± 12.6	23.5 ± 11.4	67.4 ± 32.3
FastMETRO-L [7]	115.0 ± 95.1	72.7 ± 47.4	133.6 ± 109.7	86.0 ± 55.4	25.1 ± 16.0	22.9 ± 12.7	81.3 ± 38.7
CLIFF [30]	<u>103.1</u> ± **43.7**	68.8 ± 33.8	122.9 ± **49.5**	81.3 ± **37.9**	**23.1** ± <u>9.9</u>	**21.6** ± **8.6**	<u>55.5</u> ± **17.9**
FastMETRO-L* [7]	108.1 ± 52.9	<u>66.8</u> ± 36.6	**119.2** ± 59.7	<u>81.2</u> ± 43.9	n/a	n/a	185.9 ± 51.0
HybrIK [29]	**103.0** ± <u>44.3</u>	**65.6** ± <u>33.3</u>	<u>122.2</u> ± <u>50.5</u>	**80.4** ± 39.1	<u>24.5</u> ± 11.3	23.1 ± 11.1	**49.2** ± <u>18.5</u>

Table 3: Evaluations of state-of-the-art methods on EMDB 1. Ordered descendingly by MPJPE-PA. Best results in **bold**, second best <u>underlined</u>. FastMETRO-L*: version without SMPL regression head, *i.e.*, the MPJPE is only evaluated on 14 joints as dictated by its model architecture.

Further, to shed more light onto pose diversity of EMDB compared to our closest related work, 3DPW [59], we project all poses of both datasets into VPoser's [42] latent space, run PCA and plot the first two principal components in Fig. 5. We make several observations: i) EMDB covers a larger area than 3DPW. ii) The additional area is made up of complex and diverse poses. iii) The highlighted poses of 3DPW around the lower boundary lack diversity. iv) Outliers on 3DPW can be broken poses, while the closest EMDB pose is still valid (see right-most pose pair).

We provide visualizations of our dataset's quality in Fig. 7. The recording of this dataset has been approved by our institution's ethics committee. All subjects have participated voluntarily and gave written consent for the capture and the release of their data.

7.2. Baselines on EMDB

We evaluate two tasks on EMDB: camera-local 3D human pose estimation from monocular RGB images and the emerging task of global trajectory prediction. To this end we partition EMDB into two parts: EMDB 1, which consists of our most challenging sequences (17 sequences of a total of 24 117 frames), and EMDB 2 with 25 sequences (43 120 frames) featuring meaningful global trajectories.

Monocular RGB-based Pose Estimation We evaluate a total of 8 recent SOTA methods on EMDB 1. Please refer to Tab. 3 for an overview of the results. We follow the AGORA protocol [41] and compute the MPJPE and MVE metrics with both a Procrustes alignment (*-PA) and a hip-alignment pre-processing step. In addition, we follow sensor-based pose estimation work and report the joint angular error MPJAE and the jitter metric [67].

To provide a fair evaluation and comparison between baselines, we provide ground-truth bounding boxes for methods that accept them or tightly crop the image to the human and re-scale it to the resolution the method requires. Hence only ROMP [51] takes the input images as is. Also, we exclude the few frames where the human is entirely occluded. We use the HRNet version of HybrIK [29] – an improved variant of their originally published model. For FastMETRO [7] we use their biggest model (*-L) and evaluate both with and without the SMPL regression head. None of the methods are fed any knowledge about the camera and comparisons to the ground-truth are performed in camera-relative coordinates. We use the SMPL gender(s) that the respective method was trained with.

Results: Tab. 3 reveals HybrIK [29] as the best performer. Nonetheless, an MPJPE-PA error of > 65 mm suggests that there is a lot of room for improvement. As is noted in AGORA [41], we highlight that the MPJPE-PA is a very forgiving metric due to the Procrustes alignment that removes rotation, translation, and scale. We have noticed that a good MPJPE-PA does not always translate to visually pleasing results, a circumstance that the rather high jitter and MPJPE value for all baselines supports (see also the supp. video). Similarly we observe very high standard deviations, which is a metric that tends to have been neglected by common benchmarks. Furthermore, we notice high angular errors of $> 23°$ on average for all methods. These results and the fact that we used ground-truth bounding-boxes for all methods except ROMP, suggest that there is ample space for future research in this direction using EMDB.

We show selected results for each baseline in Fig. 7 and further highlight a common failure case in Fig. 8 where the baseline method fails to capture the lower arm rotations. Note that such a failure case is not accounted for by the MPJPE metric, which is why we also report angular errors.

Global Trajectory Estimation As a second task, we evaluate GLAMR [69] on EMDB 2. We use GLAMR's publicly available code to run and evaluate its performance. This protocol computes global MPJPE, MVE, and acceleration metrics on windows of 10 seconds length, where the

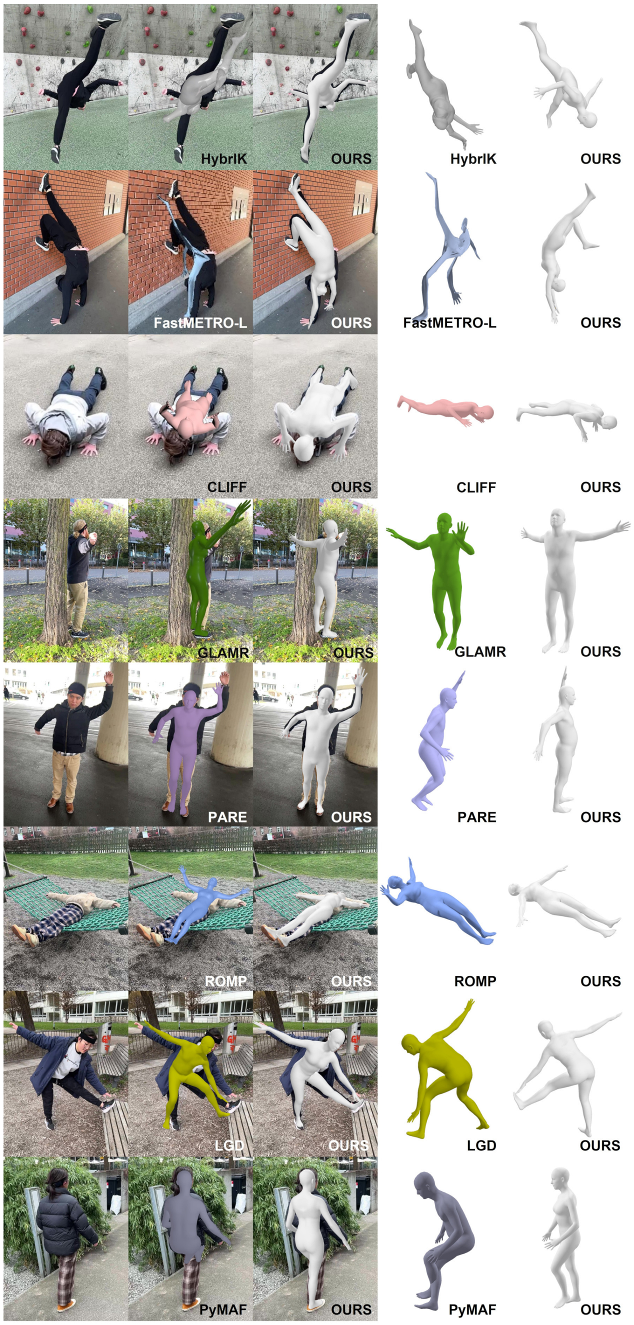

Figure 7: Example images and reference poses appearing in EMDB, alongside comparisons to the outputs of recent state-of-the-art RGB-based pose estimation methods.

beginning of each window is aligned to the ground-truth trajectory. We found that GLAMR achieves a G-MPJPE of 3 193 mm, a G-MVE of 3 203 mm and acceleration of 12.6 mm s^{-2}. We visualize one sequence in Fig. 9, where we observe that the GLAMR prediction drifts significantly from our provided trajectories. We believe EMDB will help to boost future method's performance on this task.

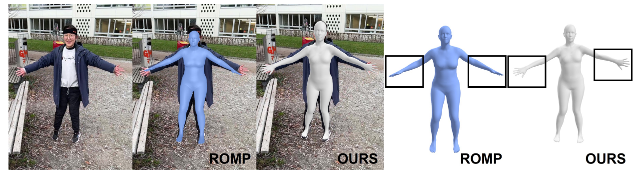

Figure 8: Common failure case where the baseline (here ROMP [51]) fails to capture the lower arm rotations.

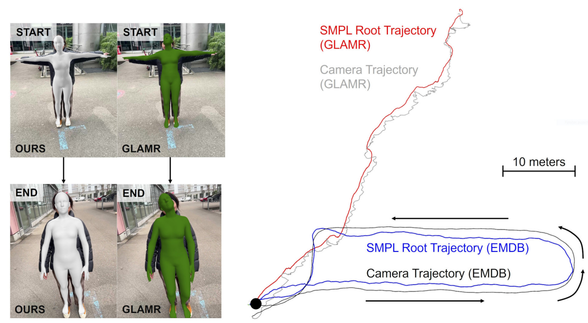

Figure 9: (Left) GLAMR [69] results projected into the camera at the start and end of a loop-closing sequence. (Right) GLAMR's global trajectories compared to ours.

8. Conclusion

Conclusion We present EMDB, the first comprehensive dataset to provide accurate SMPL poses, shapes and trajectories in an unrestricted, mobile, in-the-wild setting. Our results indicate a clear need for sensor-based performance capture to procure high-quality 3D human motion and push the boundaries of monocular RGB-based pose estimators.

Limitations EMDB does not contain multi-person sequences, because using multiple EM systems requires non-trivial changes to avoid cross-talk and interference between sensors. Furthermore, there are no sensors on the feet as indoor floors often contain metal beams that would disturb the readings. Lastly, the quality of our camera trajectories is upper-bounded by the quality of Apple's AR toolkit.

Acknowledgments We thank Robert Wang, Emre Aksan, Braden Copple, Kevin Harris, Mishael Herrmann, Mark Hogan, Stephen Olsen, Lingling Tao, Christopher Twigg, and Yi Zhao for their support. Thanks to Dean Bakker, Andrew Searle, and Stefan Walter for their help with our infrastructure. Thanks to Marek, developer of record3d, for his help with the app. Thanks to Laura Wülfroth and Deniz Yildiz for their assistance with capture. Thanks to Dario Mylonopoulos for his priceless work on aitviewer which we used extensively in this work. We are grateful to all our participants for their valued contribution to this research. Computations were carried out in part on the ETH Euler cluster.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005. 3

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3

[3] Gabriele Bleser, Gustaf Hendeby, and Markus Miezal. Using egocentric vision to achieve robust inertial body tracking under magnetic disturbances. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 103–109, 2011. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 6

[5] H. T. Butt, B. Taetz, M. Musahl, M. A. Sanchez, P. Murthy, and D. Stricker. Magnetometer robust deep human pose regression with uncertainty prediction using sparse body worn magnetic inertial measurement units. *IEEE Access*, 9:36657–36673, 2021. 2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 5

[7] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 3, 8

[8] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), jul 2015. 2, 3, 4, 6

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

[10] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 52(4):5–21, 1987. 5

[11] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *International Journal of Computer Vision*, 127:1–17, 09 2018. 2

[12] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 3

[13] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6

[14] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2, 3, 5

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

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

[17] 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)*, 37:185:1–185:15, Nov. 2018. 2

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 2, 3

[19] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery. 2

[20] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2, 3

[21] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 3

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[23] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[25] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference*

on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 1, 3, 8

[26] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 3, 6

[27] Maximilian Krogius, Acshi Haggenmiller, and Edwin Olson. Flexible layouts for fiducial tags. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2019. 4, 7

[28] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, 2022. 3

[29] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 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*, pages 3383–3393, 2021. 1, 3, 6, 8

[30] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 1, 3, 8

[31] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 3

[32] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance, 2021. 5

[33] Huajun Liu, Xiaolin Wei, Jinxiang Chai, Inwoo Ha, and Taehyun Rhee. Realtime human motion control with a small number of inertial sensors. In *Symposium on Interactive 3D Graphics and Games*, pages 133–140. ACM, 2011. 2

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3

[35] 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)*, pages 449–457, 2017. 2

[36] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 3

[37] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2, 3

[38] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 3

[39] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407. IEEE, May 2011. 4, 7

[40] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 3

[41] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3, 8

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 7, 8

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

[44] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2010. 2

[45] *record3d*, 2023. https://record3d.app/. 3

[46] 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

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

[48] L. Sigal, A.O. Balan, and M.J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal on Computer Vision (IJCV)*, 87(1):4–27, 2010. 3

[49] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 744–760. Springer, 2020. 2, 3, 6, 8

[50] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 3

[51] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 1, 3, 6, 8, 9

[52] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022.

[53] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association, 2017. 3

[54] 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

[55] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017. 3

[56] *Atomos Ultrasync One*, 2023. `https://www.atomos.com/accessories/ultrasync-one`. 4

[57] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017. 3

[58] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM Trans. Graph.*, 26(3):35–es, July 2007. 2

[59] 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 Conference on Computer Vision (ECCV)*, sep 2018. 2, 3, 6, 7, 8

[60] Timo von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs. In *Computer Graphics Forum*, volume 36, pages 349–360. Wiley Online Library, 2017. 2

[61] John Wang and Edwin Olson. AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2016. 4, 7

[62] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery. 2

[63] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^2Cap^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

[64] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019. 3

[65] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild, 2023. 1, 3

[66] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[67] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4), 08 2021. 2, 6, 8

[68] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021. 3

[69] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 8, 9

[70] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3, 8

[71] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields, 2020. 5

[72] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision (ECCV)*, Oct. 2022. 2, 3

[73] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *CVPR*, 2020. 2

[74] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019. 3

[75] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016. 3