Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging

Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz
Department of Computer Science, ETH Zürich, Switzerland

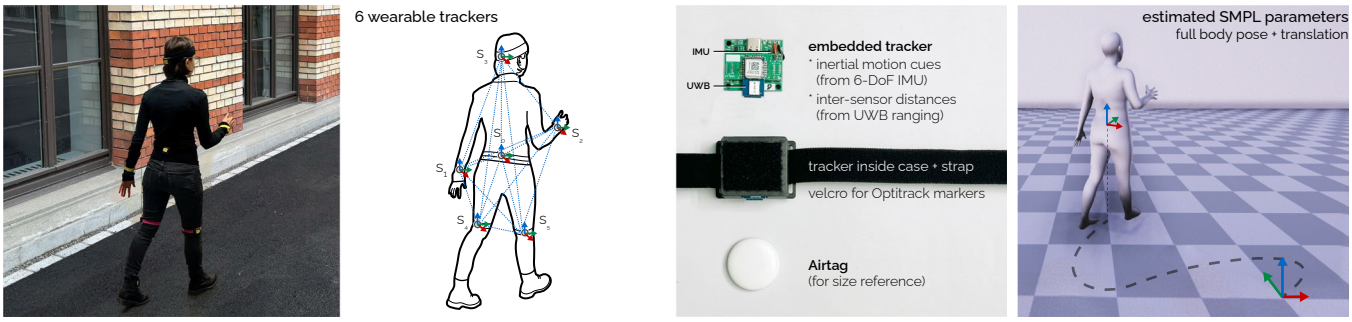

Figure 1: Our method *Ultra Inertial Poser* estimates 3D full body poses and global translation (i.e., SMPL parameters) from the inertial measurements on a sparse set of wearable sensors, augmented and stabilized by the estimated inter-sensor distances based on UWB ranging. Our lightweight standalone sensors stream *raw IMU signals,* from which we estimate each tracker's 3D state and fuse it with acceleration and inter-sensor distances in a graph-based machine learning model for pose estimation.

ABSTRACT

While camera-based capture systems remain the gold standard for recording human motion, learning-based tracking systems based on sparse wearable sensors are gaining popularity. Most commonly, they use inertial sensors, whose propensity for drift and jitter have so far limited tracking accuracy. In this paper, we propose *Ultra Inertial Poser*, a novel 3D full body pose estimation method that constrains drift and jitter in inertial tracking via inter-sensor distances. We estimate these distances across sparse sensor setups using a lightweight embedded tracker that augments inexpensive off-the-shelf 6D inertial measurement units with ultra-wideband radio-based ranging—dynamically and without the need for stationary reference anchors. Our method then fuses these inter-sensor distances with the 3D states estimated from each sensor. Our graph-based machine learning model processes the 3D states and distances to estimate a person's 3D full body pose and translation. To train our model, we synthesize inertial measurements and distance estimates from the motion capture database AMASS. For evaluation, we contribute a novel motion dataset of 10 participants who performed 25 motion types, captured by 6 wearable IMU+UWB trackers and an optical motion capture system, totaling 200 minutes of synchronized sensor data (UIP-DB). Our extensive experiments show state-of-the-art performance for our method over PIP and TIP, reducing position error from 13.62 to 10.65 cm (22% better) and lowering jitter from 1.56 to 0.055 km/s^3 (a reduction of 97%).

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07…$15.00
https://doi.org/10.1145/3641519.3657465

UIP code, UIP-DB dataset, and hardware design:
https://github.com/eth-siplab/UltraInertialPoser

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**; • **Hardware** → **Sensor devices and platforms**.

KEYWORDS

Human pose estimation; sparse tracking; IMU; UWB.

ACM Reference Format:
Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. 2024. Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657465

1 INTRODUCTION

Accurate and unrestricted motion tracking is fundamental across many domains and applications, such as animation, games, Augmented and Virtual Reality as well as fitness training and rehabilitation. High-quality motion capture systems rely on cameras, involving a tethered, comprehensive, and stationary setup [Optitrack 2023; Vicon 2023]. Compensating for the worn capture suits or markers these systems typically require, researchers have instead explored markerless motion capture from images [Chen and Ramanan 2017; Kanazawa et al. 2018] and videos [Habermann et al. 2019; Kocabas et al. 2020] from one [Habermann et al. 2020; Hu et al. 2021; Li et al. 2020b] or more cameras [De Aguiar et al. 2008].

Beyond stationary setups, tracking systems built around wearable sensors have leveraged miniaturized cameras in head-mounted

devices [Rhodin et al. 2016; Tome et al. 2019] or body-worn sensors [Li et al. 2020a; Shiratori et al. 2011; Yi et al. 2023; Zhang et al. 2020]. Such motion capture is mobile and requires less instrumentation for tracking, though at the cost of substantial motion artifacts and self-occlusion in sensor observations [Wang et al. 2021].

Substituting wearable cameras with even smaller motion sensors, dominantly inertial measurement units (IMU), has enabled less expensive setups but requires more comprehensive full-body sensor coverage, usually attached to motion capture suits to achieve comparable quality (e.g., 17–19 sensors for Xsens [Xsens 2024] or Noitom [Noitom 2024]). The multitude of IMUs is needed to compensate for the diminished motion cues, which commercial products stabilize with kinematic models to fuse cross-sensor observations.

To lessen the need for comprehensive sensor coverage, recent methods have proposed estimating full-body poses from only a sparse set of inertial sensors with learning-based techniques [Huang et al. 2018; Yi et al. 2022, 2021]. It is worth noting that the quality of their estimates thereby rely on the well-calibrated *absolute 3D state* provided by the IMU sensors, which has dominantly been from Xsens' proprietary units. Even so, without access to direct position observations and due to the nature of inertial sensing, current pose estimates suffer from overall drift and joint jitter as well as degrading accuracy when body motions are performed slowly.

In this paper, we propose *Ultra Inertial Poser*, a learning-based method for 3D full body pose estimation that fuses *raw* IMU readings with inter-sensor distances for stable predictions. Our method estimates these distances from ultra-wideband (UWB) ranging, which we add to a small, inexpensive, and untethered embedded sensor node (Figure 1). Our graph-based neural model fuses these distances with our 3D state estimations from the raw IMU signals to recover the SMPL parameters of body pose and global translation.

For training, we use AMASS [Mahmood et al. 2019] to synthesize IMU signals and distances and augment them with our empirically validated and collision-aware noise model. For evaluation, we introduce a novel motion capture dataset in which participants performed a long list of activities. Our method surpasses the current SOTA methods PIP [Yi et al. 2022] and TIP [Jiang et al. 2022b] in position accuracy (22% lower error) and jitter (97% reduction).

We make the following contributions in this paper:

(1) *Ultra Inertial Poser*, a novel method to incorporate distance constraints into a pose estimation framework for inertial measurements. Our graph-based neural network affords effective training using existing large motion capture datasets (e.g., AMASS), complemented by synthetic and noise-augmented distances.
(2) An embedded sensing platform with *inexpensive off-the-shelf components* for UWB-based ranging and IMU-based motion detection similar to those inside UWB-based item trackers. Our UWB communication protocol and distance estimation requires *no stationary nodes* and dynamically compensates for moments of occlusion between nodes as ranging crosses the human body. Compared to electromagnetic sensing, UWB ranging is lower power, less susceptible to interference, and can *scale* simply.
(3) Ultra Inertial Poser's pipeline is first to operate on the *raw signals* from IMU sensors (i.e., *just* acceleration, angular velocity), and UWB radios—without requiring proprietary components to obtain global 3D orientation as input (e.g., Xsens in prior work).
(4) *UIP-DB*, a novel motion dataset of 25 types of motion activities from 10 participants, including everyday movements as well as challenging motions for IMU-only tracking (e.g., slow transitions). UIP-DB contains synchronized 6-DoF IMU signals, UWB measurements and distances, as well as SMPL references based on a 20-camera Optitrack setup, totaling 200 minutes of data.

Taken together, Ultra Inertial Poser is a scalable approach for inexpensive full-body motion tracking using sparse wearable sensors. Our tracking approach is wireless and affords motion capture in the wild outside controlled indoor environments (Figure 11).

2 RELATED WORK

Ultra Inertial Poser is related to human body pose estimation using wearable sensors and positioning based on ultra-wideband ranging.

2.1 Motion Capture from Inertial Sensors

Inertial Measurement Units (IMUs) are popular for tracking motion due to their compact size, low power consumption, and affordability. They combine accelerometers and gyroscopes to track changes in acceleration and angular velocity, often supplemented by magnetometers to measure surrounding magnetic fields. However, IMUs cannot directly observe positions, and position estimates suffer from drift. Commercial systems address this by integrating data from 17–19 IMUs with biomechanical models (XSens [Xsens 2024]).

Pose estimation from sparse sets of IMUs has become viable using the large available motion capture (mocap) datasets to train learning-based models [Guzov et al. 2021; Huang et al. 2018; Mahmood et al. 2019; Mollyn et al. 2023; Streli et al. 2023; Trumble et al. 2017]. Early systems used offline optimization to fit the SMPL body model to data from six IMUs (e.g., SIP [Von Marcard et al. 2017]). Deep Inertial Poser's (DIP) bidirectional recurrent neural network maps IMU readings to local joint motions [Huang et al. 2018], which other methods further improved [Nagaraj et al. 2020]. More recent methods also use 6 IMUs and optimize both pose and translation estimation via ground contact points [Jiang et al. 2022b; Yi et al. 2021] and physical constraints [Yi et al. 2022]. With the advance of VR/AR applications, several methods further sparsified input requirements to just the upper body to estimate full-body pose from head and hand poses [Ahuja et al. 2021; Du et al. 2023; Jiang et al. 2023, 2022a; Yang et al. 2021; Zheng et al. 2023].

To combat the persistent challenges in estimating accurate joint angles and positions, researchers have revisited external [Pan et al. 2023; Von Marcard et al. 2018] or body-worn [Yi et al. 2023] cameras for visual-inertial tracking. Alternatively, EM-Pose [Kaufmann et al. 2021] measures the relative 3D offsets and orientations between joints using 6–12 custom electromagnetic (EM) field-based sensors.

Given the difficulties of reproducing custom EM sensing technology with commodity components, our work instead builds on previous approaches that estimate mere distances between wearable trackers (e.g., using ultrasonic sensors [Liu et al. 2011; Vlasic et al. 2007]). Our method obtains such pairwise distances from ultra wideband-based ranging and integrates them as constraints into a learning-based sparse inertial sensing pipeline. The small size, inexpensive nature, and low power consumption of UWB radios can thus complement IMUs as part of wearable and mobile devices.

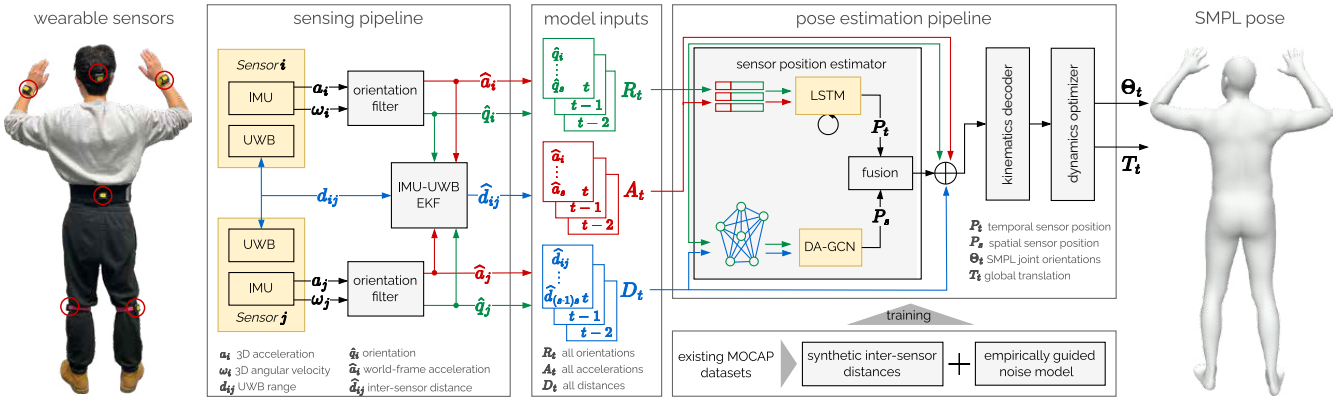

Figure 2: Overview of our method. The person is wearing 6 sensor nodes, each featuring an IMU and a UWB radio. Our method processes the raw data from each sensor to estimate sequences of global orientations, global accelerations, and inter-sensor distances. These serve as input into our learning-based pose estimation to predict leaf-joint angles as well as the global root orientation and translation as SMPL parameters.

2.2 Ultra-Wideband Ranging

Ultra-wideband (UWB) signals are characterized by their extensive bandwidth (larger than 500 MHz) and very short duration waveforms, typically in the order of a nanosecond [Sahinoglu et al. 2008]. This characteristic allows for accurate determination of the time of departure and arrival of signals, making UWB systems ideal for ranging applications through a variety protocols, including Two-Way Ranging (TWR)[IEEE 2007], Time Difference of Arrival (TDoA)[Tiemann and Wietfeld 2017; Ubisense 2023], and Angle of Arrival (AoA) [Ledergerber et al. 2019]. Additionally, UWB radios are compact and have low power consumption, making them suitable for a range of applications such as asset tracking both indoors and outdoors [Arun et al. 2022; Zhao et al. 2021], robotics localization [Cao et al. 2020; Lee et al. 2022; Ochoa-de Eribe-Landaberea et al. 2022; Zheng et al. 2022] and collaboration [Corrales et al. 2008; Hepp et al. 2016; Queralta et al. 2022], and augmented reality [Molina Martel et al. 2018].

A major challenge for accurate UWB ranging is the distortion due to obstacles in the line of sight between ranging devices. Prior methods have addressed these by analyzing the raw channel impulse response on UWB radios [Angarano et al. 2021; Barral et al. 2019; Tran et al. 2022], which is resource intensive and makes real-time operation on embedded platforms challenging. A more common approach is sensor fusion (e.g., particle or Extended Kalman Filters, EKF) to incorporate data from other sensors such as IMUs [Feng et al. 2020; Hol et al. 2009; Mueller et al. 2015; Ochoa-de Eribe-Landaberea et al. 2022], and cameras [Queralta et al. 2022; Xu et al. 2022]. We also build on an EKF-based UWB+IMU filter.

Previous work has focused on UWB positioning via trilateration with *fixed anchors* in the environment or using at least four rigidly coupled antennas on a node. In contrast, our method solely measures inter-sensor distances, thus requiring only a *single* UWB antenna and eliminating the need for stationary instrumentation.

3 METHOD

3.1 Problem Statement

Our goal is to estimate full-body pose from the sequential (raw) IMU observations and inter-sensor UWB ranging cues from a sparse set of S wearable sensors. In line with prior work, we examine this problem with 6 body-worn sensors on the pelvis, knees, wrists, and head. Given S sensor nodes with a 6-DoF IMU and a single UWB radio, we first aim to estimate each node's 3D orientation $R_t \in \mathbf{R}^{S\times 3}$, global accelerations $A_t \in \mathbf{R}^{S\times 3}$, as well as inter-sensor distances between all nodes $D_t \in \mathbf{R}^{S\times S}$. We then aim to predict the joint angle $\boldsymbol{\Theta}_t^j$ as well as the global translation T_t in SMPL human model parameters [Loper et al. 2015]. The pelvis sensor is the root of our body model, whereas the other sensors are leaf nodes.

Figure 2 shows an overview of our pipeline as detailed below.

3.2 Wearable Sensing Pipeline

3.2.1 Sensing Hardware. Our method is designed to estimate poses from commodity sensors, i.e., those commonly found in commercial devices. As such, we developed 6 wireless prototypes with a 6DoF IMU (LSM6DS) and a UWB radio (DWM1000) each, integrated into a 35×35 mm package (Figure 1 and Figure 10). An onboard microcontroller (NRF52840) runs custom firmware to sample the IMU and implement a UWB ranging protocol, streaming the results over BLE to a host computer. The host handles synchronization across the 6 devices and runs the rest of the sensing pipeline.

3.2.2 Obtaining Orientation and Acceleration Measurements. We sample each IMU at 100 Hz to obtain sensor-frame acceleration $\mathbf{a}_i$ and angular velocity ω_i signals from sensor i in S. We implement a VQF filter [Laidig and Seel 2023] to estimate gravity-compensated acceleration $\hat{\mathbf{a}}_i$ in the world frame and the absolute orientation quaternion $\hat{\mathbf{q}}_i$. At every time step, the estimates from each sensor are concatenated to form the orientation and global acceleration sequences R_t and A_t respectively. We perform a simple IMU calibration by estimating the gyroscope and accelerometer offsets as the wearer holds an initial T-pose.

3.2.3 UWB Ranging. The 6 devices implement a broadcast two-way ranging protocol to obtain the UWB distances $d_{i,j}$ between each sensor pair (i, j). In a ranging round, illustrated in Figure 8, Responder devices reply in turn to a request from an Initiator with the timestamps required to resolve time-of-flight τ_{ij} between devices.

$$\tau_{ij} = \frac{1}{2}(t_{i_{Received}} - t_{i_{Sent}} + t_{j_{Received}} - t_{j_{Sent}}) \tag{1}$$

Initiator and Responder roles are assigned at startup. In a network of 6 devices, this protocol allows us to measure all 15 pairwise distances at a rate of 25 Hz.

Individual UWB radios require an initial calibration to correct for hardware-specific variables, notably antenna delay which significantly affects ranging accuracy [Decawave 2017]. We calibrate all nodes at once, using a RANSAC regression to find an affine mapping of the raw ranges to the ground truth captured with a marker-based motion capture system. This calibration process is conducted once and applied consistently regardless of sensor placement or subject. The calibration mapping is however sensitive to temperature. with a variation of 2.15 mm/°C [Decawave 2018].

The parameters determined through this process remain valid given similar operating temperatures [Decawave 2014].

3.2.4 Estimating Filtered Inter-Sensor Distance Measurements. The raw UWB ranges exhibit increased noise when the body obstructs the direct line of sight between sensor pairs. To mitigate this, we fuse UWB ranges with the acceleration and orientation estimates for each sensor pair with an Extended Kalman Filter.

Our filter tracks the state space comprised of relative position, relative velocity, and relative orientation between a sensor pair (i, j), which we express as $\mathbf{x} = \begin{bmatrix} \boldsymbol{x}_{ij} & \boldsymbol{v}_{ij} & \boldsymbol{q}_{ij} \end{bmatrix}^\top$. We initialize the state assuming a starting position in T-pose. In the prediction step, our filter relies on dead-reckoning using the estimated acceleration and orientation as control inputs $\mathbf{u} = \begin{bmatrix} \hat{\boldsymbol{a}}_i & \hat{\boldsymbol{a}}_j & \hat{\boldsymbol{q}}_i & \hat{\boldsymbol{q}}_j \end{bmatrix}^\top$. The predicted state $\hat{\mathbf{x}}_t$ at discrete time step t becomes

$$\hat{\mathbf{x}}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_t) = \begin{bmatrix} \boldsymbol{x}_{ij_{t-1}} + \Delta T \boldsymbol{v}_{ij_{t-1}} + \frac{\Delta T^2}{2}(\hat{\boldsymbol{a}}_{j_t} - \hat{\boldsymbol{a}}_{i_t}) \\ \boldsymbol{v}_{ij_{t-1}} + \Delta T(\hat{\boldsymbol{a}}_{j_t} - \hat{\boldsymbol{a}}_{i_t}) \\ \hat{\boldsymbol{q}}_{i_t}^{-1}\hat{\boldsymbol{q}}_{j_t} \end{bmatrix} \tag{2}$$

where ΔT is the difference between consecutive time steps.

The noise in the prediction step is mainly introduced by the control input and therefore not simply additive. The VQF filter lacks explicit orientation covariance outputs. Instead, we treat we $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{q}}$ inputs as independent, and define the prediction step noise covariance $\boldsymbol{Q}_t$ as:

$$\boldsymbol{Q}_t = \boldsymbol{W}_t \Sigma_u \boldsymbol{W}_t^T \qquad \boldsymbol{W}_t = \frac{\partial f(\mathbf{x}_{t-1}, \mathbf{u}_t)}{\partial \mathbf{u}} \tag{3}$$

where Σ_u is the diagonal matrix of each of the input sensors' variance, which we determine from ground-truth. We introduce UWB ranges at the correction step. Before fusion, we remove raw UWB outliers outside a range of acceptable inter-sensor distances, to ignore measurements that are incompatible with human motion constraints. This range for each sensor pair is defined apriori assuming body heights of [1.5, 2 m] and constant across subjects. These ranges and their first derivative, form the measurement model $\mathbf{h}(\mathbf{x})$ defined in Eq. 4, and linearized as Eq. 5.

$$\mathbf{h}_t(\mathbf{x}) = \begin{bmatrix} d \\ v \end{bmatrix} = \begin{bmatrix} \|\boldsymbol{x}_{ij}\|_2 \\ \|\boldsymbol{v}_{ij}\|_2 \end{bmatrix} \tag{4}$$

$$H(\mathbf{x}) = \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\boldsymbol{x}_{ij}}{x} & \mathbf{0}_3 \\ \mathbf{0}_3 & \frac{\boldsymbol{v}_{ij}}{v} \end{bmatrix}_{2\times 6} \tag{5}$$

The measurement covariance matrix is simply the diagonal matrix of the distance and speed variances observed from raw distance measurements against ground-truth. It is different for each sensor pair, as some are LOS while others are not, but constant across subjects. Without direct observations, the relative position estimate $\boldsymbol{x}_{ij}$ is still prone to drift, however, its norm $\hat{\boldsymbol{d}}_{ij} = \|\boldsymbol{x}_{ij}\|_2$ provides a distance estimate more robust to NLOS distortions.

3.3 Pose Estimation with Distance Constraints

3.3.1 Model. The first module of our estimation pipeline is a sensor position estimator. It consists of two branches to capture the temporal and spatial information from input signals. In the first, we deploy an LSTM network [Hochreiter and Schmidhuber 1997] cells to estimate sensor positions relative to the root from acceleration and orientation estimates A_t and R_t. This module leverages temporal information from the sequence of inputs, and we denote its output as $\mathbf{P}_t \in \mathbf{R}^{S\times 3}$.

In parallel, we introduce a second branch to estimate relative sensor positions from pairwise distances using a Distance Attention Graph Convolutional Network (DA-GCN). As illustrated in Figure 3, we represent our system of wearable devices with a graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$, with vertices $\mathbf{V} \in S$ and edges $\mathbf{E}$. Each node i holds the orientation estimate $\mathbf{R}_t^i$, while edges represent inter-sensor distance estimates $\mathbf{D}_t$. Previous GCN implementations [Zhao et al. 2019] exploit a pre-defined affinity matrix with binary values to represent connectivity between joints of a skeleton model. We take it further by encoding inter-sensor distances on edge features, to capture additional correlation between nodes, to which we apply a modulated graph convolutional network [Zou and Tang 2021]. More specifically, we first normalize inter-sensor distance by the

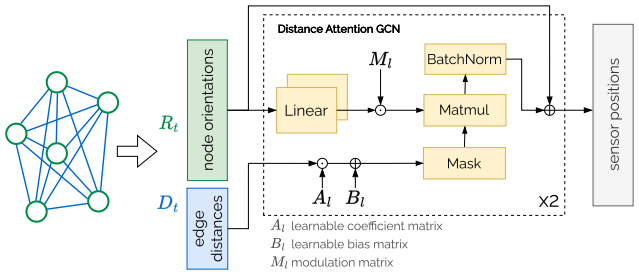

Figure 3: Architecture details for the Distance Attention Graph Convolutional Network (DA-GCN). Given node orientation R_t and edge distance D_t, DA-GCN estimates sensor positions conforming to the distance measurements. The model consists of a distance attention branch with learnable matrices A_l and B_l to represent the correlation between nodes and a weight-modulation branch M_l to learn disentangled transformation for different nodes.

distance between the head and pelvis to rule out differences in human shape. Building on the success of channel attention in image processing [Jiang et al. 2021; Woo et al. 2018], we then compute the correlation matrix as in Equation 6, where elements in matrix A_l and B_l are learnable.

$$\boldsymbol{C}_l = \boldsymbol{A_l} \odot \boldsymbol{D} \oplus \boldsymbol{B_l} \tag{6}$$

The correlation matrix is then applied as a weight to aggregate modulated node features, which can be denoted as:

$$\boldsymbol{H}_{l+1} = (\boldsymbol{M}_l \odot (\boldsymbol{W}_l \boldsymbol{H}_l))\boldsymbol{C_t} \tag{7}$$

, where H_l represents the node feature at layer l and M represents the modulation for weight W_l to learn disentangled transformation for node feature. By leveraging orientation and distance constraints, our DA-GCN outputs spatial sensor position estimation $\mathbf{P}_s \in \mathbf{R}^{S\times 3}$.

Similarly to TransPose [Yi et al. 2021], we use linear interpolation to fuse the two relative sensor position estimates. We compute a weighted sum of $\mathbf{P}_t$ and $\mathbf{P}_s$ based on the magnitude of acceleration estimates in A_t. This is driven by the fact that in scenarios with little movement IMUs have a low signal-to-noise ratio, and distinguishing between poses solely from inertial data becomes challenging.

Therefore, the fusion algorithm can be formulated as:

$$\mathbf{P}^i = \begin{cases} \mathbf{P}_s^i & \|\hat{\mathbf{a}}_i\| \le \underline{\mathbf{a}} \\ \frac{\|\hat{\mathbf{a}}_i\| - \underline{\mathbf{a}}}{\overline{\mathbf{a}} - \underline{\mathbf{a}}} \mathbf{P}_t^i + \frac{\overline{\mathbf{a}} - \|\hat{\mathbf{a}}_i\|}{\overline{\mathbf{a}} - \underline{\mathbf{a}}} \mathbf{P}_s^i & \underline{\mathbf{a}} < \|\hat{\mathbf{a}}_i\| < \overline{\mathbf{a}} \\ \mathbf{P}_t^i & \|\hat{\mathbf{a}}_i\| \ge \overline{\mathbf{a}} \end{cases} \tag{8}$$

We set the thresholds $\overline{\mathbf{a}}$ and $\underline{\mathbf{a}}$ to 8.0 and 2.0 m/s^2, respectively, which we determined via grid search to minimize position error.

This fusion approach accommodates different update rates between the inertial (100 Hz) and distance (25 Hz) measurements, eliminating the need for resampling or interpolation.

The rest of the pipeline processes the combined position estimates, acceleration, orientation, and distance measurements. These inputs are passed to a kinematics decoder, which estimates local orientation, local velocity, and foot contact, followed by a dynamics optimizer, based on previous work [Yi et al. 2022]. The pipeline outputs SMPL pose parameters Θ_t global translation T_t.

3.3.2 Loss Functions. To supervise our sensor position estimator, we design a loss function that incorporates distance constraints. From the estimated sensor positions $\mathbf{P}_s$ or $\mathbf{P}_t$, which we denote as $\widehat{\mathbf{P}}$, we compute inter-sensor distances $d(\widehat{\mathbf{P}}) \in \mathbf{R}^{S\times S}$ and the L1 reconstruction loss $\|d(\widehat{\mathbf{P}}) - \mathbf{D}\|_1$. We also pose additional constraints on its direction by computing the cosine similarity between estimated sensor positions and nearest joint positions. We can then formulate the full distance-aware loss as:

$$\mathbf{L}_d = \frac{\widehat{\mathbf{P}} \cdot \widetilde{\mathbf{P}}}{\|\widehat{\mathbf{P}}\|\|\widetilde{\mathbf{P}}\|} + \lambda \|d(\widehat{\mathbf{P}}) - \mathbf{D}\|_1 \tag{9}$$

where $\widetilde{\mathbf{P}}$ is the approximate sensor position from nearest joints. We select λ as 0.01 and leverage this distance aware loss as the objective function for sensor position estimation.

For other modules, we apply L2 loss for the rotation estimator, binary cross entropy loss for the contact estimator, and accumulated loss [Yi et al. 2021] for the velocity estimator.

3.3.3 Data Synthesis. Because training our pose estimation pipeline requires a large-scale dataset with IMU+UWB measurements and SMPL references, we first appropriate AMASS as an existing motion capture dataset [Mahmood et al. 2019] to synthesize sensor data.

Similarly to previous research [Yi et al. 2021], we synthesize global acceleration and orientation data by placing virtual sensors on SMPL mesh vertices. We deduce the rotation matrix from SMPL joint rotations and derive world-frame acceleration vectors using finite differences in vertex positions,

$$\mathbf{a}_t^i = \frac{n\mathbf{p}_{t+n}^i - 2\mathbf{p}_t^i + \mathbf{p}_{t-n}^i}{(n\Delta t)^2} \tag{10}$$

where p_t^i is the position of vertices i at time step t. To synthesize UWB data, we first compute the Euclidian distance from the virtual sensors' positions and then apply a time-varying noise model that reflects observations on real data. We conducted NLOS noise modeling experiments, in which we varied distances and obstacle sizes between sensors. We observed a noise standard deviation similar to LOS for obstacle-to-distance ratio smaller than 0.2, and larger for ratios above 0.2. As such, we model noise as dynamic Gaussian noise with standard deviation $\sigma(t)_{ij}$ given by:

$$\sigma(t)_{ij} = \begin{cases} \sigma_0 & c(t)_{ij} < 0.2 \\ \sigma_1 & c(t)_{ij} \ge 0.2 \end{cases} \tag{11}$$

where $c(t)_{ij}$ is the proportion of occlusion on the line between devices i and j. To compute this occlusion, we leverage the implicit human body model COAP [Mihajlovic et al. 2022] and query sampled points along the line segment at a constant resolution between devices. The values of σ_0 and σ_1 are determined empirically, by running our sensing pipeline with two devices first in LOS ($\sigma_0 = 0.051$ m) then in NLOS with a person as obstacle ($\sigma_1 = 0.083$ m).

3.3.4 Training Details. We adopt the Adam solver [Kingma and Ba 2015] with batch size 256 to optimize the parameters of model. We train each module separately with the learning rate at 1×10^{-3} and decays by a factor of 0.33 every 20 epochs. We train our model with PyTorch on one NVIDIA GeForce GTX 3090 GPU. It takes about 4 hours in total to train our model.

4 DATASET

To evaluate our method and to spur further work, we collected a motion capture dataset of a large variety of movement types, capturing the raw IMU+UWB measurements of our tracker nodes alongside those from an optical tracking system. We recorded more than 25 different movements from 10 participants (6 male, 4 female), with heights between 155 cm and 187 cm.

Our recording study comprised two sessions. The first focused on everyday movements that participants performed standing up, such as walking, jumping, squats, etc. The second session involves interactions with a chair and movements in various seated positions. For each participant, we recorded two takes of each session and a final freestyle session, where they acted as desired for 4–6 min.

As shown in Figure 10, participants wore a suit with 57 reflective markers and 6 of our wearable devices on the back of the head, lower back, wrists, and knees. The wearable sensors continuously

Table 1: Comparison with state-of-the-art methods on existing datasets augmented with simulated, ideal inter-sensor distances

	original TIP/PIP training set (TotalCapture ground truth included)									AMASS *without* TotalCapture								
Dataset	DanceDB			DIP-IMU			TotalCapture			DanceDB			DIP-IMU			TotalCapture		
Method	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter
TIP	18.705	8.500	1.438	17.068	5.822	0.882	11.361	5.145	0.751	18.740	8.501	2.252	17.011	5.710	1.162	13.588	5.967	1.015
PIP	20.007	8.877	2.234	15.020	6.020	0.240	12.930	7.025	0.204	20.007	8.997	2.334	15.977	6.209	0.260	15.930	7.046	0.292
with synthesized *perfect* distances:																		
TIP-D	16.243	7.624	2.220	15.910	5.442	0.863	10.762	5.142	0.722	16.225	7.648	2.221	15.914	5.259	1.128	12.183	5.507	1.004
PIP-D	16.205	8.042	2.124	13.762	5.336	0.281	11.408	5.560	0.201	16.815	8.063	2.166	13.788	5.355	0.297	13.161	6.317	0.273
UIP (ours)	**15.280**	**7.450**	**0.430**	**13.200**	**5.050**	**0.240**	**10.707**	**5.108**	**0.206**	**15.321**	**7.612**	**0.430**	**13.210**	**5.053**	**0.248**	**11.321**	**5.490**	**0.257**

Table 2: Comparisons on UIP-DB with acceleration, orientation and inter-sensor distances from real off-the-shelf sensors.

UIP-DB	overall			$\|\mathbf{a}\| \leq 1.0\,\mathrm{m/s^2}$			$\|\mathbf{a}\| > 1.0\,\mathrm{m/s^2}$		
Method	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter	SIP Err	Pos Err	Jitter
TIP	33.01	14.82	1.86	36.90	16.55	1.79	28.83	12.95	1.95
PIP	30.47	13.62	1.57	36.18	15.87	1.33	23.18	10.80	1.82
with distances *estimated* from UWB recordings:									
TIP-D	30.34	13.96	1.84	33.11	15.52	1.79	27.17	12.19	1.88
PIP-D	30.33	13.27	1.39	35.47	15.25	1.29	24.13	10.80	1.51
UIP (ours)	24.12	10.65	**0.05**	24.72	11.84	**0.051**	22.64	10.02	**0.06**
UIP finetuned	**23.85**	**10.55**	0.08	**24.51**	**11.18**	0.078	**22.43**	**9.79**	0.08

Table 3: Results of our ablation studies.

Method	SIP Err	Pos Err	Jitter
UIP (ours)	**24.115**	**10.648**	**0.055**
w/o inter-sensor distances	31.008	13.749	1.556
w/o IMU	27.005	14.086	0.098
w/o DA-GCN	27.529	12.136	0.092
w/o UWB noise	27.006	12.440	2.125
w/ Gaussian UWB noise	26.068	12.159	2.005
w/o Distance-guided loss	25.830	11.777	1.966

collected raw 3D accelerometer, 3D gyroscope, and 3D magnetometer readings at 100 Hz as well as raw UWB ranges at 25 Hz. We extended the recordings with the filtered estimates of global acceleration, orientation, and inter-sensor distances, estimated using our processing pipeline as described above.

We obtained ground truth for the poses of wearable sensors and body joints from a 20-camera Optitrack system across a capture area of 4×5 m. To extract SMPL references and global translation, we labeled the point cloud from the marker suit with SOMA [Ghorbani and Black 2021], and then used MoSh++ [Mahmood et al. 2019]. Since each of our trackers rigidly mounted a reflective marker, our dataset also includes their reference positions and orientations.

Each recording session started and ended with a T-pose to calibrate IMU readings and initialize the inter-sensor distances EKF. A jump then synchronized all sensors with the extracted SMPL poses.

5 EXPERIMENTS

We analyze our pipeline's accuracy, visually and quantitatively compare it to existing baselines, and discuss ablation studies.

5.1 Sensing Pipeline Evaluation

From the ground-truth sensor orientations and positions recorded in our dataset, we derive reference inter-sensor distances and accelerations. We use these values to evaluate our sensing pipeline and quantify the error it introduces at the input of pose estimation.

Our orientation estimate drifts at a rate of 3.21°/min, 0.42°/min and 0.31°/min for roll, pitch, and yaw angles respectively, in line with results obtained from similar commodity IMUs in previous work [Laidig and Seel 2023]. As this estimate is also used to transform sensor-frame acceleration into the global frame and filter UWB ranges, our dataset provides less accurate model inputs than those in existing datasets [Huang et al. 2018; Trumble et al. 2017] that rely on proprietary drift-free estimators in high-end sensors.

Figure 4 shows the distribution of the RMSE of the pairwise distances estimated by our filtering pipeline. The average error is largest for wrist–wrist and head–knees sensor pairs, as the body frequently occludes the line of sight between sensors.

5.2 Pose Estimation Evaluation

5.2.1 Evaluation protocol. We compare our approach against state-of-the-art IMU-based pose estimation methods: PIP [Yi et al. 2022] and TIP [Jiang et al. 2022b]. For fair comparison, we also augment both TIP and PIP by concatenating pairwise distances to their original inputs and retrain them using publicly available resources.

Our first experiment shows the benefits of adding pairwise distances, evaluating all methods on existing datasets. We form a test set from DIP-IMU [Huang et al. 2018], TotalCapture [Trumble et al. 2017], and DanceDB—held out from the AMASS training set. The acceleration and orientation measurements in TotalCapture and DIP-IMU were captured from XSens suits, while DanceDB has synthetic values. We add ideal synthetic inter-sensor distances to our test set to study their contribution independent of noise.

We first train all methods on AMASS, as originally done in TIP and PIP, with synthetic acceleration, orientation, and inter-sensor distances with our time-varying noise model. However, because this training set includes synthesized sensor readings from TotalCapture's ground truth, we additionally run these experiments with models trained on AMASS *without* TotalCapture.

A second experiment evaluates performance on our collected dataset UIP-DB, examining results both overall and for a mean acceleration split: slow movements (mean acceleration below 1

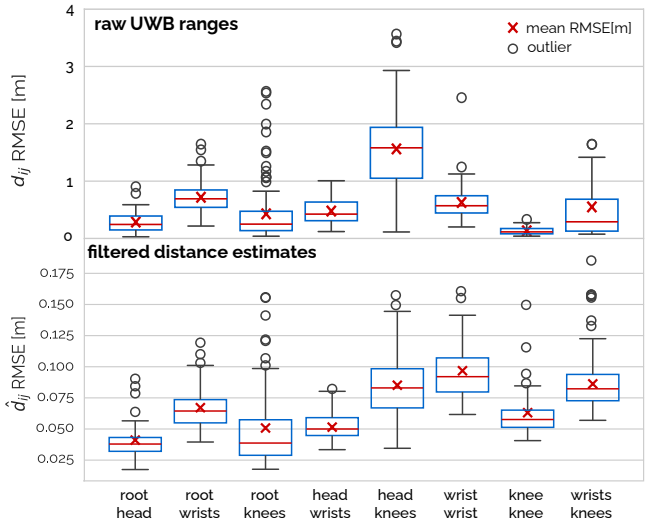

Figure 4: Inter-sensor distance RMSE in our dataset.

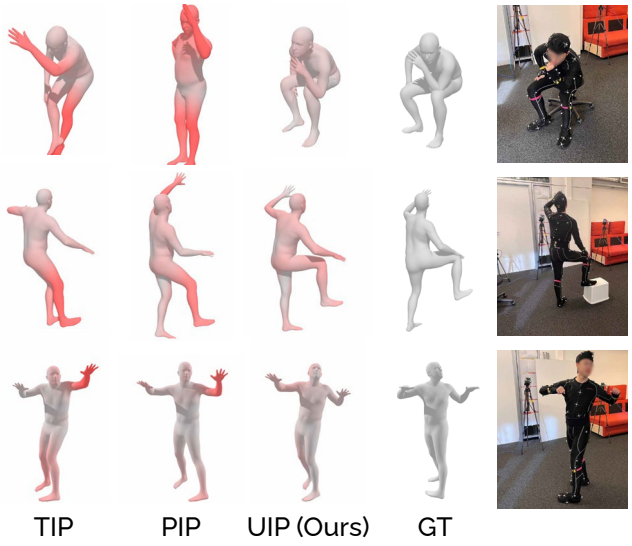

Figure 5: Qualitative comparison of pose estimates from the recordings in our dataset. Figure 6 shows more examples.

m/s²) and fast movements (mean acceleration above 1 m/s²). All inputs are from actual measurements processed by our pipeline.

5.2.2 Metrics. We evaluate model performance on three metrics. SIP Error measures the mean orientation error of the upper arms and legs in the global space in degrees. Joint position error measures the mean joint distance between the reconstructed and ground-truth joints with both root position and orientation aligned in cm. Jitter is the mean jerk (time derivative of acceleration) of all body joints in the global space in km/s^3, reflecting motion smoothness.

5.2.3 Results. Table 1 lists the evaluation results on DanceDB, TotalCapture, and DIP-IMU datasets with perfect synthetic inter-sensor distances. Table 2 has the results on our captured dataset.

Methods with distance constraints perform consistently better than inertial-only pipelines, with the most significant improvements in the SIP angle error. Within distance-augmented methods, our approach shows a moderate advantage when evaluated on existing datasets. However, its performance is markedly better on real data from our dataset, with 20% lower SIP and position errors, and an 80% reduction in jitter compared to distance-augmented PIP.

This disparity is due to two main factors. First, our evaluation uses *inexpensive off-the-shelf* sensors, such that the acceleration, orientation, and inter-sensor distance estimates from our dataset exhibit more noise and drift than synthetic datasets or those from high-end sensors. This also partially explains the larger error obtained by TIP and PIP on our dataset. In contrast, our fusion pipeline and proposed loss function actively enforces distance constraints while adaptively introducing constraints from inertial data, showing better results in all presented datasets. Figure 9 compares how varying levels of noise in the distance estimate influence SIP error.

Second, our dataset contains many sequences that are difficult to disambiguate from inertial data, such as motions with low speed. Table 2 shows that TIP and PIP—based solely on IMUs—perform significantly worse for small accelerations, whereas performance improves for quick motions. Our method effectively captures sitting poses and interactions with objects (Figure 5). Figure 7 plots the SIP error over a long sequence while a participant alternated between standing and sitting. Our method consistently tracks these state changes, whereas TIP and PIP incorrectly revert to a standing pose.

5.2.4 Ablations. We perform ablation studies on our dataset to understand the impact of each of the components we propose in our method, with results summarized in Table 3. We first assess the individual contribution of each of the inertial and distance inputs to the pose estimation pipeline, by excluding them in turn. We observe that distance constraints play a more important role in reducing SIP angle error and jitter, while inertial inputs have a more significant impact on the joint position error metric.

Next, we determine the effectiveness of our DA-GCN module. Our ablation discards DA-GCN and directly concatenates inter-sensor distances to the inertial data as input to the LSTM. This variant only considers temporal information and ignores the spatial correlation between nodes, leading to higher SIP and position errors. Investigating the effect of our collision-aware UWB noise model for training, we first train with ideal synthesized distances, then with simple additive Gaussian noise ($\sigma = 0.12$). SIP error and jitter metrics are impacted most by this ablation, showing the importance of a representative noise model in training our method.

Our final ablation study discards the pairwise distance reconstruction term from our loss function and substitutes cosine similarity with an L2 loss. This affects jitter most, indicating that the loss primarily contributes to stabilizing our predictions.

6 LIMITATIONS AND FUTURE WORK

By incorporating pairwise distances with inertial measurements, our system adds a requirement for sensing input that is UWB. Our results indicate that adding UWB to future motion sensors is worthwhile, though at the cost of requiring calibration across operating

temperatures. UWB is becoming increasingly available as part of location trackers (e.g., Airtags, Tile Ultra, Samsung SmartTag+).

Similar to other learning-based methods, our graph-based pipeline has limits predicting out-of-distribution poses. Our evaluation assumed a flat terrain and covered simple interaction with objects in the environment. A more robust method will require a dataset that captures a wider variety of motions and interactions.

Our method marginally improves global translation estimates, with a 0.318 m and 0.422 m error for ground-truth root movements of 2 m and 5 m, respectively. This is close to PIP's 0.333 m and 0.440 m error. While distances constrain cross-joint motion, global translation benefits from better pose prediction. Distance constraint-based translation estimation could be interesting in future work.

Finally, our state and distance estimates from real data are error-prone, whereas our evaluation on synthetic datasets with drift-free orientations and perfect distances shows our method's potential.

7 CONCLUSION

We have proposed a novel method for sparse, wearable sensor-based full-body pose tracking that is independent of visual input. It leverages a novel source of tracking input: inter-sensor distances estimated from UWB-based ranging, which we use to stabilize the raw signals from IMUs. We have demonstrated that based on six tracking nodes, our method affords training on existing motion capture datasets with simulated distances to robustly estimate full-body poses. Our tracking nodes integrate off-the-shelf components found in emerging UWB location trackers, and our pose estimation operates without the need for 3D state estimation from proprietary sensors or tracking systems, making Ultra Inertial Poser a scalable approach for sparse human motion tracking in the wild.

ACKNOWLEDGMENTS

We sincerely thank Xintong Liu and Max Moebus for their support and Paul Streli for feedback. We also thank all the participants of our data capture for their time.

REFERENCES

Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. 2021. CoolMoves: User Motion Accentuation in Virtual Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–23.

Simone Angarano, Vittorio Mazzia, Francesco Salvetti, Giovanni Fantin, and Marcello Chiaberge. 2021. Robust ultra-wideband range error mitigation with deep learning at the edge. *Engineering Applications of Artificial Intelligence* 102 (2021), 104278. https://doi.org/10.1016/j.engappai.2021.104278

Aditya Arun, Tyler Chang, Yizheng Yu, Roshan Ayyalasomayajula, and Dinesh Bharadia. 2022. Real-Time Low-Latency Tracking for UWB Tags. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services* (Portland, Oregon) *(MobiSys '22)*. Association for Computing Machinery, New York, NY, USA, 611–612. https://doi.org/10.1145/3498361.3538658

Valentín Barral, Carlos J. Escudero, José A. García-Naya, and Roberto Maneiro-Catoira. 2019. NLOS Identification and Mitigation Using Low-Cost UWB Devices. *Sensors* 19, 16 (2019). https://doi.org/10.3390/s19163464

Yanjun Cao, Chenhao Yang, Rui Li, Alois Knoll, and Giovanni Beltrame. 2020. Accurate position tracking with a single UWB anchor. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 2344–2350. https://doi.org/10.1109/ICRA40945.2020.9197345

Ching-Hang Chen and Deva Ramanan. 2017. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7035–7043.

J. A. Corrales, F. A. Candelas, and F. Torres. 2008. Hybrid tracking of human operators using IMU/UWB data fusion by a Kalman filter. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 193–200. https://doi.org/10.1145/1349822.1349848

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers*. 1–10.

Decawave. 2014. APS011 Application Note, Sources of error in DW1000 based two-way ranging (TWR) schemes.

Decawave. 2017. How To Use , Configure and Program the DW1000 UWB.

Decawave. 2018. APS014 Application Note, Antenna delay calibration of DW1000 based products and systems.

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. 2023. Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model. In *CVPR*.

Daquan Feng, Chunqi Wang, Chunlong He, Yuan Zhuang, and Xiang-Gen Xia. 2020. Kalman-Filter-Based Integration of IMU and UWB for High-Accuracy Indoor Positioning and Navigation. *IEEE Internet of Things Journal* 7, 4 (2020), 3133–3146. https://doi.org/10.1109/JIOT.2020.2965115

Nima Ghorbani and Michael J. Black. 2021. SOMA: Solving Optical Marker-Based MoCap Automatically. In *Proc. International Conference on Computer Vision (ICCV)*. 11117–11126.

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

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2019. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)* 38, 2 (2019), 1–17.

Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. 2020. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5052–5063.

Benjamin Hepp, Tobias Nägeli, and Otmar Hilliges. 2016. Omni-directional person tracking on a flying robot using occlusion-robust ultra-wideband signals. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 189–194. https://doi.org/10.1109/IROS.2016.7759054

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

Jeroen D. Hol, Fred Dijkstra, Henk Luinge, and Thomas B. Schon. 2009. Tightly coupled UWB/IMU pose estimation. In *2009 IEEE International Conference on Ultra-Wideband*. 688–692. https://doi.org/10.1109/ICUWB.2009.5288724

Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. 2021. Conditional directed graph convolution for 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 602–611.

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

IEEE. 2007. IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs): Amendment 1: Add Alternate PHYs. *IEEE Std 802.15.4a-2007 (Amendment to IEEE Std 802.15.4-2006)* (2007), 1–210. https://doi.org/10.1109/IEEESTD.2007.4299496

Jiaxi Jiang, Paul Streli, Manuel Meier, Andreas Fender, and Christian Holz. 2023. EgoPoser: Robust Real-Time Ego-Body Pose Estimation in Large Scenes. *arXiv preprint arXiv:2308.06493* (2023).

Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022a. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 443–460.

Jiaxi Jiang, Kai Zhang, and Radu Timofte. 2021. Towards flexible blind JPEG artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4997–5006.

Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. 2022b. Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 9 pages. https://doi.org/10.1145/3550469.3555428

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7122–7131.

Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. 2021. EM-POSE: 3D Human Pose Estimation From Sparse Electromagnetic Trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11510–11520.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

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

Daniel Laidig and Thomas Seel. 2023. VQF: Highly accurate IMU orientation estimation with bias estimation and magnetic disturbance rejection. *Information Fusion* 91 (March 2023), 187–204. https://doi.org/10.1016/j.inffus.2022.10.014

Anton Ledergerber, Michael Hamer, and Raffaello D'Andrea. 2019. Angle of Arrival Estimation based on Channel Impulse Response Measurements. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6686–6692. https://doi.org/10.1109/IROS40897.2019.8967562

Sangmin Lee, Seungho Yoo, Joon Yeop Lee, Seongjoon Park, and Hwangnam Kim. 2022. Drone Positioning System Using UWB Sensing and Out-of-Band Control. *IEEE Sensors Journal* 22, 6 (2022), 5329–5343. https://doi.org/10.1109/JSEN.2021.3127233

Shuang Li, Jiaxi Jiang, Philipp Ruppel, Hongzhuo Liang, Xiaojian Ma, Norman Hendrich, Fuchun Sun, and Jianwei Zhang. 2020a. A mobile robot hand-arm teleoperation system by vision and IMU. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10900–10906.

Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. 2020b. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6173–6183.

Huajun Liu, Xiaolin Wei, Jinxiang Chai, Inwoo Ha, and Taehyun Rhee. 2011. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*. 133–140.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. 2022. COAP: Compositional articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13201–13210.

Francisco Molina Martel, Juri Sidorenko, Christoph Bodensteiner, and Michael Arens. 2018. Augmented reality and UWB technology fusion: Localization of objects with head mounted displays. In *Proceedings of the 31st International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2018)*. 685–692.

Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation Using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. https://doi.org/10.1145/3544548.3581392

Mark W. Mueller, Michael Hamer, and Raffaello D'Andrea. 2015. Fusing ultra-wideband range measurements with accelerometers and rate gyroscopes for quadrocopter state estimation. , 1730-1736 pages. https://doi.org/10.1109/ICRA.2015.7139421

Deepak Nagaraj, Erik Schake, Patrick Leiner, and Dirk Werth. 2020. An RNN-ensemble approach for real time human pose estimation from sparse IMUs. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*. 1–6.

Noitom. 2024. *https://www.noitom.com/*. https://www.noitom.com/

Aitor Ochoa-de Eribe-Landaberea, Leticia Zamora-Cadenas, Oier Peñagaricano-Muñoa, and Igone Velez. 2022. UWB and IMU-Based UAV's Assistance System for Autonomous Landing on a Platform. *Sensors* 22, 6 (Mar 2022), 2347. https://doi.org/10.3390/s22062347

Optitrack. 2023. *https://wwww.optitrack.com/*. https://www.optitrack.com/

Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. 2023. Fusing Monocular Images and Sparse IMU Signals for Real-Time Human Motion Capture. In *SIGGRAPH Asia 2023 Conference Papers* (, Sydney, NSW, Australia,) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 116, 11 pages. https://doi.org/10.1145/3610548.3618145

Jorge Peña Queralta, Li Qingqing, Fabrizio Schiano, and Tomi Westerlund. 2022. VIO-UWB-Based Collaborative Localization and Dense Scene Reconstruction within Heterogeneous Multi-Robot Systems. arXiv:2011.00830 [cs.RO]

Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. 2016. EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.* 35, 6, Article 162 (nov 2016), 11 pages. https://doi.org/10.1145/2980179.2980235

Zafer Sahinoglu, Sinan Gezici, and Ismail Güvenc. 2008. *Ultra-wideband Positioning Systems: Theoretical Limits, Ranging Algorithms, and Protocols*. Cambridge University Press. https://doi.org/10.1017/CBO9780511541056

Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. 2011. Motion Capture from Body-Mounted Cameras. *ACM Trans. Graph.* 30, 4, Article 31 (jul 2011), 10 pages. https://doi.org/10.1145/2010324.1964926

Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. 2023. HOOV: Hand Out-Of-View Tracking for Proprioceptive Interaction using Inertial Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

Janis Tiemann and Christian Wietfeld. 2017. Scalable and precise multi-UAV indoor navigation using TDOA-based UWB localization. In *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. 1–7. https://doi.org/10.1109/IPIN.2017.8115937

D. Tome, P. Peluse, L. Agapito, and H. Badino. 2019. xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 7727–7737. https://doi.org/10.1109/ICCV.2019.00782

Vu Tran, Zhuangzhuang Dai, Niki Trigoni, and Andrew Markham. 2022. DeepCIR: Insights into CIR-based Data-driven UWB Error Mitigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 13300–13307. https://doi.org/10.1109/IROS47612.2022.9981931

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

Ubisense. 2023. *https://ubisense.com/*. https://ubisense.com/

Vicon. 2023. *https://wwww.vicon.com/*. https://www.vicon.com/

Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. 2007. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)* 26, 3 (2007), 35–es.

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

Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 349–360.

Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. 2021. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11500–11509.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

Xsens. 2024. *https://www.xsens.com*. https://www.xsens.com/

Hao Xu, Yichen Zhang, Boyu Zhou, Luqi Wang, Xinjie Yao, Guotao Meng, and Shaojie Shen. 2022. Omni-Swarm: A Decentralized Omnidirectional Visual–Inertial–UWB State Estimation System for Aerial Swarms. *IEEE Transactions on Robotics* 38, 6 (2022), 3374–3394. https://doi.org/10.1109/TRO.2022.3182503

Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. 2021. LoBSTr: Real-time Lower-body Pose Prediction from Sparse Upper-body Tracking Signals. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 265–275.

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. 2023. EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors. *ACM Transactions on Graphics (TOG)* 42, 4, Article 76 (2023), 17 pages.

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13167–13178.

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

Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. 2020. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2200–2209.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3425–3435.

Minghui Zhao, Tyler Chang, Aditya Arun, Roshan Ayyalasomayajula, Chi Zhang, and Dinesh Bharadia. 2021. ULoc: Low-Power, Scalable and Cm-Accurate UWB-Tag Localization and Tracking for Indoor Applications. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 140 (sep 2021), 31 pages. https://doi.org/10.1145/3478124

Shuaikang Zheng, Zhitian Li, Yuanli Yin, Yunfei Liu, Haifeng Zhang, Pengcheng Zheng, and Xudong Zou. 2022. Multi-robot relative positioning and orientation system based on UWB range and graph optimization. *Measurement* 195 (2022), 111068. https://doi.org/10.1016/j.measurement.2022.111068

Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. 2023. Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling. *arXiv preprint arXiv:2308.08855* (2023).

Zhiming Zou and Wei Tang. 2021. Modulated graph convolutional network for 3D human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11477–11487.

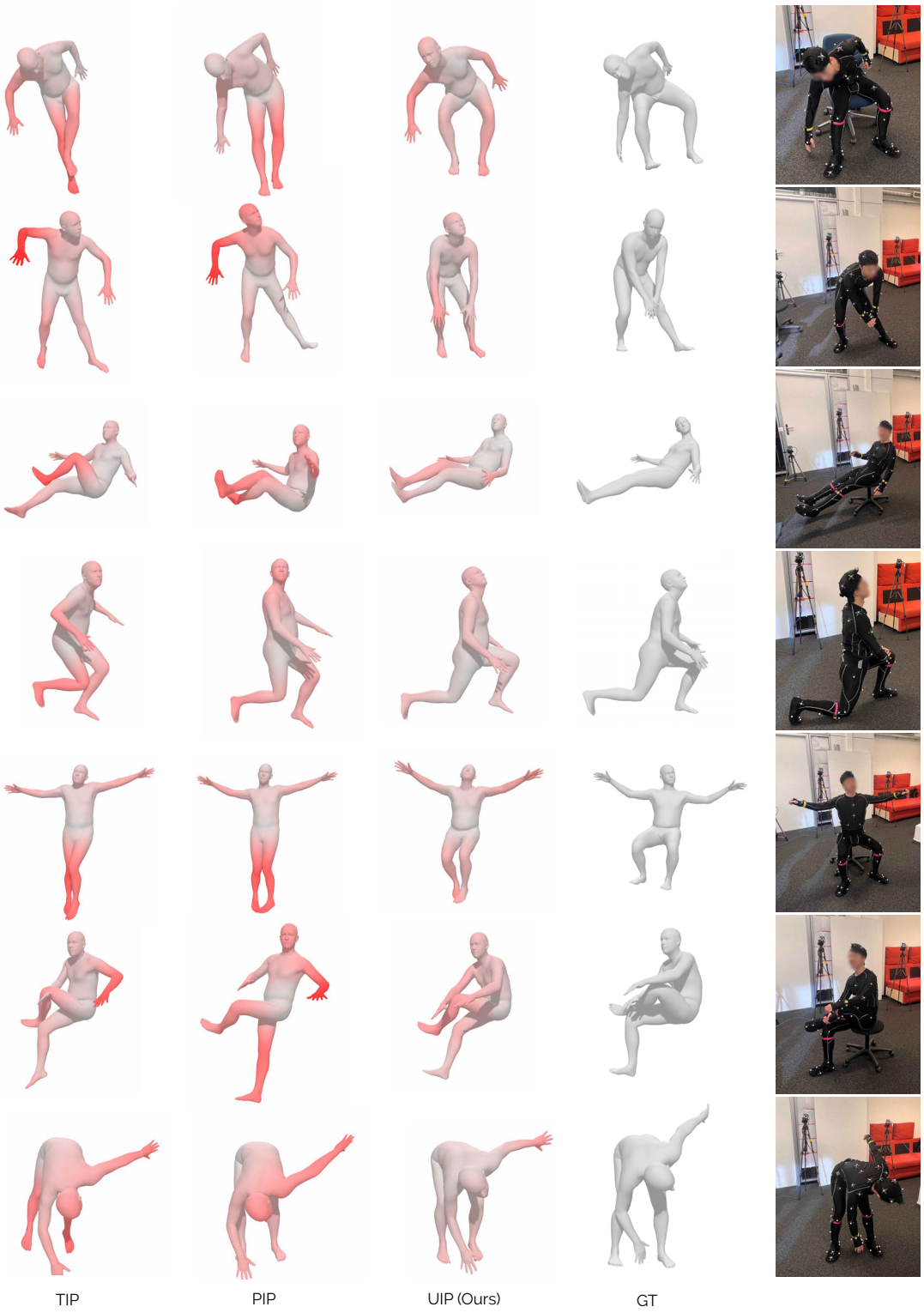

Figure 6: Qualitative comparisons among different methods on our test set.

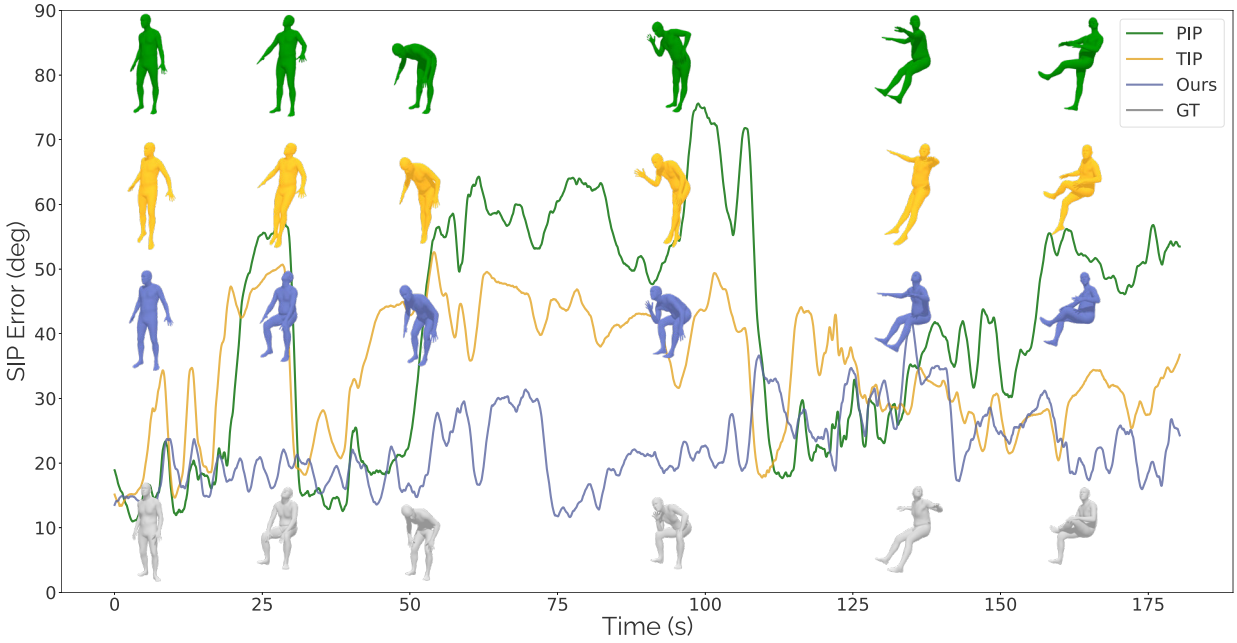

Figure 7: Visualization of SIP Error over time on a test sequence from UIP-DB.

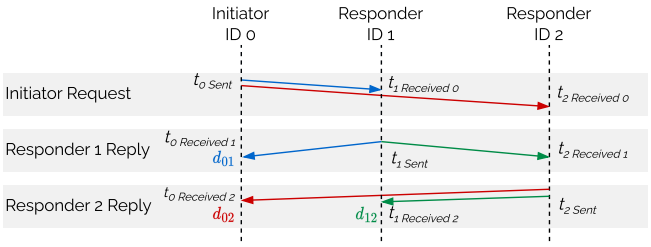

Figure 8: Ranging transaction with three devices. Timestamps to resolve time-of-flight are included in the UWB message payload and thus broadcast to all network participants.

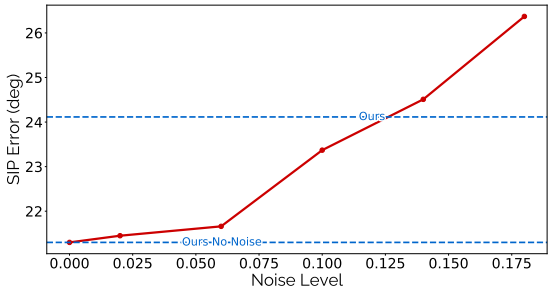

Figure 9: SIP error of UIP on our dataset at different levels of synthetic Gaussian noise on inter-sensor distances.

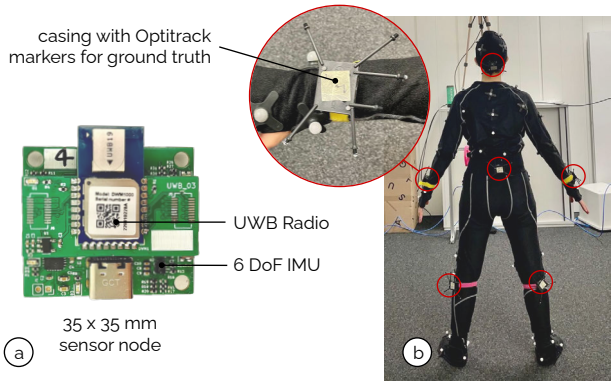

Figure 10: (a) Our embedded sensor nodes and (b) a person wearing six of them inside a 20-camera Optitrack motion capture setup for reference poses.

Figure 11: Example of motion tracking using our method of a person exercising outdoors, where no tracking infrastructure is needed for reference or anchoring.