Dynamic Inertial Poser (DynaIP): Part-Based Motion Dynamics Learning for Enhanced Human Pose Estimation with Sparse Inertial Sensors

Yu Zhang[1*] Songpengcheng Xia [1*] Lei Chu [2 †] Jiarui Yang[1] Qi Wu[1] Ling Pei[1†]
[1]Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai Jiao Tong University
[2]Wireless Devices and Systems Group (WiDeS), University of Southern California

Abstract

This paper introduces a novel human pose estimation approach using sparse inertial sensors, addressing the shortcomings of previous methods reliant on synthetic data. It leverages a diverse array of real inertial motion capture data from different skeleton formats to improve motion diversity and model generalization. This method features two innovative components: a pseudo-velocity regression model for dynamic motion capture with inertial sensors, and a part-based model dividing the body and sensor data into three regions, each focusing on their unique characteristics. The approach demonstrates superior performance over state-of-the-art models across five public datasets, notably reducing pose error by 19% on the DIP-IMU dataset, thus representing a significant improvement in inertial sensor-based human pose estimation. Our codes are available at https://github.com/dx118/dynaip

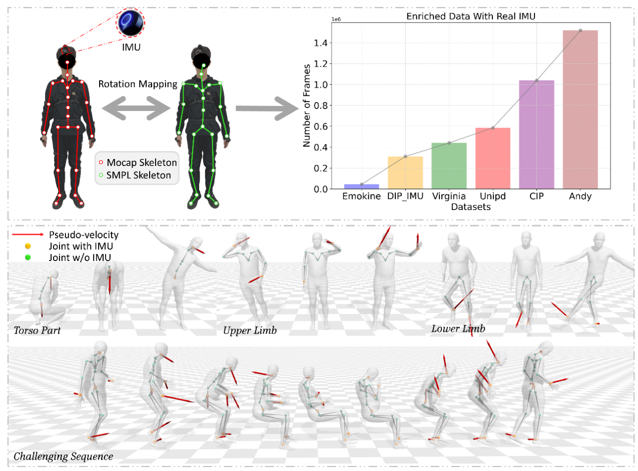

Figure 1. Our innovative data-driven approach for robust full-body pose estimation using six IMUs: unifying inertial mocap datasets across skeleton formats and enhancing challenging motion capture with local body region modeling and pseudo-velocity estimation.

1. Introduction

Human Pose Estimation (HPE) has emerged as a critical field of study, attracting considerable interest for its applications in various domains [15, 51], like aiding in sports training and analysis, and enriching interactions in Virtual and Augmented Reality (VR/AR) environments. This growing relevance underscores the importance of advancing HPE technologies to meet the diverse needs of these applications.

Our paper examines HPE, a field marked by varied sensing modalities and methodologies [2, 6, 10], divided into three categories: 1) Vision-based HPE [35, 39], using single or multi-view images, known for its significant advancements; 2) Wireless-based HPE [3, 50], which addresses some vision-based challenges but is limited by environmental factors; 3) Wearable-based HPE [14, 17, 46], our focus, which utilizes Inertial Measurement Units (IMUs) for full-body estimation. Vision-based HPE, despite its success, struggles with occlusion, privacy, and perspective issues. Wireless-based methods [3, 50] mitigate some of these issues but face performance inconsistencies. Our study centers on wearable-based HPE, harnessing wearables with sensors for joint orientation and acceleration tracking. This IMU-based method excels by being occlusion-resistant, environmentally stable, and privacy-conscious, thus addressing the drawbacks of other HPE approaches.

Recent studies [14, 26, 45] have concentrated on minimizing the number of IMU sensors for human motion reconstruction, balancing user convenience with less intrusiveness. While this has shown potential, the limited sensor placement leads to uncertainties in estimating unseen joint rotations. The integration of complex network structures [16] and physical constraints [46] has been proven effective in reducing ambiguities. Despite these developments, challenges persist, especially in complex motions. Our research aims to enhance the robustness of sparse sensor-based full-

*Equal contribution
†Corresponding authors
This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 62273229.

body pose reconstruction by identifying three key areas for potential improvements.

Multi-modal Sensor Information Utilization: The primary challenge in IMU-based HPE lies in optimally utilizing multi-modal sensor data, especially acceleration, to diminish motion ambiguity. Sole reliance on IMU orientation data for full-body pose estimation can result in ambiguities, particularly in movements like hand or leg raises. Acceleration data, with its dynamic motion features, is crucial for reducing such ambiguities. However, existing models inadequately leverage acceleration due to two main factors: 1) Acceleration data is noisier than rotation measurements, while IMUs provide accurate orientation estimates via Kalman filters; 2) Raw acceleration measurements fail to effectively capture continuous joint motion states, as exemplified during transitions from stationary to constant velocity, where accelerometers record peak values only at motion onset. To address these issues, we propose a two-stage model that estimates joint velocities with IMUs in the first stage, enhancing the utilization of acceleration data.

Spatial Relationship Exploitation of Human Body Parts and Wearable Sensors: In IMU-based HPE, previous studies have focused on using temporal information to reconstruct complex motions, such as the ambiguity between sitting and standing or long-sitting [16, 46]. However, there has been less exploration into the varied distributions within IMU-based HPE datasets. For instance, certain upper limb motions are exclusively associated with standing in the training data, leading to potential mistakenly reconstructed in scenarios where similar motions occur while sitting during testing. To tackle this issue, our model draws inspiration from two key observations noted in recent research [20, 36, 36]: 1) globally rare poses often comprise local joint configurations that are frequently represented in training datasets; 2) there is a significant dependency among nearby joints, which decreases as the distance between joints increases. Considering these insights, we propose a part-based HPE model that divides the human body into three regions: upper limbs, torso, and lower limbs, which allows the model to concentrate on the distinct characteristics of different body regions. By acknowledging the spatial relationships of body parts and sensor distribution, our model aims to enhance accuracy in pose estimation, especially in scenarios where similar motions are performed in different postures.

Human Pose Estimation Model Generalization: The scarcity of motion capture data with inertial measurement in the Skinned Multi-Person Linear (SMPL) model format [22] restricts the performance of IMU-based models. To address this, earlier research, beginning with DIP [14], utilized virtual IMU data generated from the AMASS dataset [24] to increase the diversity of training samples. While this virtual-to-reality approach has shown efficacy in various domains [23, 30, 31, 42], a notable discrepancy remains between virtual and real IMU measurements, hindering further advancements in IMU-based HPE tasks.

As human pose estimation research gains traction, the availability of motion capture datasets with real IMU measurements is increasing. However, a significant challenge in utilizing these real datasets is the variation in skeleton formats they present. Our work introduces a straightforward yet efficient mapping strategy to reconcile different skeleton formats, allowing the incorporation of additional real-world motion capture datasets with actual IMU measurements into the training of IMU-based HPE models. This integration results in improved accuracy and generalization in pose estimation. The contributions of this paper are summarized as follows:

- Our research introduces an innovative two-stage deep learning model designed for real-time and robust human pose estimation utilizing sparse IMU sensors, called DynaIP (Dynamic Inertial Poser). This method uniquely addresses motion ambiguity by learning pseudo velocities, which allows for the full utilization of acceleration data and leverage the strengths of sparse sensor data.
- Our approach divides the human body and associated IMU sensors into three local regions, forming the basis for our part-based human pose estimation model. This model incorporates low-dimensional global motion information to avoid the full-body pose inconsistencies. By focusing on individual body parts, the model minimizes the influence of less associated joints, thereby enhancing the robustness and reliability of motion tracking.
- We incorporate more real-world Mocap data with IMU measurements across different skeleton formats and applied them uniformly in IMU-based HPE model training. The extensive experimental results demonstrate that our approach significantly outperforms competitors and exhibits good generalization performance.

2. Related Work

HPE has been widely explored using different methods, including visual [8, 18, 35, 39], inertial [14, 17, 46], wireless [3, 43], and various hybrid approaches [32, 49]. Our paper specifically concentrates on IMU-based human pose estimation solutions, delving into their unique advantages and potential applications.

Human Pose Estimation with IMU sensors. Inertial motion capture (mocap) systems, known for their freedom from occlusion and lighting limitations, have witnessed significant progress in recent years. Commercial systems like Xsens [34] are accurate but require numerous sensors, making them less practical. Early attempts [37, 38] to address sensor noise and observation gaps in sparse setups involved reconstructing human motion from sparse accelerometers by referencing pre-recorded databases. Subsequent work

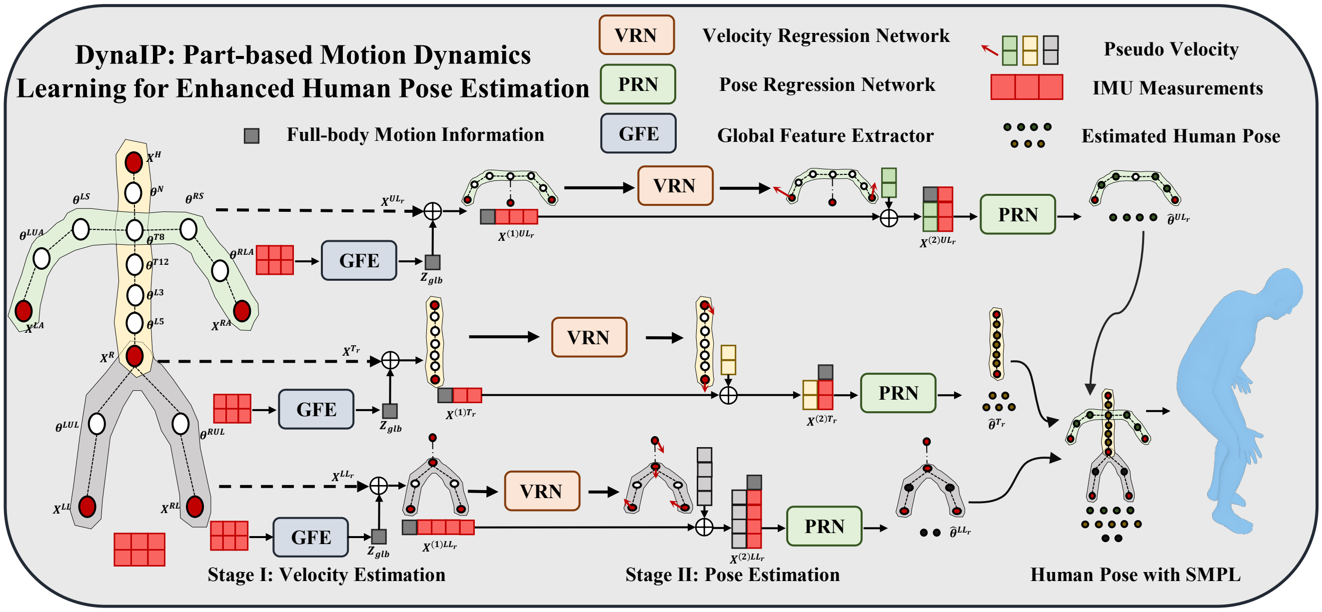

Figure 2. Overview of our proposed method, a part-based human pose estimation model with pseudo-velocity regression. Our model incorporates a two-stage structure. The first stage predicts joint velocities using IMU measurements, while the second stage focuses on predicting the entire body's joints rotation. Additionally, we partition the human body and the attached IMU sensors into three local regions. These regions are input into our proposed part-based human pose estimation model, designed to estimate each local region's pose while maintaining global coherency. This multi-stage and part-based approach enhances the accuracy and consistency of our pose estimation.

by Marcard et al. [41] introduced an offline optimization method using the SMPL model to fit sparse IMU data. Huang et al. [14] made strides by applying deep learning to real-time regression of SMPL pose parameters from IMUs, though they did not fully leverage acceleration information. Yi et al. [45, 46] improved upon this with a multi-stage birnn structure, hierarchically regressing joint locations and integrating a physical optimizer. Jiang et al. [16] introduced stationary body points (SBP) as an additional training target, utilizing zero-velocity information to address motion drift. However, existing methods have not effectively harnessed the multi-modal information within IMUs and the spatial information of the human body. In our paper, we divide the human body and the worn IMU sensors into three regions and devise a two-stage structure with three part-based branches for human pose estimation.

Handling Data Scarcity in Inertial Motion Capture. Large-scale training data plays a crucial role in the development of learning-based methods. In previous IMU-based research, the SMPL model has been commonly used to represent human pose. However, datasets that provide SMPL ground truth, such as those used in works like [14, 40], are limited in size and diversity. The acquisition of SMPL ground truth data, whether through marker-based systems with the Mosh++ [21] algorithm or offline optimizations, is prohibitively expensive, which has restricted dataset availability. To address this limitation, IMU-based methods like [14, 16, 45, 46] have predominantly utilized the AMASS dataset to generate a wealth of virtual sensor data. They have employed a virtual-to-reality transfer learning approach during model training. While virtual IMU data has been valuable in augmenting training samples, there remains a noticeable gap between virtual and real measurement noise. This gap has hindered further improvements in performance [19, 42]. In contrast to SMPL pose parameters, ground truth obtained through mocap systems with their native skeleton representation is more convenient. There are datasets available that provide inertial mocap data with Xsens ground truth, such as [4, 7, 9, 25, 29]. However, it's worth noting that these datasets use different skeleton formats for their ground truth. For example, the skeleton formats of SMPL and Xsens differ, shown in Fig. 1. While vision-based solutions have addressed this issue by combining multiple datasets with varying ground truth representations and using novel autoencoders to mitigate positional disparities in ground truth [33], our approach takes a simpler yet effective route. We introduce a mapping strategy that enables us to seamlessly incorporate additional real-world MoCap datasets with different skeleton formats into our training process. This strategy allows us to harness the rich diversity of available data for improved performance.

3. Method

Our objective is to accurately estimate human pose, denoted as $\hat{\theta}$, using data from six sensors worn on different body

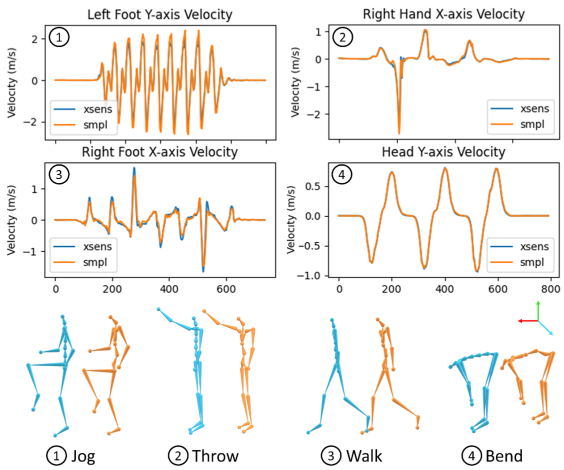

Figure 3. The evaluation of the end-effector velocity across various motions. With the mapping of joint's global orientations from Xsens to SMPL, there is no significant discrepancy in the end-effector velocities.

parts. The corresponding IMU data can be represented by $\boldsymbol{X} \in \mathcal{R}^{T \times 72} = [\boldsymbol{X^R}, \boldsymbol{X^{LL}}, \boldsymbol{X^{RL}}, \boldsymbol{X^H}, \boldsymbol{X^{LA}}, \boldsymbol{X^{RA}}]$, marked in Fig. 2, where T is the length of data samples. Additionally, our model predicts the pseudo velocity ($\hat{\boldsymbol{V}}$) of the joints with IMU measurements. As shown in Fig. 2, our model comprises three modules: 1) A method for unifying training data across different skeleton formats using a global orientation mapping strategy. 2) A two-stage human pose estimation structure with pseudo velocity regression. 3) A part-based 3D human dynamics learning module with low-dimensional full-body motion information.

3.1. Training Data Unified across Skeleton Formats

Early learning-based methods [14, 45] for IMU-based HPE encountered a significant challenge: the lack of datasets that contain real IMU data alongside corresponding ground truths. To address this challenge, these methods have taken an innovative approach by generating virtual IMU data using the AMASS dataset [24], improving the model's generalization capabilities due to its diversity. Fine-tuning the model with a small amount of collected real data resulted in impressive pose estimation results. However, recent studies [16, 19, 42] have highlighted a performance degradation issue stemming from differences in noise distributions between virtual and real IMU data. This issue has been discussed in various tasks [13, 48], indicating the need for a solution. As interest in IMU-based HPE grows, several datasets containing real IMU measurements have emerged [7, 9, 25, 29]. These datasets represent human pose ground truths but use different skeleton formats. Instead of solely addressing the gap between virtual and real domains [14, 19], our approach takes a novel direction. We propose a training method that initially integrates motion capture datasets with real IMU data from various skeleton formats, offering a promising solution to enhance pose estimation accuracy.

The SMPL [22] and Xsens [34] skeletons are two widely used skeleton representations. While they share overall structural similarities, one notable difference lies in the number of torso joints, as depicted in Fig. 1. Previous research efforts [5, 11] have achieved success in mapping human poses captured using commercial inertial mocap systems onto the SMPL skeleton. This mapping process involves replicating relative joint rotations and excluding redundant torso joints. In the skeleton model, each 3D rigid bone's motion can be denoted by a homogeneous matrix $M_{bone} \in SE(3)$ [28].

$$M_{bone} = \begin{bmatrix} \boldsymbol{R_b} & \boldsymbol{p_b} \\ \boldsymbol{0}_{1\times3} & 1 \end{bmatrix}, \tag{1}$$

where $\boldsymbol{R_b} \in SO(3)$ is a global rotation matrix and $\boldsymbol{p_b} \in \mathcal{R}^3$ is a predefined bone displacement. For inertial-based mocap, IMUs, attached to the human body, offer sensor's global orientation $\boldsymbol{R_s} \in SO(3)$. After calibration [14, 45, 46], the sensor orientation $\boldsymbol{R_s}$ serves as a direct representation of the respective bone's orientation $\boldsymbol{R_b}$. As such, although specific bone displacements $\boldsymbol{p_{bone}}$ can be defined differently within SMPL and Xsens skeleton, the bone orientations would be consistent with the IMU readings because they correspond to the same location on human body. Therefore, we've established a one-to-one ground truth mapping across skeletons, e.g., wrists to wrists, legs to legs, and so on. For the torso, we eliminate this redundant joint to maintain the consistency of our mapping. With this mapping strategy, we could unify the inertial mocap data across different skeleton formats.

In line with previous tasks related to human motion retargeting [1, 27], we utilize the velocity of end-effectors as a qualitative measure to demonstrate the practicality of our mapping process. As depicted in Fig. 3, our global orientation mapping strategy consistently maintains accuracy across various motions. This serves as a robust foundation for unifying a more extensive dataset of real IMU mocap data. By training the model with this unified dataset, we enable it to fully leverage real IMU data, resulting in improved generalization performance compared to the traditional virtual-to-reality approach.

3.2. Two-stage Human pose estimation with Pseudo Velocity Regression

Given our approach's reliance on sparse IMUs attached to limb ends for estimating full-body pose, there is an inherent challenge in accurately estimating joint rotations without direct IMU measurements. Previous models [14, 16] often overlook the rich multi-modal information contained in the IMUs, primarily relying on global orientation measurements. Therefore, it is essential to effectively integrate the dynamic information provided by acceleration into our learning-based method. To address this, we introduce a two-stage HPE model that estimates, in two distinct stages, the velocity of joints with IMUs and the global rotation of each joint in the full-body. Our method comprises two primary stages: Pseudo velocity regression (Stage I) and human pose estimation (Stage II). This two-stage approach allows us to leverage both acceleration and orientation data to enhance pose estimation accuracy.

3.2.1 Stage I: pseudo velocity regression

As previously mentioned in Sec. 1, relying on raw acceleration data may not effectively capture the continuous motion states of joints due to its sensitivity to instantaneous dynamics. In response to this challenge, Stage I of our model is dedicated to regressing the pseudo velocity of the joints with IMUs. This step is crucial for effectively extracting the dynamic information contained in acceleration data. The underlying principles guiding this design are twofold:

- **Velocity as a Motion Indicator:** Velocity has been demonstrated to better reflect the joints motion dynamics, as supported by previous research [12, 44]. It plays a crucial role in compensating for the motion ambiguity that can arise from underutilizing different modalities.
- **Velocity Estimation:** Velocity can be obtained by integrating acceleration and ensure that the dynamic information remains accurate and reliable, unlike raw acceleration. We employ a neural network to estimate the velocities of joints with IMUs, thereby harnessing the IMU's potential to express continuous body dynamics.

Our model's Stage-I takes the raw IMU measurements $\boldsymbol{X} = [\boldsymbol{X^R}, \boldsymbol{X^{LL}}, \boldsymbol{X^{RL}}, \boldsymbol{X^H}, \boldsymbol{X^{LA}}, \boldsymbol{X^{RA}}]$ as input, the velocity $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{V}}^{\boldsymbol{R}}, \hat{\boldsymbol{V}}^{\boldsymbol{LL}}, \hat{\boldsymbol{V}}^{\boldsymbol{RL}}, \hat{\boldsymbol{V}}^{\boldsymbol{H}}, \hat{\boldsymbol{V}}^{\boldsymbol{LA}}, \hat{\boldsymbol{V}}^{\boldsymbol{RA}}]$ as output via the model's Velocity Regression Network (VRN) module, denoted by $S_{VRN}(\cdot)$. Following the advanced learning-based RNN initialization strategy proposed by PIP [46], our model additionally takes the initial velocities$\boldsymbol{V_0}$ of leaf joints as the model's input. After that, we get the predicted pseudo velocity $\hat{\boldsymbol{V}}$, represented by

$$\hat{\boldsymbol{V}} = S_{VRN}(\boldsymbol{X^{(1)}}, \boldsymbol{V}_0), \tag{2}$$

where $\boldsymbol{X^{(1)}} = \boldsymbol{X}$ represents the input of the first stage.

The VRN module is comprised of two main components: a Multi-Layer Perceptron (MLP) and a two-layer Long Short-Term Memory (LSTM) network. The MLP takes the initial velocity as input. Its primary purpose is to process and transform this initial velocity information. The outputs generated by the MLP are then assigned to serve as the hidden state and cell state inputs for the first frame of the LSTM network. This design allows for the initial velocity information to be effectively processed and used as a starting point for the LSTM, which can then continue to capture and predict the velocity and motion dynamics over subsequent frames. For the intermediate variable $\hat{\boldsymbol{V}}$ to be supervised, we obtain the velocity ground-truth $\boldsymbol{V} = (FK(\boldsymbol{\theta_t}) - FK(\boldsymbol{\theta_{t-1}}))/\Delta t$ through the ground-truth of human pose $\boldsymbol{\theta}$ and forward kinematics $FK(\cdot)$ for supervision. The pseudo-velocity loss L_{vel} could be represented by:

$$\begin{aligned} L_{vel} &= \|\boldsymbol{V} - \hat{\boldsymbol{V}}\|_2 \\ &= \|\frac{FK(\boldsymbol{\theta_t}) - FK(\boldsymbol{\theta_{t-1}})}{\Delta t} - S_{VRN}(\boldsymbol{X}, \boldsymbol{V}_0)\|_2, \end{aligned} \tag{3}$$

where Δt represents the time interval and $\boldsymbol{\theta_t}$ the ground-truth of human pose at time t. For the velocity of the root joint, we only preserve its vertical component.

3.2.2 Stage II: human pose estimation

Building upon the dynamic information extracted by the VRN module, our model proceeds to Stage II, where the primary goal is to achieve robust human pose estimation. For Stage II, we design a Pose Regression Network (PRN) module, denoted by $S_{PRN}(\cdot)$, to estimate the joint rotations $\hat{\boldsymbol{\theta}}$ from the IMU measurements $\boldsymbol{X}$ and estimated leaf joints velocities $\hat{\boldsymbol{V}}$, which are concatenated as input $\boldsymbol{X^{(2)}} = [\hat{\boldsymbol{V}}, \boldsymbol{X}]$ to the Stage II. Based on the initialization strategy and the PRN module, our model could output the estimated human pose $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = S_{PRN}(\boldsymbol{X^{(2)}}, \boldsymbol{\theta_0}), \tag{4}$$

where the $\boldsymbol{\theta_0}$ is the initial human pose, and the PRN $S_{PRN}(\cdot)$ is applied with a two-layers LSTM.

With the predicted human pose $\hat{\boldsymbol{\theta}}$ and the ground-truth $\boldsymbol{\theta}$, we could get the loss L_{pose} represented by:

$$L_{pose} = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2 = \|\boldsymbol{\theta} - S_{PRN}(\boldsymbol{X^{(2)}}, \boldsymbol{\theta_0})\|_2, \tag{5}$$

Our two-stage network, which includes velocity regression, successfully addresses the ambiguities arising from the under-utilization of acceleration. However, considering the distinctive motion patterns of the upper and lower limbs, as well as the infrequent motion combinations, it is crucial to harness the spatial information of the human body to ensure robust motion tracking.

3.3. Learning Part-based 3D Human Dynamics with Three Local Body Regions

Directly using all six IMU measurements as the model's input, as done in previous approaches [14, 46], without considering the inherent spatial relationship of the human body, can lead to motion ambiguity due to weak associations between body parts. In this section, drawing inspiration from [20, 36], we introduce local region modeling to mitigate this issue. Our model divides the entire body into three local regions: upper limbs region (UL_r), torso region (T_r), and lower limbs region (LL_r), as shown in Fig. 2.

Correspondingly, the IMU measurements and the estimated human pose in our proposed model are partitioned into three local regions. To create the model's input, we group the IMU sensors on our bodies into three sets, denoted as $\boldsymbol{X^l}, l \in \{UL_r, T_r, LL_r\}$, which can be represented by $\boldsymbol{X^{UL_r}} = [\boldsymbol{X^R}, \boldsymbol{X^{LA}}, \boldsymbol{X^{RA}}], \boldsymbol{X^{T_r}} = [\boldsymbol{X^R}, \boldsymbol{X^H}], \boldsymbol{X^{LL_r}} = [\boldsymbol{X^R}, \boldsymbol{X^{LL}}, \boldsymbol{X^{RL}}, \boldsymbol{X^H}]$.

Following the previously mentioned two-stage human pose estimation structure, we establish three sub-models to acquire the three groups of inputs, and estimate the part-based pseudo velocity $\hat{\boldsymbol{V}}^{\boldsymbol{l}}, l \in \{UL_r, T_r, LL_r\}$ and the joint rotation outputs $\hat{\boldsymbol{\theta}}^{\boldsymbol{l}}, l \in \{UL_r, T_r, LL_r\}$. Our estimated joints of full-body are also output in three sub-modules, specifically expressed as $\boldsymbol{\theta^{UL_r}} = [\boldsymbol{\theta^{LS}}, \boldsymbol{\theta^{LUA}}, \boldsymbol{\theta^{RS}}, \boldsymbol{\theta^{RUA}}], \boldsymbol{\theta^{T_r}} = [\boldsymbol{\theta^{LUL}}, \boldsymbol{\theta^{RUL}}], \boldsymbol{\theta^{LL_r}} = [\boldsymbol{\theta^{L5}}, \boldsymbol{\theta^{L3}}, \boldsymbol{\theta^{T12}}, \boldsymbol{\theta^{T8}}, \boldsymbol{\theta^{N}}]$. The representation of the joints is shown in Fig. 2.

With our part-based two-stage structure, our model achieves a comprehensive representation of full-body joint rotations by synthesizing the outputs from three sub-models. This part-based design, relying on local inputs and features, effectively learns unique pose configurations for each body part and minimizes the negative impact of weakly associated joints. However, one challenge we face with this part-based approach is the potential lack of global coherence in the estimated poses. Furthermore, to address this, we incorporate global body motion information into our part-based HPE model, drawing inspiration from previous work [20, 47]. Our model utilizes a global feature extractor, denoted as S_{GLB}, to coarsely capture the full-body motion information $\boldsymbol{Z_{glb}}$ from all six IMUs $\boldsymbol{X}$. Subsequently, the global representation is concatenated with the inputs of the two-stage network featuring three local region branches. The input of Stage I ($\boldsymbol{X^{(1)_l}}$) and Stage II ($\boldsymbol{X^{(2)_l}}$) could be re-represented by

$$\begin{aligned} \boldsymbol{X^{(1)_l}} &= [\boldsymbol{X^l}, \boldsymbol{Z_{glb}}], \\ \boldsymbol{X^{(2)_l}} &= [\hat{\boldsymbol{V}}^l, \boldsymbol{X^l}, \boldsymbol{Z_{glb}}], l \in \{UL_r, T_r, LL_r\}, \end{aligned} \tag{6}$$

where the $\hat{\boldsymbol{V}}^{\boldsymbol{l}} = S^l_{VRN}(\boldsymbol{X^{(1)_l}}, \boldsymbol{V}^l_0)$ is the local pseudo velocity estimated by the local VRN module $S^l_{VRN}(\cdot)$, and $\boldsymbol{V}^l_0$ means the initial velocities of the local regions. Therefore, we get the loss function for two stages:

$$L^l_{vel} = \|\boldsymbol{V}^l - \hat{\boldsymbol{V}}^l\|_2, L^l_{pose} = \|\boldsymbol{\theta}^l - \hat{\boldsymbol{\theta}}^l\|_2, \tag{7}$$

where the $\boldsymbol{V}^l, \boldsymbol{\theta}^l, l \in \{UL_r, T_r, LL_r\}$ are the ground-truth of the velocity and human pose for three local regions, and the $\boldsymbol{\theta^l}$ is the local human pose estimated by the local PRN module $S^l_{PRN}(\boldsymbol{X^{(2)_l}}, \boldsymbol{\theta^l_0})$ of three local branches. By integrating local regions with global information, our model ensures that while each part-based branch effectively learns localized motion dynamics, the overall pose estimation remains coherent and consistent with the global motion patterns of the human body. This approach strikes a balance between local and global information, resulting in robust and accurate pose estimation.

With the technical analysis provided above, we can formulate the final objective function for training our model as follows:

$$L = \sum_{l \in \{UL_r, T_r, LL_r\}} (L^l_{pose} + L^l_{vel}) \tag{8}$$

4. Experiments

Experiment Setup. Our experimental evaluation is structured into three main parts: Firstly, we illustrate the benefits of using unified inertial mocap data compared with a virtual-to-real training scheme. Secondly, we show the effectiveness of our proposed model with other state-of-the-art methods on our unified mocap data. Finally, we conduct ablation studies on key components of our model.

Datasets. We utilized a combination of datasets for both training and evaluation in our experiments. The datasets include DIP-IMU [14] and Xsens datasets, which encompass AnDy [25], Emokine [4], Virginia Natural Motion [7], UNIPD [9], and CIP [29]. Additionally, we used the AMASS [24] dataset for evaluating the effectiveness of virtual and real mocap data. Detailed information about these datasets can be found in the supplementary document.

Metrics. Following [45, 46], we use the following metrics for evaluation: *1) SIP error [°]*: the mean global rotation difference of upper arms and upper legs between the estimation and ground truth; *2) Global Angular error [°]*: the mean global rotation error between estimated joints and ground truth; *3) Position error [cm]*: the mean Euclidean distance error between all joints and ground truth, root position is aligned. *4) Mesh error [cm]*: the mean vertex distance between estimated meshes and ground truth.

4.1. Impact of the Unified Inertial Mocap Data and Virtual-to-Real Training Scheme

Comparing performance of our model using different training settings. To evaluate the impact of unified training data and virtual dataset, we conducted an experiment on these four training settings: *1)* Synthetic AMASS data only; *2)* Synthetic AMASS data with DIP-IMU Fine-tuning; *3)* Real Xsens data only (DynaIP); *4)* Combined DIP-IMU and Xsens data (DynaIP*). With these four settings, we evaluated our model on the DIP-IMU test set. Furthermore, for models only trained on AMASS or Xsens data, additional evaluations were conducted on the full DIP-IMU dataset (includes the training set defined by the previous methods) and specifically on DIP-IMU challenging sitting sequences.

The experiment results are presented in Tab. 1. Comparing the performance of models *1)* and *3)* on the DIP-IMU full set, it becomes evident that while synthetic data from AMASS demonstrates some transferability, real mocap data

	DIP IMU Test			
	SIP Err(°)	Ang Err(°)	Pos Err(cm)	Mesh Err(cm)
*1)*AMASS	23.80	8.25	6.04	7.18
*2)*AMASS+DIP	14.41	5.90	5.03	6.05
*3)*Xsens(DynaIP)	17.31	7.66	5.79	7.01
*4)*Xsens+DIP(DynaIP*)	**13.67**	**5.83**	**4.84**	**5.82**
	DIP IMU Full			
	SIP Err(°)	Ang Err(°)	Pos Err(cm)	Mesh Err(cm)
*1)*AMASS	24.50	8.16	6.87	7.82
*3)*Xsens	**18.98**	**7.85**	**6.73**	**7.80**
	DIP IMU Sitting			
	SIP Err(°)	Ang Err(°)	Pos Err(cm)	Mesh Err(cm)
*1)*AMASS	40.72	11.98	13.85	14.56
*3)*Xsens	**29.28**	**10.76**	**12.52**	**13.60**

Table 1. The performance comparison with different training data settings on our model.

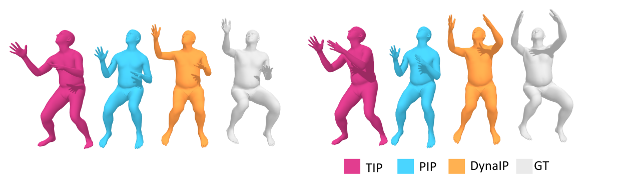

Figure 4. Qualitative comparisons on DIP-IMU [14] test set.

from Xsens exhibits significantly better performance and generalization ability, resulting in a 22% reduction in the SIP error. This improvement is particularly pronounced in the DIP-IMU sitting sequence, where there is a 28% reduction in the SIP error. Additionally, when the DIP training set is integrated into our model training, two noteworthy observations emerge: First, the inclusion of the DIP training set enhances the performance of models *1)* and *3)*. Second, a comparison between models *2)* and *4)* reveals a clear advantage in using a more comprehensive unified real mocap dataset over the virtual-to-real scheme. The additional diversity and realism provided by the real mocap data contribute to enhancing the model's generalization capabilities.

Comparing performance of our method and previous virtual-to-real SOTAs. We further compare the results trained on the unified real mocap dataset with previous virtual-to-real transfer approaches. In Tab. 2, when trained with the DIP-IMU training set, DynaIP* demonstrates an 8% reduction in SIP error and a 19% reduction in global pose error compared to PIP [46], highlighting the accuracy and robustness of our approach. Furthermore, DynaIP, which only uses Xsens data, achieves comparable performance to TransPose [45], which employs DIP-IMU for fine-tuning. This illustrates that having access to abundant real inertial mocap data is beneficial for generalizing to new subjects and unconstrained motions. Fig. 4 provides a visual comparison of our model's performance on the DIP-IMU test set against state-of-the-art methods trained with the AMASS and DIP-IMU. We specifically focus on challenging poses within the DIP-IMU test set, particularly hand-raising motions with relatively low velocities. In these scenarios, our method shows superior performance, effectively capturing the nuances of these motions where the utilization of acceleration is crucial, highlighting the excellence of subtle motion dynamics understanding.

	DIP IMU Test			
	SIP Err(°)	Ang Err(°)	Pos Err(cm)	Mesh Err(cm)
DIP [14]	17.10	15.16	7.33	8.96
TransPose [45]	16.68	8.85	5.95	7.09
TIP [16]	16.20	9.17	5.49	6.61
PIP [46]	15.02	8.73	5.04	**5.95**
DynaIP	17.43	8.90	5.93	7.71
DynaIP*	**13.78**	**7.07**	**4.98**	5.99

Table 2. The performance comparison between our model with the SOTAs reported in their papers on DIP-IMU [14] test set. For a fair comparison, we transform each result into a local rotation representation, consistent with the PIP [46].

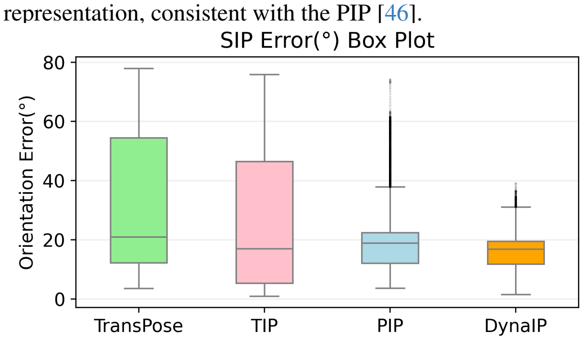

Figure 5. Qualitative results of SIP error box plot for three competing methods and DynaIP on Natural Motion [7] dataset.

4.2. Overall Performance Comparison on the Unified Inertial Mocap Data

For fair and meaningful evaluation, we retrained previous models [14, 16, 45, 46] using Xsens data and evaluated their performance on both the Xsens and DIP-IMU test sets, and compared their results to our model's performance.

The evaluation results in Tab. 3 demonstrate the superior performance of our method compared to competing methods on various datasets. Notably, our model achieves a relative 28% reduction in SIP error on the Natural Motion dataset and an 18% reduction in global pose error on the CIP dataset compared to the state-of-the-art structure PIP [46]. These significant improvements in performance highlight the robustness and generalization capability of our model, which can be attributed to our effective velocity estimation strategy and part-based modeling modules.

Fig. 5 presents a box plot of the SIP error for various models [16, 45, 46] and our model on the Natural Motion dataset [7]. It's worth noting that our model not only achieves the lowest maximum SIP error but also has fewer outliers, indicating that our method consistently produces more robust performance across various motions. This further emphasizes the effectiveness of our approach in handling challenging and diverse pose estimation tasks. In Fig. 6, we provide a visualization comparison between our

	DIP IMU			AnDy			UNIPD			CIP			Natural Motion		
	SIP Err(°)	Ang Err(°)	Pos Err(cm)	SIP Err(°)	Ang Err(°)	Pos Err(cm)	SIP Err(°)	Ang Err(°)	Pos Err(cm)	SIP Err(°)	Ang Err(°)	Pos Err(cm)	SIP Err(°)	Ang Err(°)	Pos Err(cm)
DIP [14]	21.71	10.14	7.92	11.39	5.73	4.34	12.02	5.47	4.22	19.13	8.61	6.86	33.43	11.91	13.33
TransPose [45]	22.53	10.28	8.42	12.15	6.29	4.91	15.11	6.05	4.82	20.06	8.75	6.86	30.62	11.26	11.99
TIP [16]	19.22	8.94	6.91	10.11	4.55	3.56	9.85	4.06	2.78	13.05	5.67	4.30	22.06	7.90	7.92
PIP [46]	17.62	8.33	6.21	9.49	4.09	**3.29**	8.90	3.59	2.66	12.68	5.52	4.12	19.62	7.49	6.93
DynaIP	**17.31**	**7.66**	**5.79**	**8.93**	**3.45**	3.41	**7.29**	**2.77**	**2.21**	**11.42**	**4.54**	**3.69**	**15.78**	**7.18**	**5.83**

Table 3. Evaluation results of the state-of-the-art models on DIP-IMU [14], AnDy [25], UNIPD [9], CIP [29] and Natural Motion [7] when trained only with real inertial mocap data. All models run in real-time setting.

Figure 6. Qualitative comparisons of different model from CIP [29] (Left) and Natural Motion [7] (Right) test set.

	DIP IMU		CIP		Natural Motion	
	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)
①Baseline	15.26	6.17	13.18	5.13	33.22	11.20
②w/o Part	14.97	6.02	13.00	4.86	31.62	9.60
③w/o Vel	14.87	6.11	12.54	4.89	29.15	10.49
DynaIP*	**13.67**	**5.83**	**11.67**	**4.63**	**18.88**	**8.03**

Table 4. The ablation study on pseudo-velocity estimation and the part-based modeling approach.

reconstructed poses and those inferred by state-of-the-art methods using selected frames from CIP and Natural Motion datasets. Our model accurately maintains the sitting or standing pose, even during arm movements or prolonged periods of crossed legs. In contrast, methods like TransPose [45] and PIP [46] intermittently oscillate between sitting and standing poses. The superior performance of our model in these instances can be attributed to two factors: firstly, the retention of root vertical velocity information, which helps capture motion pattern transitions; and secondly, the effectiveness of our part-based modeling approach. This approach proves particularly advantageous for un/rare-seen poses, as it is less influenced by imbalanced training distributions and better at resolving ambiguities.

4.3. Ablations on the Components of our Model

To evaluate the effectiveness of key components in our model, we compared three additional variants with our

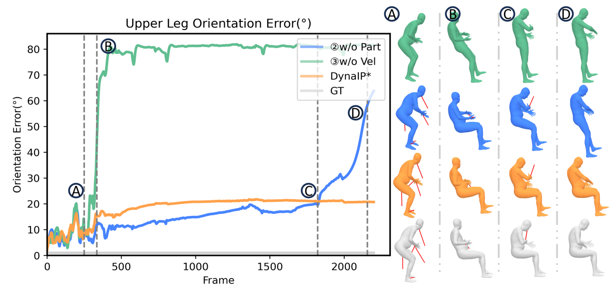

Figure 7. Visualization of upper leg orientation error over time on a test sequence from Natural Motion [7].

method: ①Baseline: a naive LSTM network with RNN-initialization [46]; ②w/o Part: same as DynaIP* but without partition modeling; ③w/o Vel: same as DynaIP* but without learning velocity. The results of these variants on CIP, DIP-IMU, and Natural Motion in Tab. 4 demonstrate a clear increase in testing errors as each component is progressively removed. This trend emphasizes the significance of both pseudo velocity learning and part-based modeling for achieving robust human pose estimation. These components work in synergy to improve the overall performance of our model. The visualization of a test sequence from Natural Motion in Fig. 7 highlights the performance differences between the variants. Variant③ initially struggles to perform a sitting-down motion, while variant②, which includes velocity modeling, can perform the sitting motion but experiences gradual drift over time due to sensor noise. In contrast, our complete model (DynaIP*) combines both velocity and part-based modeling, resulting in accurate and robust pose estimation throughout the sequence. This demonstrates the effectiveness of our approach in handling complex and dynamic motions.

5. Conclusions

This study focuses on improving the robustness and accuracy of learning-based HPE utilizing sparse inertial sensors. The methodology involves the integration of real-world motion capture data from diverse skeleton formats. It employs pseudo-velocity as an intermediary and introduces a novel part-based approach. This method effectively leverages acceleration data and local body correlations, leading to enhanced pose estimation results. As evidenced in the case studies, this method significantly outperforms existing techniques on all performance metrics across five datasets.

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020. 4

[2] Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems(NeurIPS)*, 35:27414–27426, 2022. 1

[3] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 3501–3510, 2022. 1, 2

[4] Julia F. Christensen, Andres Fernandez, Rebecca Smith, Georgios Michalareas, Sina H. N. Yazdi, Fahima Farahi, Eva-Madeleine Schmidt, Nasimeh Bahmanian, and Gemma Roig. Emokine: A kinematic dataset and computational framework for scaling up the creation of highly controlled emotional full-body movement datasets. https://doi.org/10.5281/zenodo.7821844, 2023. 3, 6

[5] 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, 2022. 4

[6] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 682–692, 2023. 1

[7] Jack H Geissinger and Alan T Asbeck. Motion inference using sparse inertial sensors, self-supervised learning, and a new dataset of unscripted human motion. https://doi.org/10.7294/2v3w-sb92, 2020. 3, 4, 6, 7, 8

[8] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, pages 14676–14686, 2021. 2

[9] Mattia Guidolin, Emanuele Menegatti, and Monica Reggiani. Unipd-bpe: Synchronized rgb-d and inertial data for multimodal body pose estimation and tracking. https://doi.org/10.17605/OSF.IO/YJ9Q4, 2022. 3, 4, 6, 8

[10] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4318–4329, 2021. 1

[11] Vladimir Guzov, Torsten Sattler, and Gerard Pons-Moll. Visually plausible human-object interaction capture from wearable sensors. *arXiv preprint arXiv:2205.02830*, 2022. 4

[12] Sachini Herath, Hang Yan, and Yasutaka Furukawa. Ronin: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3146–3152. IEEE, 2020. 5

[13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 9924–9935, 2022. 4

[14] 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 (TOG)*, 37(6):1–15, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[15] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Proceedings of European Conference on Computer Vision(ECCV)*. Springer, 2022. 1

[16] 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)*, pages 1–9, 2022. 1, 2, 3, 4, 7, 8

[17] 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 *Proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, pages 11510–11520, 2021. 1, 2

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

[19] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies(IMWUT)*, 4(3):1–29, 2020. 3, 4

[20] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 12375–12384, 2021. 2, 5, 6

[21] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph. (TOG)*, 33(6):220–1, 2014. 3

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 4

[23] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual mark-

ers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 534–543, 2023. 2

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, pages 5442–5451, 2019. 2, 4, 6

[25] Pauline Maurice, Adrien Malaisé, Clélie Amiot, Nicolas Paris, Guy-Junior Richard, Olivier Rochel, and Serena Ivaldi. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *The International Journal of Robotics Research(IJRR)*, 38(14):1529–1537, 2019. 3, 4, 6, 8

[26] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 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(CHI)*, pages 1–12, 2023. 1

[27] Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. Humor: Human motion representation using topology-agnostic transformers for character animation retargeting. *arXiv preprint arXiv:2305.18897*, 2023. 4

[28] Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017. 4

[29] Manuel Palermo, Sara M Cerqueira, João André, António Pereira, and Cristina P Santos. From raw measurements to human pose-a dataset with low-cost and high-end inertial-magnetic sensor data. *Scientific Data*, 9(1):591, 2022. 3, 4, 6, 8

[30] Ling Pei, Songpengcheng Xia, Lei Chu, Fanyi Xiao, Qi Wu, Wenxian Yu, and Robert Qiu. Mars: Mixed virtual and real wearable sensors for human activity recognition with multidomain deep learning model. *IEEE Internet of Things Journal*, 8(11):9383–9396, 2021. 2

[31] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free domain adaptive human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4826–4836, 2023. 2

[32] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2337–2347, 2023. 2

[33] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(WACV)*, pages 2956–2966, 2023. 3

[34] Martin Schepers, Matteo Giuberti, Giovanni Bellusci, et al. Xsens mvn: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8):1–8, 2018. 2, 4

[35] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 17038–17047, 2023. 1, 2

[36] Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 45(4):4122–4135, 2022. 2, 5

[37] Ronit Slyper and Jessica K Hodgins. Action capture with accelerometers. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 193–199, 2008. 2

[38] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics(TOG)*, 30(3):1–12, 2011. 2

[39] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4713–4725, 2023. 1, 2

[40] 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(BMVC)*, pages 1–13, 2017. 3

[41] 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(CGF)*, pages 349–360. Wiley Online Library, 2017. 3

[42] Songpengcheng Xia, Lei Chu, Ling Pei, Zixuan Zhang, Wenxian Yu, and Robert C Qiu. Learning disentangled representation for mixed-reality human activity recognition with a single imu sensor. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 2, 3, 4

[43] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems(SenSys)*, pages 391–406, 2022. 2

[44] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum(CGF)*, pages 265–275. Wiley Online Library, 2021. 5

[45] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 3, 4, 6, 7, 8

[46] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 13167–13178, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[47] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in

3d human pose estimation with a split-and-recombine approach. In *Proceedings of European Conference on Computer Vision(ECCV)*, pages 507–523. Springer, 2020. 6

[48] Lei Zhang and Xinbo Gao. Transfer adaptation learning: A decade survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 4

[49] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. 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(CVPR)*, pages 2200–2209, 2020. 2

[50] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pages 7356–7365, 2018. 1

[51] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 1