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

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

This supplementary material provides additional information to complement our main paper. Sec. A outlines the network architecture and training details of our approach. Sec. B offers an overview of the datasets utilized in our research. Sec. C presents extended ablation studies, encompassing various training data settings and further analysis of component design. Sec. D features additional qualitative comparisons between our model and existing state-of-the-art methods. Finally, Sec. E summarizes our work, discusses its limitations, and proposes potential directions for future research.

A. Implementation Details

Network Details Our model architecture comprises distinct sub-networks, each equipped with a Multilayer Perceptron (MLP) layer for initializing hidden states, followed by the Long Short-Term Memory (LSTM) networks. The first two-layer LSTM network is designed for regressing the pseudo-velocity of joints, and the second two-layer LSTM network focuses on the final pose estimation. An additional LSTM layer is used to extract global context from all six sensors, which is shared by all three sub-networks. Notably, the output for each joint is represented as a global 6D-rotation relative to the root joint [14], and this relational approach is also applied to intermediate pseudo-velocity predictions.

Specifically, we directly regress the orientations of 11 joints on the Xsens skeleton that lack IMU measurements. To align with the SMPL model during evaluation (on DIP-IMU) and visualization, we remove one redundant torso joint (labeled as 'L3' in our implementation) and map the remaining predicted results. When integrating with the DIP-IMU dataset, we duplicate the 'Spine1' joint in the SMPL model to correspond with our model's output dimensions.

A difference from prior methodologies in our work is the treatment of joints such as the head, root, forearms, and forelegs, which are directly attached to IMUs. Rather than predicting their orientations, we directly utilize the orientations provided by their respective IMU measurements, ensuring a more straightforward and accurate representation.

Datasets	Motion types	Subjects	Minutes
DIP-IMU [6]	jumping, sitting, walking, lifting arm	10	86
CIP [12]	grabbing, reaching, sitting	10	288
Natural Motion [4]	long time sitting, exercising, walking	17	692
Emokine [1]	upper body motion	1	12
AnDy [10]	walking, kneeling, manipulating loads	13	421
UNIPD [5]	sitting, pointing, bending, walking and jogging	15	162
Real IMU data	-	-	1661
AMASS [9]	-	-	2122

Table 1. Dataset Overview

B. Datasets Details

We use a subset of AMASS [9] to generate synthetic IMU as previous works. Simultaneously, we train and evaluate our model using several publicly available Xsens Inertial Mocap dataset. The following part will briefly introduce these public datasets we used.

1) AnDy [10]: The Inertial Mocap data from AnDy were collected from 13 participants in an industry environment. A total number of 195 trails consist of commonly seen motions like raising arms, walking, bending torso, crouching and etc.. We use the last two subject's data (ID: 9266 and 9857) for evaluation and the others for training.

2) CIP [12]: The sensing data of the CIP dataset were collected from 10 participants for 6 different types of movement sequences. These sequence types involve motions such as factory assembly tasks, office dynamics and random free movements. We select data from the subject 4 and 8 for evaluation and the rest data for training.

3) Natural Motion [4]: The dataset consists of motions from 17 participants, with 13 performing unscripted daily activities like walking, working at a computer, exercising,

*Equal contribution
†Corresponding authors

This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 62273229.

	DIP IMU		AnDy		UNIPD		CIP		Natural Motion	
	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)
*1)*AMASS	23.80	8.25	20.80	7.72	14.98	5.64	25.27	9.08	40.58	13.56
*2)*AMASS+DIP	14.41	5.90	18.79	8.10	14.29	5.49	18.65	7.51	20.72	9.26
*3)*Xsens	17.62	8.33	8.93	3.45	7.29	2.77	**11.42**	**4.54**	**15.78**	**7.18**
*4)*Xsens+DIP	**13.67**	**5.83**	9.17	3.56	7.60	2.83	11.67	4.63	18.88	8.03
*5)*AMASS+Xsens+DIP	13.77	5.92	**8.84**	**3.42**	**7.08**	**2.67**	12.01	4.66	19.89	8.39

Table 2. Performance Comparison on Different Training Data Settings.

	DIP IMU		AnDy		UNIPD		CIP		Natural Motion	
	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)	SIP Err(°)	Ang Err(°)
①Baseline	15.26	6.17	9.89	3.76	9.05	3.24	13.18	5.13	33.22	11.20
②w/o Part	14.97	6.02	9.62	3.85	7.63	2.92	13.00	4.86	31.62	9.60
③w/o Vel	14.87	6.11	9.27	**3.50**	7.61	**2.81**	12.54	4.89	29.15	10.49
④CP	14.64	6.35	10.74	4.14	**7.54**	2.88	12.52	4.87	20.58	**7.90**
⑤DynaIP*	**13.67**	**5.83**	**9.17**	3.56	7.60	2.83	**11.67**	**4.63**	**18.88**	8.03

Table 3. Performance Comparison of Ablation Variants.

etc., and the remaining 4 executing actions within industrial settings. Notably, an essential characteristic of this dataset is the exceptional duration of each raw capture sequence, varying from half an hour up to three hours. This allows for a wealth of extended long-time sitting or standing sequences to be incorporated. Nevertheless, we find some raw data may exhibit certain drift due to prolonged capture times and lack of precise re-calibration. We manually selected 9 (ID: 1, 2, 3, 4, 5, 6, 10, 11 and 13) out of the 13 participants engaged in daily activities and extracted clean data segments to be utilized for training and testing.

4) Emokine [1]: This dataset contains 63 sequences captured from a dancer performing different body movements with emotions like anger, contentment, fear, joy, neutrality, and sadness, this includes a variety of rapid and slow movements of the upper and lower limbs. Since the total duration of this data only amounts to 12 minutes, we have decided to use it solely for training purposes.

5) UNIPD [5]: This dataset captures detailed body poses from 15 participants performing 12 scripted activities in laborious environment, such as walking, sitting, jogging and bending. We use data from last two subjects (ID: 14 and 15) as test set and the rest for training.

The overview of these dataset along with DIP-IMU [6] is shown in Tab. 1

C. Additional Analysis

In this section, we present more experimental results of our proposed method (DyanIP).

Additional Comparison on Unified Mocap Data and Virtual-to-Real (V-to-R) Training Scheme. We extend our analysis by including results from the Xsens test set for the four training settings previously discussed in Sec. 4 on more real mocap datasets. Additionally, we report a extra setting marked as model *5)*: AMASS+Xsens+DIP. This setting involves pretrain on AMASS synthetic data and then retrain with a mix of Xsens and DIP-IMU data.

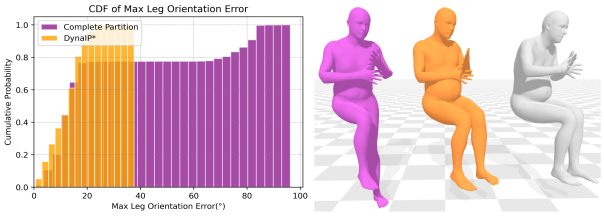


Figure 1. CDF of the largest upper leg orientation error. Complete partition model predicts unnatural result for excessive separation.

The findings, as shown in Tab. 2, reveal that relying solely on AMASS (model *1)*) leads to subpar performance in real-world scenarios. Fine-tuning with a smaller real dataset, such as DIP-IMU, offers limited improvements (model *2)*).

A notable observation is that the performance of models *4)* and *5)* is quite similar, indicating that pre-training on AMASS does not significantly affect the outcome. This similarity may stem from the test set's focus on everyday motions, which are well-represented in the current real-world training data. However, the gap between virtual and real IMU data remains a concern and could be further investigated. Virtual data's potential may lie in capturing more diverse and uncommon motions, such as those found in sports like basketball, swimming, parkour, or specific dance styles, which are not adequately covered by existing real datasets.

Additional Ablation Study. We show the comparison results of ablation models on more datasets in Tab. 3. Along with aforementioned variants (①, ②, ③), We additionally

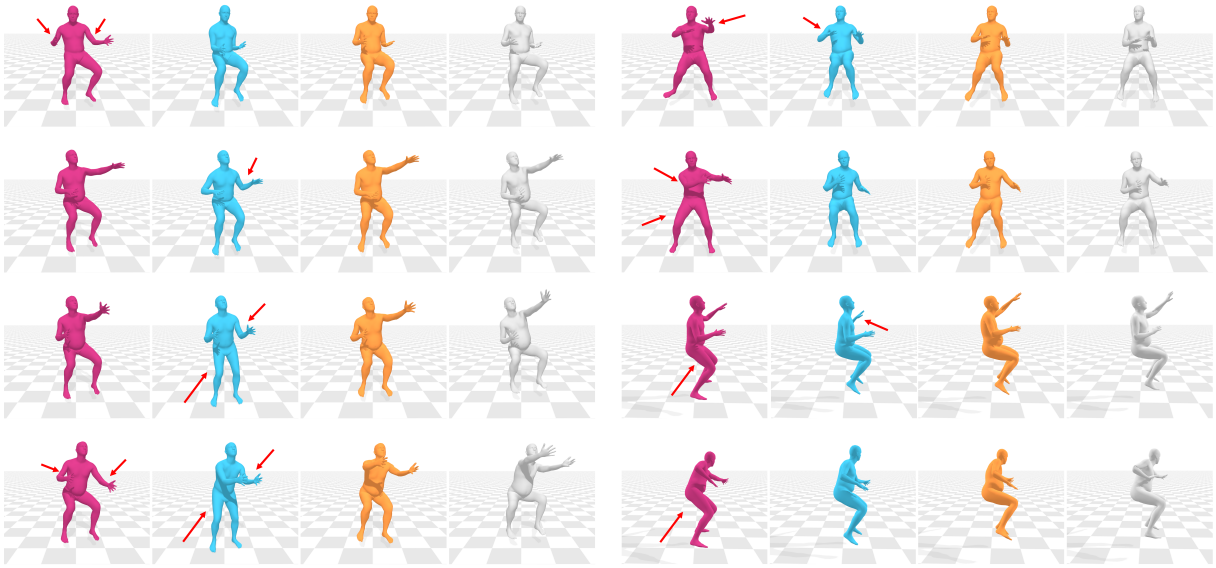


Figure 2. Additional qualitative results on DIP-IMU with previous virtual-to-real SOTAs. From left to right: TIP [8], PIP [15], DynaIP and GT.

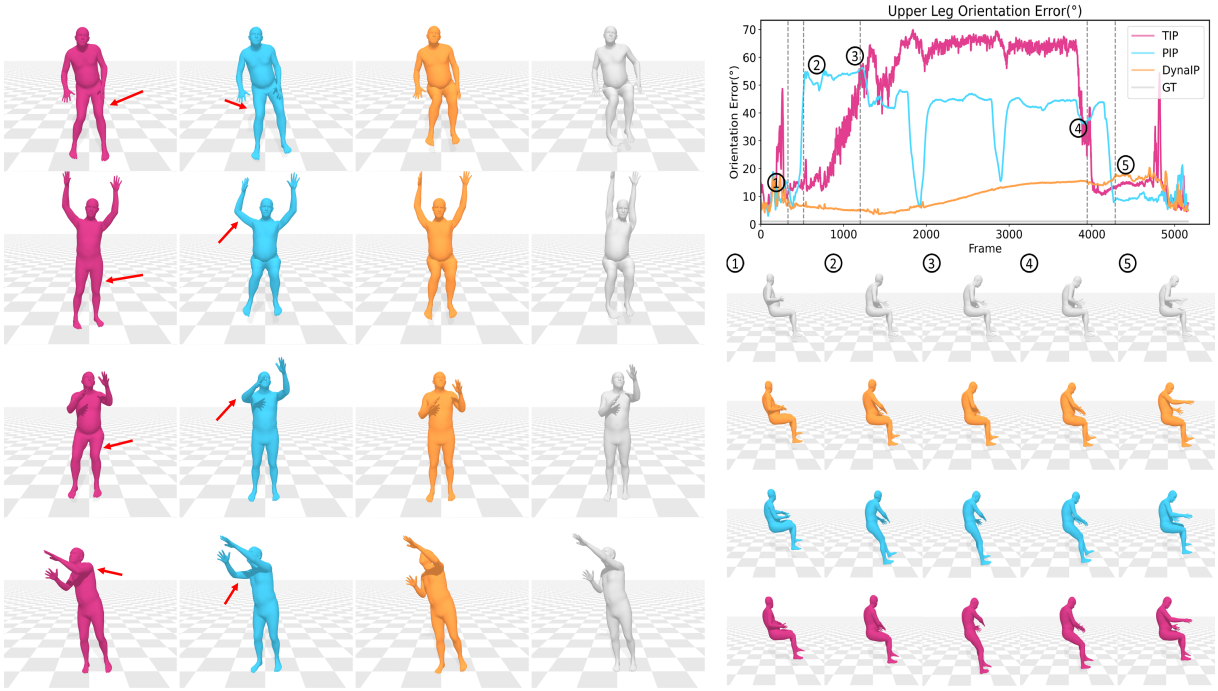


Figure 3. Additional visualization when models all trained using Xsens mocap data. Left: results from CIP sequence, Right: results from Natural Motion sequence.

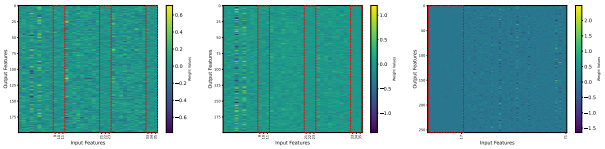
evaluate a variant of our model denotes as "④ Complete partition (CP)", which further divides body into left and right sides: four limb groups and one torso group, as it is interesting to examine whether more local parts can bring better performance.

The results, detailed in Tab. 3, demonstrate that DynaIP* maintains the best overall performance on the most test set, particularly on the DIP-IMU, Natural Motion, and CIP datasets. These datasets include more complex motion scenarios, where the unique benefits of pseudo-velocity estimation and part-based modeling are most evident.

In our experiments, the complete partition model, marked as ④, reveals a subtle decline in SIP error performance, as indicated in Tab. 3. Additionally, we present

Dataset	Method	SIP Err(°)	Ang Err(°)	Pos Err(cm)	Mesh Err(cm)
DIP_IMU	PIP	15.02	8.73	5.04	**5.95**
	DynaIP†	**14.11**	**7.00**	**4.97**	5.97
Total Capture	PIP	12.93	12.04	5.61	6.51
	DynaIP†	**12.42**	**11.06**	**5.11**	**5.79**

Table 4. The performance comparison on DIP-IMU and Totalcapture. DynaIP† ia trained on AMASS and DIP-IMU, same as PIP.



(a) DynaIP* (b) DynaIP*-Position (c) PIP-official

Figure 4. Visualization of first linear layer weights, Accs on marked cols while the rest are Oris. (PIP Accs on col 0 to 17)

the Cumulative Distribution Function (CDF) of the largest orientation error between the two upper legs on the Natural Motion test set in Fig. 1. While the introduction of more sensors and joint groups theoretically enhances the model's capacity to capture local motion information, oversegmentation of the body into numerous parts may inadvertently compromise the inter-connectivity between the subnetworks. This disconnect is not entirely compensated by the incorporation of low-dimensional global context. As a result, this can lead to inconsistencies in the pose estimates, manifesting as increased errors in complex motions, as depicted in Fig. 1. Our observations suggest that while segmenting the body into local parts enhances local correlation, an excessive division may inadvertently weaken the model's ability to maintain global coherence, thus impacting its performance negatively in certain challenging scenarios.

Additional Comparison on Model performance. To isolate data impact and assess model performance, we provide a additional comparison with our model marked as DynaIP† using a data set consistent with PIP (combined with AMASS and DIP data) in Tab. 4. The performance enhancement highlights our architecture's efficiency.

D. Additional Qualitative Results

In this section, We show more visualization results of our method compared to the state-of-the-art methods

Additional Qualitative Comparison with the V-to-R SOTAs. We present a comprehensive qualitative comparison between our method, DynaIP*, and previous state-of-the-art models using virtual-to-real (V-to-R) training scheme, such as TIP [8] and PIP [15]. Additional frames from the DIP-IMU dataset are visualized in Fig. 2, where our approach, DynaIP*, displays marked improvements over the previous V-to-R models in various situations, highlighting the advantages of incorporating real inertial mocap data.

This comparison points out that past V-to-R models mainly focused on real IMU datasets with SMPL ground truths. As a result, the proportion of 'real' components in these models was limited, which constrained their overall performance. By contrast, as shown in Fig. 2, our approach is built on a wide array of real inertial data, which enhances the model's adaptability and precision across various motion contexts.

Additional Qualitative Comparison on network structure. This visualization presents a deeper qualitative analysis that compares the performance of our model, DynaIP, with state-of-the-art models (TIP [8] and PIP [15]) trained solely on Xsens data. In Fig. 3, we exhibit more instances to highlight the enhanced capabilities of our model.

On the left side of the image, DynaIP's shows better ability to capture the movements of the upper leg and arm. This precision can be attributed to our model's advanced network structure, which effectively integrates dynamic motion information and spatial correlations within the body.

Turning to the right side of the image, our model demonstrates its robust in tracking a sitting pose, even when it involves complex hand movements. This scenario often poses a challenge for other models due to their limited capacity in localized region modeling. However, DynaIP's part-based approach allows for a more focused and accurate interpretation of motions within each region.

Effect of using velocity as intermediate outputs. Previous works, such as DIP, identified a trend where networks tend to discard much of the acceleration, and introduced an auxiliary task to reconstruct the acceleration for alleviating this problem. To better utilize acceleration, we first regress the pseudo velocity as an intermediary output, supervised by the ground-truth velocity obtained from the human pose. Compared with PIP/Transpose that uses leaf joints positions as intermediate outputs, our method has certain advantages in terms of using acceleration. Since noise in acceleration easily diverges with quadratic integration, it's more difficult to capture the spatial relationship in noisy acceleration. Instead, the network learns to infer the position more from the orientation input due to human kinematic relationship. Following DIP's explanation, we visualize the weights of the first linear layer of the network in the Fig. 4, where both the weights of PIP and ours-w.position in the corresponding dimension of the acceleration input is almost zero.

Tracking robustness of long-time sitting motion. In assessing the robustness of our model, DynaIP, particularly in prolonged motion scenarios, we conducted a qualitative analysis using a real sitting sequence from the Natural Motion dataset. This sequence, lasting for 14 minutes, is showcased in Fig. 5 and exemplifies the model's capability in long-time motion tracking.

The accurate modeling of a real seating posture is notably challenging. It entails not only managing the inherent

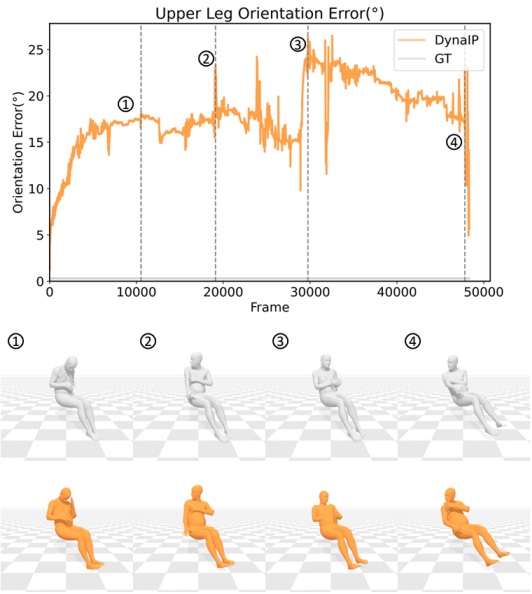


Figure 5. We've selected four timestamps with relatively high upper leg orientation errors from a 50,000 frame sitting sequence and have visualized their respective mesh representations with ground truth.

noise in IMU readings but also addressing the subtle and occasional body movements humans exhibit in unscripted situations, as opposed to a perfectly stationary posture. Our model's part-based approach plays a crucial role in its success in this context.

Furthermore, to reinforce our findings and demonstrate the practical applicability of our model, we have also evaluated DynaIP in real-world scenarios using an inertial mocap device. Please refer to the supplementary videos for more details.

E. Discussions and Future Works

To our knowledge, we are the first to utilize pre-existing publicly available inertial mocap datasets in learning-based sparse inertial mocap with the global orientation mapping strategy across skeleton formats. Despite our efforts to expand the training data with real IMU readings, we acknowledge a challenge: the model's accuracy dips when it encounters poses that are underrepresented in the training set. To address this, we recognize the potential of integrating additional datasets featuring real IMUs, which cover a broader spectrum of motion types. Recently, there comes several multi-modality datasets such as [2, 3] that make extensive use of IMU data as part of their sensor modality. Regrettably, raw IMU data synchronized with these resources aren't presently available for public access. We are of the strong belief that, once made available, these datasets could offer immense value for inertial sensor-based motion capture.

At present, our system does not incorporate improvements in global root translation and operates based on the assumption of flat ground conditions. This limitation stems from the inherent tendency of purely inertial-based global trajectory estimation methods to experience drift. Additionally, the accuracy of our current methods for estimating translation can be adversely impacted by inaccuracies in pose prediction, a problem also linked to the susceptibility of inertial-based systems to drift. Looking ahead, we aim to tackle the issue of long-term drift in IMU-based solutions by integrating environmental constraint information or by incorporating additional sensor modalities.

In summary, our research points towards exciting avenues for future work in sparse IMU-based human pose estimation. Expanding the dataset diversity, dividing distinct body regions, and enhancing velocity estimation are key areas that hold the potential to advance this field significantly. As we continue to explore these possibilities, we would also be committed to using widely used smart devices containing IMUs, such as VR [7, 13], mobile phones, smart-watches [11], etc., to develop robust and accurate motion capture solutions that can thrive in a variety of real-world applications.

References

[1] 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. 1, 2

[2] 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. 5

[3] 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. 5

[4] 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. 1

[5] 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. 1, 2

[6] 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

[7] 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. 5

[8] 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. 3, 4

[9] 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. 1

[10] 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. 1

[11] 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. 5

[12] 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. 1

[13] 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

[14] 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

[15] 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. 3, 4