Physical Shape-aware Poser (PSP): Tracking Various-shaped Human Motion From Sparse Inertial Sensors

Lu Yin 30920241154581[1], **Shifan Jiang 30920241154567**[2], **Ben Zhang 36920241153278**[2]

[1]Informatics Class
[2]AI Class

Abstract

IMU-based human motion capture is a task with great potential. Existing methods show effective results using a template-shaped adult body model. However, when applied to a subject with a large difference in size(such as a child), motion capture accuracy will decrease significantly. To this end, we propose a sparse inertial motion capture solution that achieves motion tracking of various-shaped humans, from small kids to adults. Including a learning-based method to convert motion data (e.g., joint acceleration and velocity) between SMPL and non-SMPL human shapes. Given the shape information of a body, by converting signals captured from 6 IMUs to a SMPL-shaped body, we then apply the template-based pose estimator and output motion-related data such as local pose and joint velocities. To further obtain global motion for the target human shape and ensure its physical plausibility, we utilize a shape-aware optimization strategy, which recovers motion to target characters and improves translation accuracy remarkably. For validation, we collected a dataset with various-shaped subjects. Experiments demonstrate that our approach outperforms state of the art on motion tracking accuracy and produces plausible results for various body shapes.

Introduction

Human motion capture (MoCap) has seen a surge in research in the past few decades(Nagymáté and Kiss 2018; eas 2021; Yi, Zhou, and Xu 2021), focusing primarily on optical-based methods. Both marker-based and camera-based tracking approaches have achieved high accuracy. The main objective of these tasks is to reconstruct the 2D/3D pose and global translation of the human body. However, because of the diversity in body shapes, the reconstructed pose is often expected to be compatible with the individual's shape. Targeting parametric body models, many researchers also consider shape estimation as one of the goals of motion capture. For example, using images and videos, (Kocabas, Athanasiou, and Black 2020; Wang and Zhang 2023; **?**) extract human body shape information from visual data, enabling motion capture for various body shapes.

Although effective, optical-based methods are highly dependent on cameras and require carefully adjusted surroundings. In recent years, motion capture using wearable sensors has gained widespread attention due to its advantages in portability and lower cost. Compared to camera-based systems, inertial motion capture is not restricted by environmental factors such as lighting and occlusion. Such approaches use only six IMU sensors attached to key body parts (legs, arms, head, and waist) to track human motion in real time.

However, in a lack of real-world data, state-of-the-art systems are mostly trained on synthesized IMU data from a template-shaped human body with SMPL parameters. We observed that although they are able to perform efficient motion tracking on many adult subjects, their accuracy in pose and translation significantly decreases when applied to subjects that differ greatly from the template shape (such as small children). This problem occurs especially in acceleration-dominated poses, e.g., raising hands or legs, for when performing the same poseure, different-sized characters demonstrate very similar joint orientations.

We assume that with the same pose (joint rotation), the IMU measured orientation is also the same for different body shapes. Thus, *1) by "correcting" IMU measured accelerations, we can estimate the pose for any given body shape*. For translation regressing, we follow the baseline method in (Yi et al. 2022) and (Yi, Zhou, and Xu 2024) by estimating joint velocities and foot-ground contact probabilities and apply a physics-based motion optimization. In this process, *2) estimated template-body velocities need to be restored to the target body shape* and on top of the recovered pose and velocity, *3) various-shaped subjects call for a size-aware physical optimization technique to ensure the physical correctness of captured motion*.

To this end, we propose Physical Shape-aware Poser (PSP), the first sparse IMU real-time motion capture solution that can handle subjects with various body shapes. To address *1)* and *2)*, we propose a learning-based kinematic signal retargeting method. This algorithm is for estimating template-shaped acceleration from target-shaped acceleration, and in reverse, target-shaped joint velocity from template-shaped velocity. Since the human body shape expressed in weighted models such as SMPL is often a parameter from a PCA latent space, its representational range is limited to the latent space composed of adult subjects. We scale the human skeleton in AMASS to obtain bodies shaped differently from the parametric space. Then, by adding physical constraints to the global translation, we obtained motion of the scaled hu-

man bodies, along with the "unscaled" AMASS motion data. We train our data retargeting model on this dataset. To address *3)*, we utilize proportional derivative(PD) rules on a shape-aware dynamic model. Compared to previous methods, our PD controller takes shape-differed joint positions into consideration and is able to perform optimization on various-shaped subjects.

To validate the effectiveness of our method, we collected a real-world dataset from 3 subjects with heights 139cm, 153cm and 180cm. Experimental results demonstrate that our method is applicable to various body sizes and significantly outperforms state-of-the-art solutions in terms of accuracy on both synthesized and real data. In summary, our main contributions are as follows.

- Physical Shape-aware Poser (PSP): The first real-time motion capture solution using sparse IMU that achieves motion tracking of various-sized subjects, from small kids to adults.
- A learning-based method to convert human motion data (e.g., joint acceleration and velocity) between a template-shaped adult body and human bodies with various sizes.
- A shape-aware physical optimization strategy that recovers physically plausible motion from different-shaped human motion data, which also improves translation accuracy significantly.

Related Work

Optical-based Motion Capture

Optical-based motion capture has a long history in research. Commercial-level systems like (Point 2011; Nagymáté and Kiss 2018), use multiple cameras and dense marker points to achieve the golden standard human mocap. With the rapid development of deep learning, studies such as (Sun et al. 2019; Kocabas et al. 2021; Zhao et al. 2024) based on single camera input have made remarkable progress. Using RGB-D data (Kehl et al. 2016; Yu et al. 2021) and multiple view images (Wu et al. 2021; Ye et al. 2022; Zhang et al. 2021) has also been a research focus in this task. Furthermore, motion tracking studies often use parametric 3D human body models (Loper et al. 2023; Pavlakos et al. 2019) as output target, as capturing statistics of human shape from images(Wang and Zhang 2023; Xu et al. 2020) or videos (Zhang et al. 2023; Kocabas, Athanasiou, and Black 2020) can result in not standard but various shaped target avatars when representing different subjects.

However, the above methods all rely on cameras. Although optical methods have become the golden standard for human motion capture, their cost and experimental environment requirements make them unsuitable for everyday life.

Motion Tracking from Sparse IMUs

IMU sensors can be used to measure acceleration and rotation. Due to their portability and independence from cameras, IMU-based motion capture solutions have received widespread attention in recent years. Von Marcard et al. (2017) proposed a method using six IMUs for motion capture, but it operates offline and is not suitable for real-time applications. Huang et al. (2018) developed a method using bidirectional RNNs to reconstruct human body posture in real-time with six IMUs, although it only estimates local pose without considering global translation.

To achieve real-time motion capture, Yi, Zhou, and Xu (2021) introduced the TransPose model, which captures both human pose and global translation simultaneously. An extended version of this system was later proposed by Yi et al. (2022), incorporating a physical dynamics optimization module to enhance motion capture accuracy and ensure physical plausibility of motion. By introducing a dual PD controller, PIP is able to gain global control of the character and is the first to leverage explicit physics-based optimization into spare IMU-based motion capture. However, this technique only works on a template-shaped character. Not long after, Jiang et al. (2022b) demonstrated the application of Transformers for motion capture using sparse IMUs, also generating terrain maps during the motion process. Some studies also focus on practical application. For instance, Zuo et al. (2024) introduce a loose-wear jacket with IMU integrated for wearing comfortableness. In VR/AR domains, researchers such as Jiang et al. (2022a) and Ponton et al. (2023) use IMUs in VR devices to track human movement. Most recently, to cope with undeterministic acceleration measurements, Yi, Zhou, and Xu (2024) introduced non-inertial root frame and fictitious force modeling in inertial-based motion capture.

Although these approaches show great potential in the task of inertial-mocap, when the mocap target body shape differs from the adult body shape used in training, the estimating error becomes apparent. Additionally, for algorithms with a physical optimization strategy, the proposed technique is also limited to a template-shaped character.

Plan

Our main contributions are as described in the Introduction section. Our proposed plan include various-shaped motion tracking data synthesize, design and training of our motion data retargeting neural network and our shape-aware physical optimization strategy.

We also plan to conduct thorough experiments as follow:

- Comparision with state-of-the-art (Yi et al. 2022; Yi, Zhou, and Xu 2024) on both synthesized data and real-world data of our own collecting using the noitom PN studio system.
- Evaluation on our main modules acceleration retargeter and velocity retargeter, as well as the physical optimization module.
- Other related experiments.

Method

Our goal is to estimate real-time human motion from 6 Inertial Measurement Units (IMUs) placed on the leaf joints (forearms, lower legs, head) and the root joint (pelvis). The inputs to our system are IMU measurements including accelerations, angular velocities, and orientations. We also input subject height to obtain human shape information in the

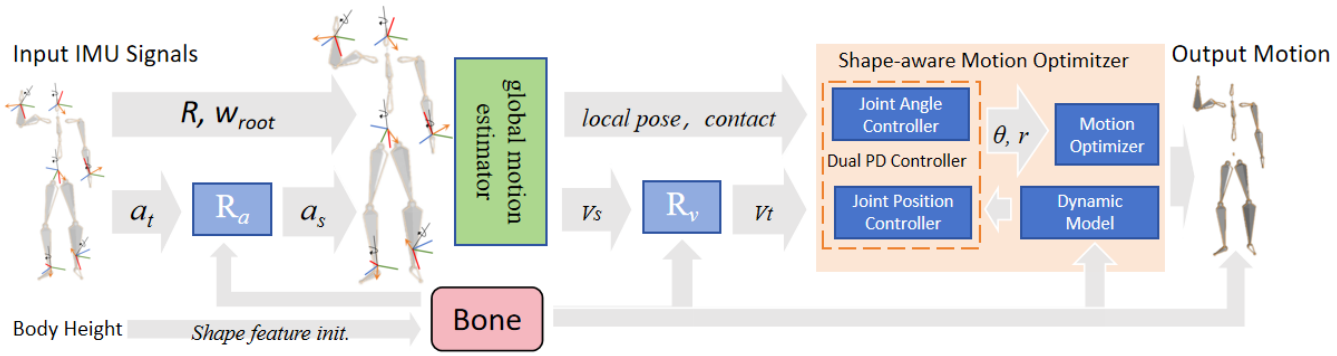

Figure 1: Pipeline of our method: input IMU signals of a non-SMPL shaped human body(e.g. a child), are first retargeted to SMPL-space IMU measurements. Then, the global motion estimator based on PNP estimates local pose, foot-ground contact probability and joint velocity of the template SMPL body. A second retargetor R_v is applied to transform the estimated velocity back to target body shape. Our shape-aware motion optimitzer utilizes a dynamic model corresponding to the target human body to obtain physically corrected motion.

process. The output of our system are the local pose and global movements. In the following, we first introduce our learning-based motion signal retargeter and then explain our shape-aware physical optimazation module. See Fig. 1 for an overview.

Leaning-based Kinematic Signal Retargeting

Shape-aware Nerual Retargeter Consider two individuals with significantly different body sizes (an adult and a child), we can divide their joint kinematic data into two categories: rotation-related signals and position-related signals. When performing the same pose, rotation-related signals are also the same, such as angular velocity and leaf IMU orientation. However, due to their difference in skeleton, position-related signals such as joint velocity and acceleration will vary. When feeding the IMU input of a child to state-of-the-art networks trained on adult data, the estimated motion inevitably exhibits error. To address this issue, we first use an RNN-based kinematic signal retargeting network to regress the input position-related signal(IMU accelrations) to the corresponding acceleration for the template SMPL body that produces the same local pose. Input of this network R_a, is the target subjects' IMU acceleration and shape feature. The input to R_a consists of the IMU accelerations $\mathbf{a} \in R^{6\times3}$ and the target human shape $\mathbf{b} \in R^{23\times1}$: $\mathbf{a}_s = R_a(\mathbf{a}_t, \mathbf{b})$. Where $\mathbf{b}$ is the length vector of the 23 bones in the SMPL body model, t and s denote the target and SMPL. We compute the target body shape vector by adding a scale to the template SMPL body computed with the height input:

$$\mathbf{Bone} = \left(\frac{height}{height_s}\right) \cdot \mathbf{Bone}_s \tag{1}$$

Next, using the IMU input angular velocities and rotations, along with the regressed template-shaped accelerations, we first transform these global coordinates to the root joint coordinate system. For IMU data that do not contain shape information, we split the motion reconstruction task into local pose estimation and global translation estimation, as in previous works.

For local pose estimation, we first apply a neural auto-regressive estimator that learns the physically correct fictitious forces arising from modeling the non-inertial human root's coordinate frame, and outputs the fictitious force acceleration a_{fic}, following (Yi, Zhou, and Xu 2024). Following (Yi, Zhou, and Xu 2021; Yi et al. 2022; Yi, Zhou, and Xu 2024), we first estimate leaf joint positions $\mathbf{p}_{RL}$, then all joint positions, and finally all joint rotations. This is achieved through three Long Short-Term Memory (LSTM) recurrent networks with jump connections on the IMU input. The resulting joint rotations correspond to the local pose, which is the same for both the SMPL body and the target human body.

For global translation estimation, we follow (Yi et al. 2022; Yi, Zhou, and Xu 2024) to regress joint velocities and foot-ground contact probabilities. However, while body shape changes do not affect foot-ground contact, joint velocities differ for individuals with different bone lengths. A second shape-related kinematic signal retargeting network R_v is used to regress the predicted joint velocities for a SMPL-template human model back to the joint velocities of the target body shape, as shown in Figure 1. Symmetrically to R_a, the input to velocity retargeter is $v_s R^{24*3}$, $\mathbf{Bone}$, and the output is the target human's joint velocities $v_s R^{24*3}$.

Training Data Synthesize To train our kinematic signal retargeters, we need motion data of SMPL-shaped adults and paired motion data of various-shaped bodies. We acquired the former from AMASS dataset whereas the latter requires synthesization. AMASS dataset contains motion data for adults and their corresponding SMPL body shapes, which is used to train state-of-the-art methods as well as out pose estimator. To enrich body shapes, we scale the original AMASS adult skeletons within a range of 0.5 to 1.2, resulting in skeletons with heights ranging from 0.8m to 2.0m, covering a much wilder range of human bodies, including preteen children. Next, to calculate global translation of the scaled bodies, we add physical constraints based on the following conditions. In any given motion frame i, velocity of

the ground-contact point Δv_i:

$$\Delta v_i = v_i - v_{i-1} \tag{2}$$

where

$$v_i = \mathrm{FK}(Bone, pose_i) \tag{3}$$

and we calculate the root joint translation $tran_i$ by adding translation of frame i to the original translation $tran_{i-1}$. However, when there is no detected foot-ground contact, we use ground truth translation to calculate root joint velocity, and

$$tran_i = tran_{i-1} + \Delta tran_i^{GT} \tag{4}$$

Note that this ensures, in actions where both feet are off the ground, the root joint acceleration follows physical laws (e.g., when jumping into the air, MMM and SSS should land at the same time with the root node acceleration equal to gravity, meaning they should jump to the same height). This does not prevent the motion data we generate from being visually reasonable and suitable for training motion tracking algorithms.

For foot-ground contact (global movement), we select two mesh vertices on both feet of the SMPL model, located at the toe and heel positions. If the movement distance of one vertex in a frame is less than 0.5cm, we consider the joint to be in contact with the ground (global rest position).In this way, we obtain motion data for bodies with different shapes in the AMASS dataset.

The IMU data simulation follows the PNP data generation method, by simulating the raw signals, including the accelerations, angular velocities, and magnetic field measurements, and employing the IMU fusion algorithm utilizing the error-state Kalman Filter.

Shape-aware Physics Model

We use the dynamics module to explicitly apply the physical constraints following PIP, in order to obtain the motion, internal joint torques, and ground reaction forces that align with the reference but also satisfy physical constraints. Input of the physics module is local pose, foot-ground contact predicted by motion estimator and target-shaped joint velocity output from R_v. However, the dynamic model used in previous works are based on a single human model proposed in Physcap. This directly leads to the fact that the greater the difference between the target body shape and the model, the more inconsistent the calculated translation will be. To address this, we propose a shape-aware dynamic model and kinematic model initialized by $\mathbf{Bone}$.

Our joint rotation controller computes the desired joint angular acceleration $_{des}$ from the estimated reference joint rotations using:

$$\theta_{des} = k_p\,(E_\theta - \varphi) - k_d\,\dot{\theta} \tag{5}$$

Where and are the current joint angles and angular velocities; $\mathbf{E}(\cdot)$ transforms the reference pose to local Euler angles; k_p = 2400 and k_d = 60 are the gain parameters.

Joint position controller utilizes the proposed shape-aware kinematic model and follows (Yi et al. 2022) in dynamic state updating.

Experiments

In this section, we present the implementation details. We then compare our method with state-of-the-art in motion capture from sparse IMUs and evaluate the key contributions of our method.

Implementation Details

Networks Our method incorporates 8 neural networks, including 2 recurrent networks for kinematic signal retargeting, 5 recurrent networks for local pose and global motion estimation following (Yi et al. 2022) and 1 fully connected network for the fictitious force estimator following (Yi, Zhou, and Xu 2024) The kinematic signal retargeters each contains 4 layers with a hidden width of 512, activated by ReLU, and optimized by the Adam optimizer. We train our retargeting networks on a windows PC with NVIDIA RTX 4090D graphics card.

Hardware and Performances For real world data collecting, we use noiton PN studio system with 17 IMUs. Our framework is implemented in Pytorch and the physics-based optimization is implemented using Rigid Body Dynamic Library (RBDL). Our live demo also uses noitom Perception Neuron series IMUs.

Comparisions

We compare our method with the state-of-the-art works in motion capture from sparse IMUs including PIP(Yi et al. 2022) and PNP(Yi, Zhou, and Xu 2024). When evaluating local pose accuracy, we align the root joint position and orientation with the ground truth, and use the same metrics as in (Yi, Zhou, and Xu 2021; Yi et al. 2022; Yi, Zhou, and Xu 2024), including:

- **SIP Error (°)**: the global rotation error of hips and shoulders.
- **Angular Error (°)**: the global rotation error of all joints.
- **Positional Error (cm)**: the position error of all joints.
- **Mesh Error (cm)**: the vertex error of the posed SMPL meshes.
- **Jitter ($10^3 m/s^3$)**: the average jerk of all joints *w.r.t.* the world.

The pose comparison results are presented in 1. Our method consistently outperforms previous works in pose accuracy on both synthesized dataset DanceDB and the real world data of our own collection. Note that we also report results of (Yi et al. 2022) on the unscaled adult DanceDB dataset, which demonstrate that state-of-the-art baseline methods are able to perform motion reconstruction as long as subjects are SMPL-shaped adults. This also shows the effectiveness of our motion data retargeting networks. We show qualitative comparison results in Fig.2, selected poses are all from non-SMPL subjects. Our method is visually the most accurate over all the methods.

Table 1: Comparison with state-of-the-art on the synthesized DanceDB dataset (DanceDB*) with non-SMPL shaped subjects and a real dataset of our own collection (PSP dataset). We also report the performance of PIP on the original DanceDB dataset.

Method	SIP Error (deg)	Angle Error (deg)	Joint Error (cm)	Vertex Error (cm)	Jitter Error (km/s^3)
PSP Dataset (height 153cm)					
PIP	13.84	7.48	4.54	5.29	0.11
PNP	13.51	9.64	5.17	6.09	**0.10**
PSP (ours)	**12.76**	**7.44**	**4.43**	**4.78**	0.12
DanceDB* (height 100cm-120cm)					
PIP	15.82	10.93	7.03	8.23	0.47
PNP	12.85	9.39	5.64	6.64	0.58
PSP (ours)	**12.32**	**8.67**	**5.24**	**6.09**	**0.47**
DanceDB (adults over 165cm)					
PIP	11.79	8.27	4.98	5.82	0.43

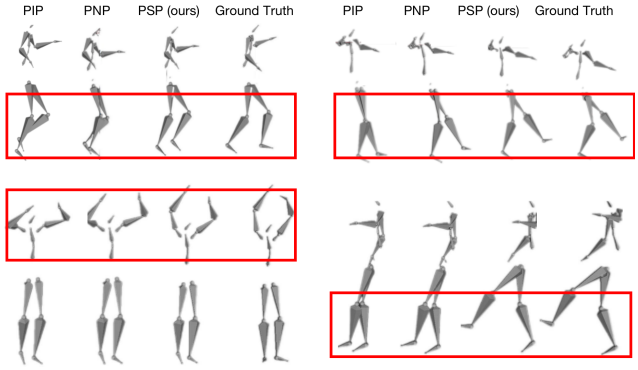

Figure 2: Qualitative comparisons with prior works. The examples are picked from DanceDB* dataset.

Table 2: Ablation study on retargeting networks (R) and shape-aware physics model (P) on PSP dataset.

Method	**Angle Err**	**Joint Err**	**Vertex Err**	**Jitter**
PNP(baseline)	9.64	5.17	6.09	**0.10**
PNP w/ R	9.36	4.53	5.04	0.11
PNP w/ R&P	**7.44**	**4.43**	**4.78**	0.12

Evaluations

Shape-aware Motion Signal Retargeting In Fig.3 we evaluate our shape-aware motion signal retargeting by visualizing accelration data input and output of R_a. (a) demonstrates the difference in IMU input of an adult and a child when performing the same pose and, thus, necessary to utilize our network before using baseline methods. The latter three figures show the output of R_a compared to the SMPL-shaped accelerations. In figure (b), the output a_s is obviously close to the ground truth accelrations. We also show a more clear comparison by separate the two signals in (c) (ground truth) and (d) (a_s).

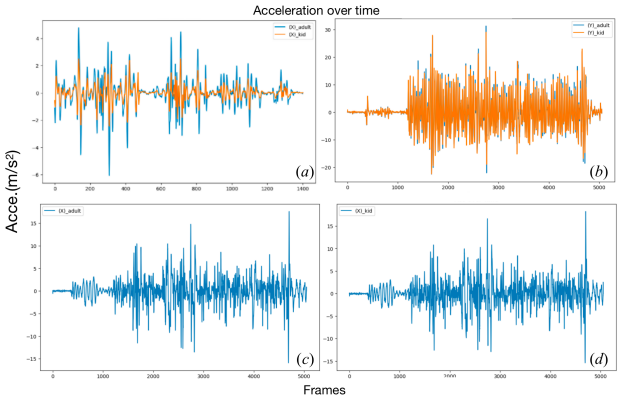

Figure 3: We demonstrate 4 figures of acceleration over time. (a): Root joint IMU acceleration data of adult(DanceDB, blue) and child(DanceDB*, orange); (b): Our acceleration retargeting results(orange) and original SMPL-body accelerations(blue) of root IMU Y-axis; (c) and (d): X axis of acceleration in (b). The examples are picked from DanceDB*.

Shape-aware Physics Model In Tab.2 , we show ablation study results on the real-world PSP dataset. We use PNP without its physics model as our motion estimator. Note that although our retargeting networks decrease joint error greatly, our shape-aware physics model has a significant effect on angular errors. We attribute this to our shape-aware dynamic model.

Conclusion We address the issue of non-SMPL shaped human motion tracking with sparse inertial sensors. We propose to use two separate learning-based motion signal retargeting networks before and after performing motion estimation. Additionally, we designed a shape-aware motion optimization technique to replace the physics model used in prior works. Our proposed PSP method outperforms the state-of-the-art in pose accuracy on both synthesized dataset and a real-world dataset of our own collection.

References

2021. EasyMoCap - Make human motion capture easier. Github.

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; Hilliges, O.; and Pons-Moll, G. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6): 1–15.

Jiang, J.; Streli, P.; Qiu, H.; Fender, A. R.; Laich, L.; Snape, P.; and Holz, C. 2022a. AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing. In *European Conference on Computer Vision*.

Jiang, Y.; Ye, Y.; Gopinath, D.; Won, J.; Winkler, A. W.; and Liu, C. K. 2022b. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Kehl, W.; Milletari, F.; Tombari, F.; Ilic, S.; and Navab, N. 2016. Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 205–220. Springer.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 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.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11127–11137.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Nagymáté, G.; and Kiss, R. M. 2018. Application of OptiTrack motion capture systems in human movement analysis: A systematic literature review. *Recent Innovations in Mechatronics*, 5(1): 1–9.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985.

Point, N. 2011. Optitrack. Natural Point, Inc. *Natural Point Inc*.

Ponton, J. L.; Yun, H.; Aristidou, A.; Andujar, C.; and Pelechano, N. 2023. SparsePoser: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1): 1–14.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, 349–360. Wiley Online Library.

Wang, D.; and Zhang, S. 2023. 3D Human Mesh Recovery with Sequentially Global Rotation Estimation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14907–14916.

Wu, S.; Jin, S.; Liu, W.; Bai, L.; Qian, C.; Liu, D.; and Ouyang, W. 2021. Graph-based 3d multi-person pose estimation using multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11148–11157.

Xu, X.; Chen, H.; Moreno-Noguer, F.; Jeni, L. A.; and la Torre, F. D. 2020. 3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning. *ArXiv*, abs/2007.13666.

Ye, H.; Zhu, W.; Wang, C.; Wu, R.; and Wang, Y. 2022. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision*, 142–159. Springer.

Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 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.

Yi, X.; Zhou, Y.; and Xu, F. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Yi, X.; Zhou, Y.; and Xu, F. 2024. Physical Non-inertial Poser (PNP): Modeling Non-inertial Effects in Sparse-inertial Human Motion Capture. In *SIGGRAPH 2024 Conference Papers*.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5746–5756.

Zhang, H.; Meng, Y.; Zhao, Y.; Qian, X.; Qiao, Y.; Yang, X.; and Zheng, Y. 2023. 3D Human Pose and Shape Reconstruction From Videos via Confidence-Aware Temporal Feature Aggregation. *IEEE Transactions on Multimedia*, 25: 3868–3880.

Zhang, J.; Cai, Y.; Yan, S.; Feng, J.; et al. 2021. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34: 13153–13164.

Zhao, Q.; Zheng, C.; Liu, M.; and Chen, C. 2024. A single 2d pose with context is worth hundreds for 3d human pose estimation. *Advances in Neural Information Processing Systems*, 36.

Zuo, C.; Wang, Y.; Zhan, L.; Guo, S.; Yi, X.; Xu, F.; and Qin, Y. 2024. Loose Inertial Poser: Motion Capture with IMU-attached Loose-Wear Jacket. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2209–2219.