EgoPoser: Robust Real-Time Ego-Body Pose Estimation in Large Scenes

Jiaxi Jiang Paul Streli Manuel Meier Christian Holz
Department of Computer Science, ETH Zürich, Switzerland
{firstname.lastname}@inf.ethz.ch

Abstract

Full-body ego-pose estimation from head and hand poses alone has become an active area of research to power articulate avatar representation on headset-based platforms. However, existing methods over-rely on the confines of the motion-capture spaces in which datasets were recorded, while simultaneously assuming continuous capture of joint motions and uniform body dimensions. In this paper, we propose EgoPoser, which overcomes these limitations by 1) rethinking the input representation for headset-based ego-pose estimation and introducing a novel motion decomposition method that predicts full-body pose independent of global positions, 2) robustly modeling body pose from intermittent hand position and orientation tracking only when inside a headset's field of view, and 3) generalizing across various body sizes for different users. Our experiments show that EgoPoser outperforms state-of-the-art methods both qualitatively and quantitatively, while maintaining a high inference speed of over 600 fps. EgoPoser establishes a robust baseline for future work, where full-body pose estimation needs no longer rely on outside-in capture and can scale to large-scene environments.

1. Introduction

Current Mixed Reality (MR) systems derive tracking cues and user input mainly from cameras embedded inside the headset, which observe the environment as well as the user's hand motions when inside the field of view [15, 16]. This enables them to track their own position inside the world and at the same time derive input commands from the user's actions. Due to the constrained nature of the input signals, which primarily rely on data from the user's head and hands, contemporary Mixed Reality (MR) systems are limited in their ability to generate comprehensive virtual representations, confining the avatar to encompass only the upper body. Consequently, this restriction undermines the sense of immersion and results in a reduction of the overall experiential fidelity.

Because holistically embodying users as 3D avatars in Mixed Reality is desirable for presence [17], immersion, and thus user experience [31] in AR and VR alike, several recent methods have attempted to estimate full-body poses from the sparse tracking cues current systems provide [7, 9, 19, 35]. These efforts have all relied on large motion-capture datasets to estimate realistic body poses and animations, leveraging the robust, continuous, and high-fidelity recordings across a large variety of environments. However, as we demonstrate in this paper, previous methods over-rely on the nature of their training data, specifically the continuous motions recorded in outside-in tracking setups. Their conditions are not representative of the intended use: Not only are the user's hand poses continuously available in motion-capture datasets and, thus, during testing and evaluation, the stationary setup of motion capture leads previous methods to overfit to global coordinates.

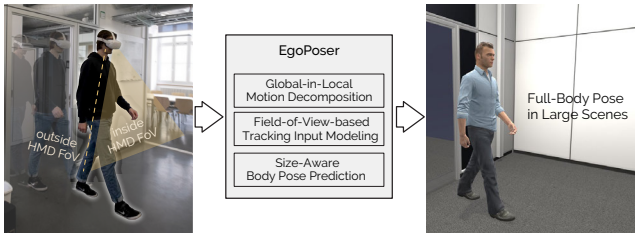


Figure 1. Our proposed EgoPoser overcomes the reliance of previous pose-estimation methods on global position, outside-in tracking, and uninterrupted availability of tracking information, while simultaneously adapting motions to the user's body dimensions,

Specifically, existing methods exhibit several limitations in considering real-world applications. (1) Prior approaches directly employ the global pose in world space as the network input, causing the trained model to overfit to training data that is typically concentrated near the origin. Our paper reveals that using a global input representation results in significantly worse predictions, even for slight meter-scale offsets. (2) Current methods assume that hands will always remain within the field of view. However, for more portable inside-out tracking systems, hands might occasionally move out of the field of view, resulting in intermittent input signals. Although recent work proposed a random

masking strategy [7], it fails to accurately model the temporal and spatial characteristics of the real scene. (3) Existing methods only account for a mean body shape, disregarding the natural variations in body size among different subjects. This limitation prevents the model from adapting to real-world inputs and accurately representing the body. Consequently, motion artifacts such as floating and ground penetration may arise.

To solve these problems, we propose EgoPoser, an exclusively headset-centered estimation method for full-body poses that robustly performs on the sparse and intermittent tracking cues provided by today's inside-out tracking systems. As shown in Figure 1, EgoPoser comprises several components that jointly enable its robust performance on real-world data and live motions outside motion-capture datasets: Our novel Global-in-Local motion decomposition retains the critical relative global information in a local representation, making it robust to position changes by encoding motion priors from sparse inputs. We sample the original signals at different rates to capture longer motion time series as the input to the Transformer encoder, while improving prediction accuracy without increasing the computational burden. EgoPoser's realistic field-of-view modeling captures both spatial and temporal information to smoothly estimate accurate full-body poses even when the user's hands leave the camera's view frustum. To support personalized use, we predict the body size to accurately anchor each user's representation within the world.

To summarize, the main contributions of this paper are:

(1) We introduce EgoPoser, a novel systematic approach for full-body pose estimation using sparse motion sensors. To the best of our knowledge, We are the first to study the HMD-based ego-body pose estimation in large scenes. Our method remains robust even when hands are out of the field of view, and generalizes well to various body sizes.

(2) We have identified a notable issue with existing methods, wherein they tend to overfit to the training data due to the global input representation of the neural network. To address this concern, we emphasize the significance of position-invariant prediction and present an effective Global-in-Local motion decomposition strategy.

(3) Unlike existing methods that assume a mean body shape, EgoPoser stands out by effectively accommodating different body sizes, showcasing remarkable input adaptability, and delivering accurate output avatar representation. Moreover, our proposed strategies significantly reduce motion artifacts such as floating and ground penetration.

(4) We showcase superior numerical and visual performance compared to state-of-the-art methods on public datasets AMASS and HPS. Our demo also shows that EgoPoser seamlessly integrates into real-world MR systems, affirming its practical viability and effectiveness.

2. Related Work

Pose Estimation from Sparse Sensors. Previous research on full-body pose estimation from sparse inputs has utilized up to six body-worn inertial sensors [18, 30, 36, 37]. However, these sensors are distributed across the body, making motion capture inflexible and unwieldy. As egocentric vision [13, 39] has recently attracted more attention, an increasing amount of research is focused on full-body pose estimation using head-mounted devices. However, early methods [7, 9, 35] assume the pose of the root is available, which requires an additional tracker attached to the pelvis in real usage. For the practical 3-Point tracking problem, AvatarPoser [19] was the first method to train a single model for various motion types. It combines a Transformer-based method with IK optimization to make the prediction realistic and match the observation. QuestSim [33] combined a reinforcement learning-based method with a physical simulator to make the prediction physically plausible. Recently, diffusion model-based methods AGRoL [10] and EgoEgo [21] were proposed, which synthesized smooth predictions. However, they both relied on future input signals to make current predictions, and diffusion models inherently have a slow sampling speed by design. These two factors pose significant challenges for real-time applications.

Pose Estimation under Field-of-View Constraints. Estimating human pose when parts of the body are outside the cameras' field of view is a challenging problem [3, 8, 26, 34]. To alleviate the visibility problem, many designs were tried in terms of hardware such as mounting a camera *et al*. [22] or IMUs [28] to the wrist, adding cameras to controllers [5], mounting downward facing fisheye cameras to a specially designed hat [6, 27, 32] or glasses [40]. But these designs often come at an additional cost and are often not portable or aesthetically pleasing. In terms of the algorithm solution, FLAG [7] retained the original constraints of headset-only capture and augmented the training data by randomly masking the hands with a certain probability instead. However, this strategy does not consider the actual spatial relative pose between the hand and headset—a hand can be masked out even if it is actually inside the FOV. In this paper, we model the FoV realistically by considering the spatial relative pose and the temporal continuity.

Data Normalization in Deep Learning. Normalization is a technique utilized in deep learning to adjust input data such that all features have similar scales. This is particularly beneficial when features in a dataset have varying scales. Normalization can enhance the model's generalization ability by making it more robust to variations in the input data.

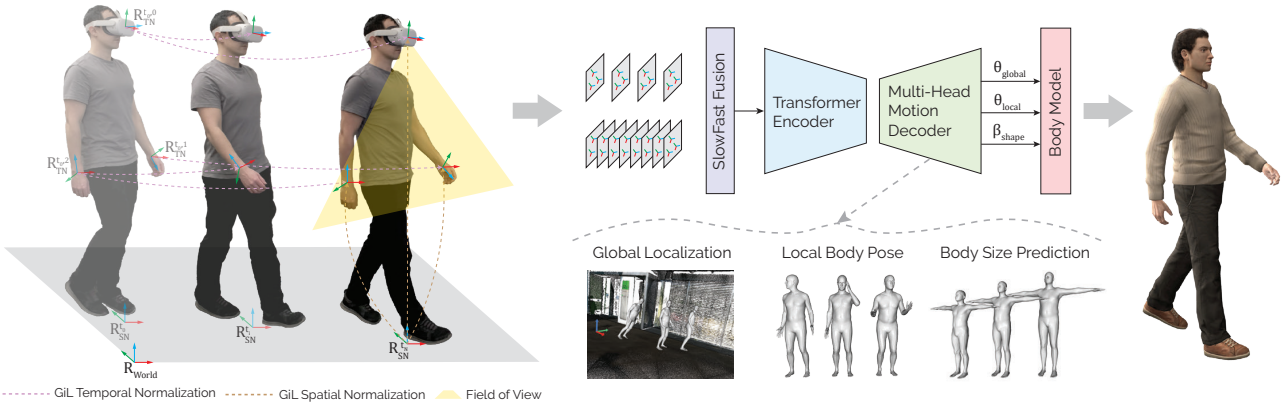


Figure 2. **The architecture of our proposed EgoPoser for full-body pose estimation from a single HMD device.** Our proposed Global-In-Local (GiL) strategy enables us to decompose global motion from input tracking signals, making the model robust to different user positions. We sample these signals at different rates, capturing both dense nearest information and sparse but longer information. The resulting preprocessed signals are then fused by SlowFast Fusion module and fed into a Transformer Encoder. The Multi-Head Motion Decoder outputs parameters for global localization, local body pose, and body size prediction. Given N=80 frames as input, we generate the last frame to function as the full-body representation for each timestamp, facilitating real-time applications.

However, it can also cause a loss of information in the original data if the original scale of the input features is important. In our work on HMD-based human pose estimation, we have found that prior approaches [10, 19] which use the global pose as input do not generalize well to different positions. Although common normalization techniques, such as subtracting the head pose from the global pose [4, 7, 9, 35], are translation-invariant, they result in information loss and reduced performance.

3. Method

In this section, we describe our method EgoPoser for the real-time estimation of the global full-body pose based on the tracked hand and head poses of an HMD.

3.1. Overview

While various MR systems may diverge in the tracking technology they depend upon, The global positions p and orientations Θ of the headset, along with those of the two hands, are typically accessible. Given this 3-point pose information as input, the goal of full-body pose estimation is to estimate a mapping from the input and the position of J full-body joints:

$$\{p_t^j\}^{j=1:J} = f(\{p_t^j, \Theta_t^j\}^{j=1:3}) \tag{1}$$

This is a challenging under-determined problem because the same input may correspond to multiple possible outputs. To make the predicted results conform to the human skeleton, most existing work used first 22 joints defined in the kinematic tree of the SMPL-H [23, 24] skeleton models for the output full-body representation, ignoring the pose of fingers. Besides, SMPL-H models provide 16 shape parameters β to control the overall shape of a 3D human mesh.

Existing methods suffer from limitations in problem formulation, impacting their performance in practical scenarios. First, these methods feed global poses directly into the network, causing overfitting to specific data recording environments. Second, they assume hands are consistently visible in the headset's field of view, relying on uninterrupted signals. Third, they assume a standard body shape and disregarding natural body size variations. Addressing these limitations is vital for enhancing the applicability and accuracy of proposed approaches.

We show an overview of our approach in Figure 2. The core components of our approach include our proposed Global-in-Local (GiL) representation, our temporal SlowFast feature fusion module, a Transformer Encoder as well as a human motion decoder that forwards its output consisting of the root orientation, the local joint rotations and the shape parameters β together with the decomposed global head position to the differentiable SMPL body model for the estimation of the global joint positions.

3.2. Global-in-Local Motion Decomposition

MR tracking systems usually directly provide the global pose of the headset and hands/controllers. Based on that, AvatarPoser [19] and AGRoL [10] directly take the global poses in world space as input to its network. However, most existing datasets like AMASS are recorded in a limited physical space near the origin, so it remains unclear whether it can generalize well to a different location. Data augmentation could be a remedy, yet it makes the training process less efficient, and it is impossible to traverse the whole infinity 3D space. In addition, the global position could leak training data-specific scene information into the learning process. For example, a sit-down motion might be more likely within locations where a chair was placed dur-

ing the data capture, and the action of throwing a basketball is likely to happen close to the basketball hoop affixed in the capture studio.

Another common strategy is to decompose the global motion into a rigid body motion in global world space and the local motion relative to a root capturing the current body pose. This is also the default representation of many body motion capture representations like SMPL [24] and has been used in previous works on self-body pose estimation [4, 7, 35]. This is an intuitive design choice when root information, which provides a rich global pose, is available. Besides, since models trained with local representations do not utilize global position, it can generalize well to different locations. However, in 3-point tracking problem, simply converting the reference frame to head is easy to make the prediction sensitive to the head rotation. Besides, removing the global information can lead to information loss, making the ill-posed problem even more challenging.

To combine the advantages of both representations, we introduce a motion Global-in-Local motion decomposition strategy (GiL). GiL is designed to be position-invariant, thereby enhancing its resilience in pose estimation across expansive scenes. GiL also extracts crucial relative global information and leverages it to encode motion data effectively. The GiL approach encompasses two key operations: spatial and temporal normalization, detailed as follows.

(1) Spatial Normalization (SN): Instead of subtracting the head's 3D translation information from head and hand positions to make the hand position relative to the head, we only make the horizontal translation relative to the head. We retain the global vertical translation as an important feature to encode motion priors.

(2) Temporal Normalization (TN): we do temporal normalization by subtracting the horizontal translation of all the input joints at the first frame from the horizontal translation of the corresponding joints within a temporal window so that we capture the relative global trajectory of each joint within a temporal window.

The left of Figure 2 illustrates how GiL works. The Spatial and temporal normalization of GiL convert the positions of headset and hands from the world frame R_{World} into a new frame R^t_{SN} on the ground and R^t_{TN}, j on the i-th joint, which are the input to our pose estimation network. The positions given by SN and TN at time t_i are written as:

$$\begin{aligned} p^{t_i,hand}_{SN,h} &= p^{t_i,hand}_{W,h} - p^{t_i,head}_{W,h} \\ p^{t_i,j}_{TN} &= p^{t_i,j}_{W} - p^{t_0,j}_{W} \end{aligned} \tag{2}$$

In addition to the orientation and decomposed position information, we calculate their corresponding linear and angular velocity to enrich the input data. 6D representation [41] is adopted for rotation due to its continuity. Finally, a total of 59 input features are provided to the network.

3.3. Realistic Field of View Modelling

We find that previous work does not adequately address the inherent limitations of the inside-out hand tracking on today's state-of-the-art headsets like *Meta Quest 2* and *HoloLens 2*. They usually model tracking failures through random frame drops uniformly sampled over the complete tracked motion but fail to take into account that whole regions outside the headset's cameras field of view exist where the hand tracking fails.

Based on the head pose, which determines the viewing angle of the cameras mounted to the headset, and the relative position of the hands, we simulate the tracking of headsets with varying field of views (FoVs).

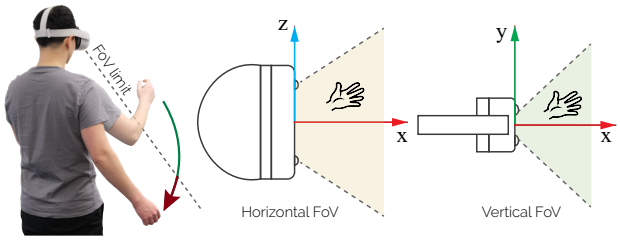


Figure 3. **An illustration of HMD's field-of-view.** Left: A user is moving the hand from inside FoV to outside FoV. Middle: A hand is in the horizontal FoV. Right: A hand is in the vertical FoV.

For an HMD with a horizontal FOV of α_h and a vertical FOV of α_v, a hand is visible if the following conditions are satisified,

$$\begin{aligned} \tan(\frac{\alpha_h}{2}) &\geq \left|\frac{z_{hand}}{x_{hand}}\right|, \\ \tan(\frac{\alpha_v}{2}) &\geq \left|\frac{y_{hand}}{x_{hand}}\right|, \\ x_{hand} &> 0. \end{aligned} \tag{3}$$

Here, x_{hand}, y_{hand} and z_{hand} are x, y and z coordinate of the hand position in a head-centered coordinate system with the x-axis pointing through the eyes (see Figure 3).

We train our method to robustly handle inputs with continuous tracking gaps by setting the input features of joints outside the field of view to 0. Since the hand pose serves as a reference frame in the temporal normalization process of GiL, we employed the predicted hand pose as the reference frame when the real hands are positioned outside the FoV.

3.4. SlowFast Feature Fusion

Based on the information from a single frame, a multitude of plausible body poses exist that would fit a given set of head and hand poses. However, the problem converges towards a more unique solution as we consider the head and hand motions over a longer temporal context. Yet, simply adding more frames to the input would significantly increase computational overhead as for example, the

computational complexity of a Transformer's self-attention module scales quadratically with input sequence length. Thus, inspired by SlowFast networks originally proposed for video recognition [11], we propose a SlowFast feature fusion module that increases the context of considered past tracking frames in a more efficient manner. Given an input window of τ past frames, the SlowFast module concatenates the linear embeddings for the $\frac{\tau}{2}$ last frames (FAST) with $\frac{\tau}{2}$ frames sampled with a stride of 2 over the complete window (SLOW). In this way, we reduce the length of the input sequence by a factor of 2 while keeping the temporal context over the whole window. In addition, we still capture the temporal information contained within the higher temporal resolution of the FAST input frames.

3.5. Considerations of Different Body Sizes

One common limitation of existing methods [9, 10, 19] is that only the mean shape skeleton is used, so that the difference of body sizes are ignored. This assumption can introduce two main problems: First, since all the training and testing data share the same body shape. It is unclear whether a trained model can still work well to input data from users with diverse body sizes in real-world usage. Second, the final animated result can not reflect the real body size. Even for perfect joint angle rotation, wrong body skeleton can lead to ground penetration or floating artifacts.

To address these issues, we introduce an approach that combines data augmentation with T-pose calibration. We augment the training data by incorporating ground truth shape parameters. During the testing phase, we measure the body height and arm length, subsequently computing the ratio between the measured dimensions and the corresponding mean shape values. The average of these ratios is then utilized as the scaling factor for the entire body. This scale factor is subsequently applied to adjust the output representation accordingly.

Nevertheless, this method requires an extra calibration step to accurately measure the body sizes, adding to the overall effort involved. To this end, we introduce a calibration-free method via estimating the size of the user from the tracking input. However, simply predicting the shape parameters is very challenging because different body shapes can map to the same input pose. Besides, estimation of fine-grained body shape such as if the body is fat or thin is not important, and can be easily adjusted by post processing. However, the body size plays an important role because it can determine the scale of the input signals, and also affect the output avatar representation, because wrong body size can lead to ground penetration or foot floating artifacts.

Instead of directly supervising the shape parameters β in the loss function, we implicitly optimize the estimated β through the error in the joint positions that are the output of the shape-aware differentiable SMPL body model that takes β and the estimated joint rotations as input. In addition, we perform L1 regularization on β to encourage sparsity and the estimation of zero-valued mean shape parameters in case they are not needed for the accurate estimation of the joint positions. As our method estimates β for each frame, we can apply the median predicted shape parameters after some sequences to enforce consistency, although we did not observe frequent or sudden deviations in β for a given input sequence.

The loss function for body size estimation is written as:

$$\mathcal{L}_{BS} = \lambda_{pos}\mathcal{L}_{pos} + \lambda_{\beta} \left\|\beta\right\|_1 \tag{4}$$

where the positional loss is calculated through forward kinematics:

$$\mathcal{L}_{pos} = \left\|FK(\theta, \beta) - FK(\theta_{GT}, \beta_{GT})\right\|_1 \tag{5}$$

The final loss function is composed of an L1 local rotational loss, an L1 global orientation loss, an L1 positional loss, and a L1 regularization of β denoted by:

$$\mathcal{L}_{total} = \lambda_{ori}\mathcal{L}_{ori} + \lambda_{rot}\mathcal{L}_{rot} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{\beta} \left\|\beta\right\|_1 \tag{6}$$

We set the weights λ_{ori}, λ_{rot}, λ_{fk}, and λ_{β} to 0.05, 1, 1, and 0.01 respectively.

3.6. Design Choices for Real-Time Application

Although AGRoL [10] can generate smooth motion and be accelerated via sampling only 5 times, we find its benefit can only be utilized in an offline fashion but still challenging for real-time application because (1) the sampling speed of diffusion models is very slow by design (2) future inputs signals are used to generate the current pose.

The total delay t_d is calculated as:

$$\begin{aligned} t_d &= t_n + t_q \\ &= t_n + (s-1) \cdot v_{play} \end{aligned} \tag{7}$$

where t_n is the inference time of the network for each pass, t_q is the queue time for the output frames to be played. The queue time t_q depends on the number of output frame s and the play speed of the MR system.

In AGRoL, given the inference time of each pass as 35ms, if all the 196 output frames are used, the total delay played at a 30Hz MR system will be $t_d = 35 + 1000 \times (196-1)/30 = 6535$ ms, which is far from real-time application. Therefore, for real-time application, only one frame can be used. We follow the design choice of AvatarPoser [19] to use a lightweight Transformer backbone (t_{net} = 1.6ms) and only generate one frame as output.

4. Experiments

4.1. Datasets and Training Details

Following prior work [19], we mainly utilized three subsets of the AMASS [24] dataset, namely CMU [1], BML-

Table 1. **Evaluation of different methods on AMASS dataset.** A study to evaluate the robustness of various methods to offset from the origin. The evaluation metrics are position error MPJPE [cm] and velocity error MPJVE [cm/s]. *: AGRoL was tested in an offline fashion in its original paper because the future frames are used for current prediction. To ensure a fair comparison, we run the its public code and set the step of sliding window as 1. Observations reveal that the AvatarPoser [19] and AGRoL [10], which use global pose as input, experiences a substantial decrease in performance with increasing offset from the origin.

Methods	Offset = 0 m		Offset = 1 m		Offset = 2 m		Offset = 5 m		Offset = 10 m		Offset = 50 m	
	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE
Final IK	18.09	59.24	18.09	59.24	18.09	59.24	18.09	59.24	18.09	59.24	18.09	59.24
CoolMoves	7.83	100.54	7.83	100.54	7.83	100.54	7.83	100.54	7.83	100.54	7.83	100.54
AvatarPoser	4.18	29.40	4.78	29.77	5.49	30.81	11.39	41.88	18.40	66.19	31.67	68.79
AGRoL	3.86	50.94	4.89	61.83	6.73	85.02	11.90	161.08	20.74	252.54	53.45	1795.78
AGRoL-Offline*	**3.71**	**18.59**	4.76	**21.86**	6.58	26.33	11.77	45.50	20.94	138.00	63.64	2087.78
EgoPoser (Ours)	4.14	25.95	**4.14**	25.95	**4.14**	**25.95**	**4.14**	**25.95**	**4.14**	**25.95**	**4.14**	**25.95**

Table 2. **Comparisons to state-of-the-art methods on HPS.** HPS is a large-scene dataset captured in large scenes. The evaluation metrics are position error MPJPE [cm] and velocity error MPJVE [cm/s].

Methods	BIB_EG_Tour		MPI_EG		Working_Standing		UG_Computers		Go_Around		UG_Long	
	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE
AvatarPoser	22.53	60.25	16.54	36.39	19.08	52.95	23.24	40.65	19.50	59.54	16.65	43.59
AGRoL	28.95	166.34	19.41	55.52	17.67	53.97	20.90	109.12	14.16	98.34	12.81	74.13
EgoPoser (Ours)	**9.55**	**49.39**	**11.05**	**35.60**	**8.70**	**46.49**	**10.34**	**37.63**	**6.90**	**45.10**	**8.95**	**38.30**

rub [29], and HDM05 [25], for both training and testing. We adopted the same data splitting which was generated by randomly allocated 90% of the data to the training set and 10% to the test set. To evaluate the performance in large-scale scenarios, we also used HPS [14] for testing, using the pre-trained model trained on the three subsets in AMASS. HPS is a large-scene dataset captured in large scene. We utilize the high-quality results from the joint optimization described in [14], which integrates camera localization, IMU pose estimates, and scene constraints, as our ground truth data.

To optimize the parameters of EgoPoser, we adopted the Adam solver [20] with a batch size of 256. We consider the previous 80 frames as input ($\tau = 80$), resulting in an input window with 40 frames after SlowFast fusion. The learning rate starts from 1×10^{-4} and decays by a factor of 0.5 every 2×10^4 iterations. We trained our model using PyTorch on a single NVIDIA GeForce GTX 3090 GPU.

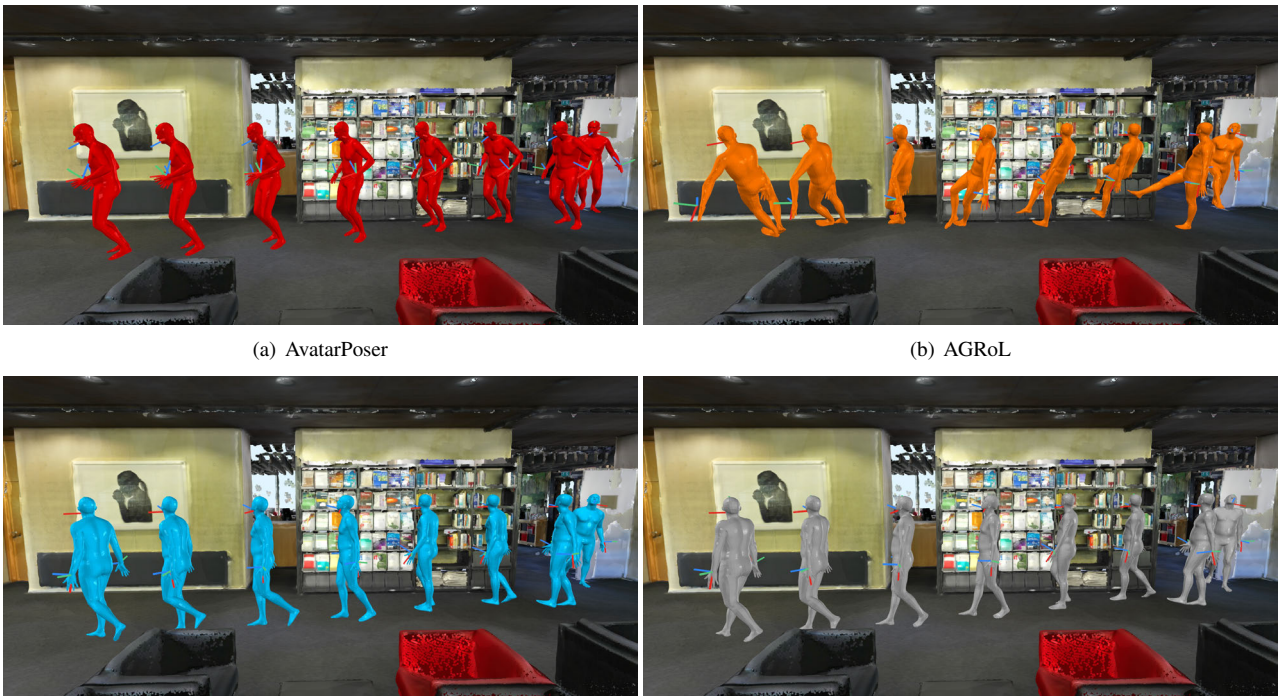
4.2. Evaluation Metrics

We use Mean Per Joint Position Error (MPJPE) and Mean Per Joint Velocity Error (MPJVE) as our main evaluation metrics to measure the estimation accuracy and smoothness. Due to page limitations, we excluded the Mean Per Joint Rotation Error (MPJRE), as the MPJPE stands as a more representative metric for assessing pose accuracy, and the trends in MPJRE and MPJPE across various methods are fundamentally aligned. When evaluating the size-aware pose estimation, we adopt the Mean Per Joint Position Error (MPJPE), Mean Vertex Error (MVE), mean errors of predicted heights and bone lengths. Besides, we compute the average distance to the ground for mesh vertices below the ground to evaluate ground penetration [38]. To evaluate foot floating artifacts, we calculate the mean distance between the ground and the lowest point on the mesh within a given sample when it is above the ground.

To ensure a fair comparison with state-of-the-art methods and to provide a clear demonstration of each proposed component, we assume the full hand visibility and use the mean body shape following prior work when directly comparing the results with them (Table 1 and 2). We evaluate the hand partial visibility problems and size-aware pose estimation independently in Table 3 and 4, respectively.

4.3. Evaluation Results

Pose Estimation in Large Scenes. To test the robustness to different position in large scenes, we add an offset to the position and use it as the input. We synthesize the offset by setting it as constants in different scales, ranging from 0 to 50. To test the robustness of our approach to different positions in real-world scenarios, we introduce an offset to the position and use it as input. We synthesize the offset by setting it as constants in various scales, ranging from 0 to 50. We compare our method with AvatarPoser [19] and AGRoL [10], which use the global input representation, the classical KNN-based method CoolMoves [4], and the traditional optimization-based FinalIK [2]. It's worth noting that the latter two methods use a local input representation. Besides, AGRoL was tested in an offline fashion in its original paper because the future frames are used for current prediction, but this setting is not practical in real-world MR scenarios, as we have discussed in Section 3.6. To ensure a



(a) AvatarPoser (b) AGRoL

(c) EgoPoser (Ours) (d) Ground Truth

Figure 4. **Visual results on HPS dataset.** The provided example depicts a user walking from an outdoor setting into a library and subsequently walking within the library. As illustrated in the figure, the outcomes from AGRoL are notably incorrect, and AvatarPoser maintains a static local pose. In contrast, our proposed EgoPoser generates natural and visually pleasing motions that closely resemble the ground truth data. Despite being trained only on the AMASS dataset, EgoPoser demonstrates remarkable performance and robustness, surpassing existing state-of-the-art methods.

Table 3. **Evaluation of different strategies for outside the FoV pose estimation on AMASS.**

Strategies	FoV = 180°		FoV = 120°		FoV = 90°	
	MPJPE	MPJVE	MPJPE	MPJVE	MPJPE	MPJVE
Full Visibility [19]	24.75	183.84	38.99	144.42	41.24	95.66
Random Masking [7, 10]	7.09	49.91	13.29	64.09	14.84	58.33
Ours	**5.31**	**39.69**	**6.07**	**46.01**	**6.60**	**48.25**

fair comparison, we run the its public code and set the step of sliding window as 1. We report both its online and offline results in Table 1 and a dropped performance of AGRoL can be observed especially on MPJVE in online testing.

We show the position error and velocity error against different scales of offsets in Table 1. Our observations indicate that AvatarPoser and AGRoL, which use global pose as input, experiences a significant decrease in performance as the offset from the origin increases, although AGRoL achieved the best performance when the offset is zero. On the other hand, methods with local representation remain stable in performance.

We further compare our method with state-of-the-arts in a real-world large-scene MoCap dataset HPS [14]. The numerical and visual results can be found in Table 2 and Figure 4.

Figure 5. **Visual comparisons of different strategies for the scenario where the hands can go out of FoV.** The horizontal and vertical FoVs are set as 120°. Hands outside the FoV are rendered in ●. Default: model trained with full hand visibility: RM: Random Masking strategy. Ours: Realistic FoV Modelling. GT: Ground Truth. Our method achieves the best performance.

Outside-the-FoV Pose Estimation. We have evaluated various strategies for scenarios where hands are tracked by a headset and may go out of field of view (FoV). To simulate real-world scenarios, we have considered different angles of FoV, including 180°(using fisheye cameras), 120°(as in

Table 4. **Evaluation on size-aware pose estimation.** Best results are highlighted in **bold** for each metric. Ground truth shape is used when calculating the evaluation metrics. GP is for ground penetration error and FF is for foot floating error. All metrics are measured in centimeters [cm].

Strategies	MPJPE	Vertex	Height	Arm	GP	FF
Mean Shape [7, 9, 10, 19]	6.36	6.74	7.67	7.42	3.87	5.38
Ours 1 - DA + Calibration	5.26	4.69	**1.36**	**1.24**	**2.06**	1.67
Ours 2 - Size Prediction	**4.79**	**4.08**	1.78	1.66	2.31	**1.64**

Table 5. Ablation studies on global motion decomposition. Best results are highlighted in **bold** for each metric.

Configurations	MPJPE	MPJVE
Mean Normalization (All Features)	6.25	42.69
Mean Normalization (Horizontal + Vertical Positions)	6.24	42.75
Mean Normalization (Horizontal Positions)	6.25	42.87
Spatial Normalization (Horizontal + Vertical Positions)	4.96	29.59
Spatial Normalization (Horizontal Positions)	4.45	27.56
Temporal Normalization	4.58	28.01
Ours - GiL (Temporal + Spatial Horizontal Norm.)	**4.14**	**25.95**

Table 6. Ablation studies on SlowFast design. Best results are highlighted in **bold** for each metric.

Configurations	MPJPE	MPJVE	MACs	#Parameters
length 40	4.36	28.12	160.66M	4.12M
length 80	**4.11**	29.27	326.14M	4.12M
length 80 , stride 2	4.13	30.02	160.66M	4.12M
Ours	4.14	**25.95**	160.66M	4.12M

Oculus Quest 2), 90°(as in Hololens 2). We have tested the results on a model trained on hands with full visibility, denoted as 'Full visibility,' as well as fine-tuned models with random hand masking using a probability p = 0.2 as proposed in FLAG [7] (denoted as 'Random Masking' or 'RM'), and our realistic FoV modeling (denoted as 'Ours').

Table 3 and Figure 5 present the numerical and visual results of models trained with different strategies. When testing the performance in various FoVs using the default model that assumes hands are always visible during training, we observe two main trends. Firstly, as the FoV becomes smaller, the position error MPJPE increases. This is intuitive since a smaller FoV means there are more chances that hands are outside the FoV, making the problem more challenging. Secondly, with a smaller FoV, the velocity error MPJVE initially increases and then decreases. This trend can be explained by when FoV is 180°or 120°, switching between going out and coming back into the FoV can cause strong discontinuity in predictions. When FoV is even smaller, the hands are always outside the field of view, having a smoother but less accurate predictions.

While the random hand masking strategy can improve results, our realistic FoV modelling strategy sets the visibility status based on the actual position of the hand relative to the head, and captures the real temporal dependencies of hand visibility. Consequently, it achieves the best performance in terms of both position accuracy and smoothness.

Size-Aware Pose Estimation. We evaluate the performance of size-aware pose estimation on the same test data as Table 1 from AMASS dataset but with the true shape parameters β. This test set includes over 175 subjects with heights ranging from 145 to 207 cm. As indicated in Table 4, the model trained using the mean body shape achieved a mere 6.36cm in MPJPE. Moreover, the average error pertaining to body dimensions, such as height and arm length, exceeds 7cm. These discrepancies arise from an inaccurate shape representation, consequently giving rise to issues like ground penetration (GP) and foot floating (FF).

Conversely, as our first solution, data augmentation (DA) with ground truth body shapes and subsequently re-scaling the standardized model output by a body size factor obtained via T-pose calibration reduced the MPJPE to 5.26 cm. It also led to considerably enhanced performance across various metrics concerning body sizes and motion artifacts. Furthermore, our calibration-free size prediction approach yielded further improvements in both MPJPE and mean vertex error, while delivering comparable outcomes in metrics related to body sizes and motion artifacts. The size prediction method performs better for MPJPE and vertex error because our model can uncover the latent correlation between human size and shape. Calibration works slightly better for arm length and height as they are directly measured. A visual comparisons to show the importance of size-aware pose estimation can be found in Figure 6.

4.4. Ablation Studies

We make a through ablation studies to show the effectiveness of each proposed component. It should be noted that different strategies for FoV-aware pose estimation and size-ware pose estimation have already been discussed Table 3 and Table 4 in previous section.

Ablation studies for different strategies for global motion decomposition are presented in Table 5. Mean normalization refers to the operation of removing the mean value of each component of all features, or only horizontal features, or horizontal and vertical positions, across the window. Spatial Normalization (Horizontal + Vertical Positions) denotes the method of subtracting the 3D position information of the head from all the inputs, while (Horizontal Positions) denotes that this subtraction is only applied to horizontal translation. The results reveal that retaining vertical information leads to significantly better predictions. We also showed results of applying temporal normalization only. Our GiL motion decomposition method combines temporal normalization and spatial horizontal normalization, resulting in the best performance.

Table 6 shows ablation studies for the SlowFast design,

which demonstrates that our design can improve the results without increasing the model size or computational cost.

4.5. Test on Real HMD devices

To assess the robustness of our method on real-world data, we executed our algorithm on streaming data from *Meta Quest 2*. We use a custom framework based on Velt [12] and Unity to interface with the VR system to handle communication. As shown in Figure 6, EgoPoser have better foot contact than AvatarPoser because of the accurate output size representation. More results can be found in the supplementary videos.

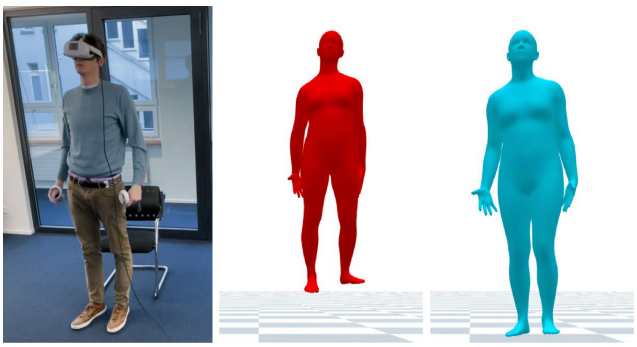


Figure 6. **A visual comparison between AvatarPoser and EgoPoser on a real VR system.** EgoPoser have better foot contact because of the accurate body size representation.

5. Conclusion

We have proposed EgoPoser, a novel systematic approach for 3D full-body pose estimation based alone on the tracking information on contemporary Mixed Reality head-mounted devices. We address the challenges existing efforts face using such platforms, specifically scaling robust estimations to arbitrary real-world settings, handling hands even when they are outside the field of view , and robustness to varying body dimensions. EgoPoser achieves new state-of-the-art performance for accurate motion estimation under these challenging circumstances by combining our novel Global-in-Local motion decomposition method, SlowFast fusion strategy, robust field-of-view modeling, and size-aware pose estimation method. We believe that our proposed strategies can significantly contribute to the advancement of 3D full-body pose estimation and its integration into various VR/AR applications.

Acknowledgments: We sincerely thank Andreas Fender for data recording and manuscript proofreading.

References

[1] CMU MoCap Dataset. http://mocap.cs.cmu.edu/, 2004. 5

[2] RootMotion Final IK. https://assetstore.unity.com/packages/tools/animation/final-ik-14290, 2018. 6

[3] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. Mecap: Whole-body digitization for low-cost vr/ar headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 453–462, New York, NY, USA, 2019. Association for Computing Machinery. 2

[4] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021. 3, 4, 6

[5] Karan Ahuja, Vivian Shen, Cathy Mengying Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. Controllerpose: Inside-out body capture with vr controller cameras. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022. 2

[6] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2

[7] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13253–13262, 2022. 1, 2, 3, 4, 7, 8

[8] Gilles Bailly, Jörg Müller, Michael Rohs, Daniel Wigdor, and Sven Kratz. Shoesense: a new perspective on gestural interaction and wearable applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1239–1248, 2012. 2

[9] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021. 1, 2, 3, 5, 8

[10] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 5, 6, 7, 8

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 5

[12] Andreas Fender and Jörg Müller. Velt: a framework for multi rgb-d camera systems. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, pages 73–83, 2018. 9

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson

Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[14] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 6, 7

[15] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020. 1

[16] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1

[17] Paul Heidicker, Eike Langbehn, and Frank Steinicke. Influence of avatar appearance on presence in social vr. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 233–234, 2017. 1

[18] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, Nov. 2018. 2

[19] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 443–460. Springer, 2022. 1, 2, 3, 5, 6, 7, 8

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[21] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 2

[22] Shuang Li, Jiaxi Jiang, Philipp Ruppel, Hongzhuo Liang, Xiaojian Ma, Norman Hendrich, Fuchun Sun, and Jianwei Zhang. A mobile robot hand-arm teleoperation system by vision and imu. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10900–10906. IEEE, 2020. 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 3, 4, 5

[25] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007. 6

[26] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: Egocentric marker-less motion capture with two fisheye cameras. *ACM Trans. Graph.*, 35(6), nov 2016. 2

[27] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[28] Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. Hoov: Hand out-of-view tracking for proprioceptive interaction using inertial sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023. 2

[29] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2–2, 2002. 6

[30] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, pages 349–360. Wiley Online Library, 2017. 2

[31] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1643–1652, 2018. 1

[32] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11500–11509, 2021. 2

[33] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 2

[34] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M. Kitani. Back-hand-pose: 3d hand pose estimation for a wrist-worn camera via dorsum deformation network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 1147–1160, New York, NY, USA, 2020. Association for Computing Machinery. 2

[35] 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*, volume 40, pages 265–275. Wiley Online Library, 2021. 1, 2, 3, 4

[36] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Phys-

ical 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*, pages 13167–13178, 2022. 2

[37] 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. 2

[38] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7159–7169, 2021. 6

[39] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 180–200. Springer, 2022. 2

[40] Dongxu Zhao, Zhen Wei, Jisan Mahmud, and Jan-Michael Frahm. Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In *2021 International Conference on 3D Vision (3DV)*, pages 32–41. IEEE, 2021. 2

[41] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4