# Self-Avatar's Animation in VR: Extending Sparse Motion Features with Cartesian Coordinates in Transformer-based Model

Antoine Maiorca
antoine.maiorca@umons.ac.be
Thierry Ravet
thierry.ravet@umons.ac.be
Thierry Dutoit
thierry.dutoit@umons.ac.be

ISIA Lab, University of Mons,
31 Boulevard Dolez,
7000 Mons, Belgium

**Abstract**

Animating virtual characters in VR applications is crucial for creating immersive experiences in various use cases and particularly in video games industry. This paper addresses the challenge of accurately animating a virtual avatar's full-body motion using sparse inputs from head-mounted displays (HMDs) and handheld controllers. However, the sparse nature of the input data leads to pose ambiguity issue which can affect the realism of the animation. The paper investigates the impact of extending these sparse motion features with Cartesian coordinates that can be estimated from RGB-D cameras setup, hence without relying on additional tracking devices on the user's body. Experiments are conducted employing *AvatarPoser* [9], a state-of-the-art Transformer-based model in the full-body avatar's animation. The results indicate that augmenting the sparse input, even with a single-view 2D pose, enhances the accuracy of the avatar's full-body motion reconstruction, especially in lower-body tracking. More importantly, our analysis reveals that, when combined with sparse motion signals, 2D Cartesian coordinates from two different perspectives are sufficient to reconstruct the motion at least as accurately as 3D positional data. We believe that this paper impacts positively the development of VR applications since the user experience overall quality is enhanced by the improvement of the self-avatar's motion reconstruction.

## 1 Introduction

Animating virtual characters is a crucial task in a broad range of industrial applications, especially in video games. Various deep learning-based methods have been proposed to synthesize diverse and human-like motion samples [24, 43, 44]. In the Virtual Reality (VR) paradigm, a key component in the design of immersive spaces is the animation of the self-avatar *i.e.*, the virtual representation of the user's body designed to mirror its movements and actions. It helps the users to feel more connected to the virtual space and can enhance the overall sense of immersion and realism in VR experiences. In this context, the subject wears a head-mounted displays (HMD) and handheld controllers whose position and orientation of are computed by integrated tracking sensors. Since a VR commercial system usually provides only these 3 devices, the lack of available information makes the self-avatar's full-body animation challenging. Indeed, it is necessary to estimate lower-body information from the hands and head motion features and this can lead to a pose ambiguity issue since many poses can be estimated from the sparse motion signals.

On the one hand, some works estimate the full-body motion features given this set of sparse inputs [6, 9]. These algorithms animate the lower body without any tracked information from this subset. The main advantage of these methods is that no external tracking device is involved, making it suitable for their implementation in a consumer-grade level.

On the other, to resolve the pose ambiguity, additional information can be used to extend the sparse sensors data, such as external Inertial Measurement Unit (IMU) sensors [8, 14], depth cameras [34] or RGB videos [41]. These methods aim to extract motion features from body limbs that are not tracked by HMDs and controllers. However, such methods come with several drawbacks such as the presence of external devices which can impede the widespread adoption of these technologies. Moreover, equipping the users with such sensor might have an impact on its comfort and affect negatively the overall experience. Then, the process of merging motion capture data from various sources, each with its own unique framerate, can potentially lead to an increase in inference delay. This, in turn, might have consequences for the real-time responsiveness of the animation system, which is an essential feature in this context.

The objective of this research is to analyze the influence of supplementary motion cues on mitigating pose ambiguity challenges in the estimation of complete-body articulated avatar movements. This investigation is conducted employing *AvatarPoser* [9] [1], a state-of-the-art Transformer-based animation model. More precisely, we extend the sparse information from HMD and handheld controllers by Cartesian positions that can be estimated from a setup of RGB(-D) cameras such as 3D [26, 31] or 2D coordinates [4, 5].

# 2 Related Work

## 2.1 Articulated Avatar Animation from Sparse Sensors in VR

The virtual character's full-body pose estimation from sparse sensors is a task widely studied in the literature in various configurations [10, 28, 36, 39, 40]. Typically, the user's body is equipped with several IMU sensors that are used to accurately retrieve its pose. In VR scenarios, the typical approach involves three tracking sensors with 6 degrees of freedom (DoF) to estimate the avatar's pose.

Machine Learning approaches have been proposed to estimate the pose of a virtual avatar using only the sparse sensors signals [1, 25]. These methods employ respectively k-NN and motion matching [3] to fetch the appropriate pose from a motion database that closely aligns with the user's input-defined pose. However, the performance of such algorithms relies importantly on the motion database from which motion samples are selected. This dataset should gather high-quality motion samples that encompass smooth transitions and blending between different motions as well as the wide range of desired actions. Moreover, in VR animation, Ponton *et al.* point out [25] the difficulty of animating upper body gestures with motion matching since the motion of the user's arms are not constrained. This leads to a unmanageable large-scale motion database in order to cover the variety of plausible poses.

[1] https://github.com/eth-siplab/AvatarPoser

Sequential models based on Deep Neural Networks have been utilized to animate the avatar's full-body. *LoBSTr* [37] employs Gated Recurrent Units and an Inverse Kinematics solver to animate respectively the lower- and upper-body pose. Inspired by the popularity of Transformers [32], *AvatarPoser* [9] make use of that family of neural network to predict the local orientations of each joint in the avatar's kinematic tree based on the positions, orientations, linear and angular velocities of the 3 tracking sensors. *DualPoser* [45], built upon *AvatarPoser*, proposes two Transformer-based encoders to handle global and local information and further fuse the processed data. This method improves the accuracy of the pose reconstruction in comparison to *AvatarPoser*.

A last set of solution leverages this problem in kinetic *i.e.*, involving physical simulation of the virtual character. *QuestSim* [33] and *QuestEnvSim* [16] use Reinforcement Learning techniques to estimate the pose from sparse tracking sensors of a physical avatar and make it adaptive to its virtual environment.

## 2.2 Articulated Avatar Animation from Multimodal Data

Due to the difficulty to animate a virtual character's full-body from sparse sensors, often leading to pose ambiguity, some approaches have been proposed to combine these signals with motion features from other sources. Since RGB videos have been successfully exploited to tackle the problematic of real-time human pose reconstruction [12, 13, 21, 22, 42], videos of the scene have been used to extend IMU sensors information, even with a single point of view. This procedure has shown encouraging results in accurate motion tracking [11, 17, 20, 29]. More recently, *EgoLocate* [41] fuses the data from the egocentric view captured by a monocular camera and the signals from the IMU tracking devices to precisely animate and localize the virtual character in the environment.

Additional external sources such as LiDAR [27] or optical markers placed on strategic positions on the subject's body [2] are also employed in this context. The 3D points clouds data from the LiDAR is fused with IMU sensors to achieve a robust and accurate motion tracker for collecting motion data in large-scale scenarios.

VR applications also benefit from such data fusion. Wu *et al.* set up, additionally to the HMD and the handheld controllers, 4 RGB-D cameras (Kinect) and a LeapMotion for full body and hands tracking [35]. However, this setup can be cumbersome regarding the use case and a calibration process for each Kinect, which hinders a large-scale deployment of this method. To leverage this issue, a single external of-the-shelf RGB web camera has been integrated in the animation external setup [38]. This application employs a 2D pose estimator to reconstructs the human full-body positions from the RGB videos and extend the data from the VR trackers to estimate the virtual avatar's full pose.

# 3 Analysis

## 3.1 Dataset

Similarly to [9], we conduct our experiments on 3 subsets of the large-scale motion capture AMASS Dataset [19]: BMLrub [30], CMU [15] and HDM05 [23]. AMASS Dataset unifies optical-based motion capture datasets into a standard kinematic tree and use SMPL approach [18] to provide realistic 3D human meshes represented by a rigged body model. These subsets gather around 5200 motion samples for a duration of more than 20 hours. The standardized kinematic tree is structured into 22 joints.

## 3.2 Approach

The proposed method aims to study the effect of additional motion features on the full-body self-avatar's pose estimation and is described in Figure 1. We employ *AvatarPoser* [9], a Transformer-based architecture that encodes the sparse motion signals tracked by the VR devices. Then, the local orientations and the global motion navigation are learnt form the encoded motion representation. We refer the sparse motion signals as $X_{0,\ldots,T-1}$ gathering the 3D Cartesian positions $p = \{p_{Head}, p_{LeftHand}, p_{RightHand}\}$, the orientation $r$, the linear and angular velocity respectively denoted as $v^p$ and $v^r$, of the head and handheld controllers. The motion samples are considered in a temporal window of $T$ frames. In our experiments, we set $T = 40$ frames.

$$X_{0,\ldots,T-1} = \begin{bmatrix} p_0 & r_0 & v_0^p & v_0^r \\ p_1 & r_1 & v_1^p & v_1^r \\ \cdots & \cdots & \cdots & \cdots \\ p_{T-2} & r_{T-2} & v_{T-2}^p & v_{T-2}^r \\ p_{T-1} & r_{T-1} & v_{T-1}^p & v_{T-1}^r \end{bmatrix} \tag{1}$$

Our method proposes to concatenate the sparse motion signals from the VR tracking devices by $X^F_{0,\ldots,T-1}$. Hence, since $X_{0,\ldots,T-1} \in \mathbb{R}^{T\times 54}$ in *AvatarPoser* [9], $X_{0,\ldots,T-1} \oplus X^F_{0,\ldots,T-1} \in$ respectively $\mathbb{R}^{T\times 120}$ and $\mathbb{R}^{T\times 98}$ if $X^F_{0,\ldots,T-1}$ represents 3D or single-view 2D features.

$$X^F_{0,\ldots,T-1} = \begin{bmatrix} f_0 \\ f_1 \\ \cdots \\ f_{T-2} \\ f_{T-1} \end{bmatrix} \tag{2}$$

Here, $f$ plays the role of (a) the full-body avatar's 3D global Cartesian positions, then (b) the 2D local projection of these positions using calibrated RGB cameras parameters. These feature sets can be estimated from a setup of one or multiple cameras [7]. Therefore, no additional cumbersome and costly trackers that hamper the user experience are required. Concerning the 2D positions, we built a set up of 4 virtual calibrated cameras for the multi-view projection of the 3D Cartesian coordinates. This set up is shown in Figure 2. Our projection takes into account the extrinsic and intrinsic camera parameters, but we ignore the radial and tangential distortion induced by the cameras.

Our approach feeds $X_{0,\ldots,T-1} \oplus X^F_{0,\ldots,T-1}$ to *AvatarPoser* as proposed in [9] to generate the virtual character's full-body pose *i.e.*, the local orientations and the global root displacement. Then, the reconstruction error is measured between the ground truth and the predicted pose to study the impact of $X^F_{0,\ldots,T-1}$ in the full-body pose estimation.

## 3.3 Experiments

The experiments are split into 3 cases: $X^F_{0,\ldots,T-1}$ contains (1) the ground truth 3D Cartesian positions, then (2) a single-view 2D projection of these coordinates in each camera defined in the proposed setup and (3) the 2D projection from two distinct cameras to investigate

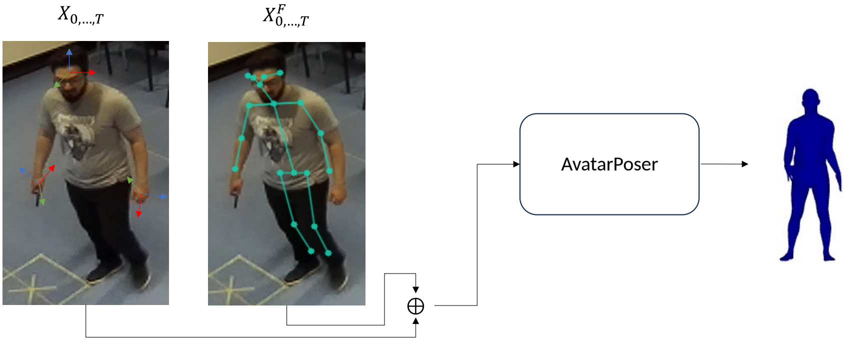

Figure 1: Overview of the proposed analyzes. The experiments aim to measure the impact of extending HMD and controllers sparse signals on the self-avatar's full-body motion reconstruction error in Transformer-based model. These analyzes employ *AvatarPoser* [9] architecture to estimate the local rotations and global displacement from the input motion features. In our experiments, $X^F_{0,\ldots,T-1}$ is either the 3D ground truth Cartesian positions or their 2D projection in a camera space.

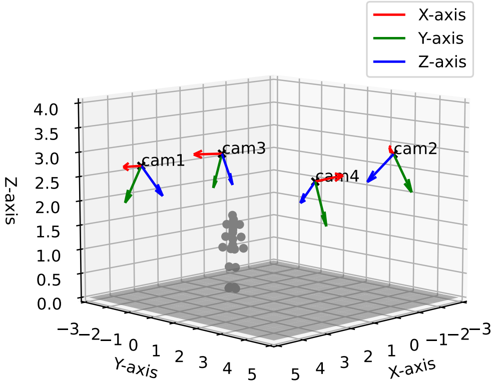

Figure 2: Virtual cameras set up. The 3D Cartesian positions are projected into each camera space and provide a set of multi-view 2D poses. The coordinate system in each camera space is represented in red (X-axis), green (Y-axis) and blue (Z-axis).

the influence of dual projection viewpoints. In each configuration, we train *AvatarPoser* [9] 10.000 epochs with a batch size of 256 on a Nvidia RTX3090 GPU. Then, the mean error of position, rotation and velocity between the model output and the ground truth are computed for each scenraio. This analysis is divided into the upper and lower body segments, aiming to accentuate potential dissimilarities in behavior between these two regions.

# 4 Results

To evaluate each configuration, we report the mean per joint positional (MPJPE), rotational (MPJRE) and velocity (MPJVE) error computed between the ground truth and the predicted motion for upper and lower body. Table 1 shows the results of the motion reconstruction in each tested configuration. The *3 inputs* row refers to the original experiments proposed in [9].

First of all, in every tested configuration, the Cartesian positions improve the full-body motion estimation: the reconstruction error is decreased for upper- and lower-body subsets. Considering the 3D Cartesian positions, the MPJPE and MPJVE metrics are not relevant. Indeed, although providing ground truth 3D Cartesian positions helps to accurately estimate the local rotations of the articulated avatar, the positions estimated by the model should not be used to animate the avatar since the ground truth positional information is available.

Then, our observation reveals that supplementing the inputs from the HMD and handheld controllers with 2D projected positions from two distinct perspectives yields reconstruction errors that are either comparable to or inferior than those attained by the model utilizing 3D Cartesian positions. This means that, estimating the full-body motion with *AvatarPoser* [9], it appears not necessary to make a 3D reconstruction of the 2D poses estimated on the RGB cameras if two different points of view are provided as additional information.

We also observe that providing even one single view of the 2D full-body positions, in addition to the sparse HMD and controllers inputs, guides efficiently the avatar's motion estimation. More importantly, the precision of the motion reconstruction is barely affected by the camera point of view. The largest error difference regarding each metric between the single camera configurations is (0.04cm, 0.13°,0.34cm/s) and (0.1cm,0.04°,1.04cm/s) for respectively upper- and lower-body motion. Indeed, the motion dataset gathers a large diversity of global motion direction. This ensures that variations in camera angles have minimal influence on the precision of the motion reconstruction process, as soon as the 2D projection in the camera space, regardless its position and orientation in world space, is used as training data.

Finally, the positional features impact more importantly the reconstruction of the lower-body motion compared to the upper-body motion. We believe that this effect comes from the fact that the lower-body information is not tracked using sparse sensors highlighting the pose ambiguity issue. Hence, providing lower-body information helps to resolve this pose ambiguity and efficiently guides the full-body poses estimation.

Examples of estimated pose samples can be found in Figure 3. The positional errors are highlighted on the SMPL model. We observe that, using only the sparse inputs, the positional errors are mainly located in the avatar's legs and feet due to the lack of tracking sensors in lower-body region. Augmenting the sparse information with the 3D or projected 2D Cartesian coordinates of the avatar's joints decreases the reconstruction error, especially in that region of the body. Figure 3 shows that this improvement is more significant providing 3D Cartesian positions than the 2D coordinates in a single viewpoint.

# 5 Discussion and Perspectives

The experiments described in Section 3 aims to study the behavior of a Transformer-based model that estimate the self-avatar's full poses regarding the motion features provided by the tracking solution. We consider that solution based on RGB-D cameras for the user to avoid

| | MPJPE up [cm] | MPJRE up [°] | MPJVE up [cm/s] |
|---|---|---|---|
| 3 inputs | 1.65 | 5.64 | 12.86 |
| 3 inputs + 3D positions | - | 2.52 | - |
| 3 inputs + 2D (cam 1) | 0.86 | 2.8 | 7.66 |
| 3 inputs + 2D (cam 2) | 0.89 | 2.93 | 8.00 |
| 3 inputs + 2D (cam 3) | 0.86 | 2.82 | 7.73 |
| 3 inputs + 2D (cam 4) | 0.9 | 2.90 | 7.92 |
| 3 inputs + 2D (cam 1 + 4) | **0.72** | 2.52 | 6.78 |
| 3 inputs + 2D (cam 2 + 4) | 0.73 | 2.55 | 6.81 |
| 3 inputs + 2D (cam 3 + 4) | **0.72** | **2.49** | **6.71** |

| | MPJPE low [cm] | MPJRE low [°] | MPJVE low [cm/s] |
|---|---|---|---|
| 3 inputs | 6.79 | 6.4 | 44.35 |
| 3 inputs + 3D positions | - | 2.02 | - |
| 3 inputs + 2D (cam 1) | 1.81 | 2.36 | 16.17 |
| 3 inputs + 2D (cam 2) | 1.88 | 2.38 | 16.66 |
| 3 inputs + 2D (cam 3) | 1.8 | 2.34 | 16.04 |
| 3 inputs + 2D (cam 4) | 1.9 | 2.34 | 17.08 |
| 3 inputs + 2D (cam 1 + 4) | 1.46 | 1.92 | 13.37 |
| 3 inputs + 2D (cam 2 + 4) | 1.47 | 1.9 | 13.4 |
| 3 inputs + 2D (cam 3 + 4) | **1.42** | **1.89** | **13.22** |

Table 1: Reconstruction error in each tested configuration. The Transformer-based animation model performance is positively impacted by the additional motion features, especially in the lower-body reconstruction.

wearing any tracking devices or sensors other than the HMD and the handheld controllers. The additional motion features are available in the dataset used for our experiments. Since the motion datasets have been recorded by optical systems, the 3D Cartesian positions are tracked with a high fidelity. However, estimating 3D Cartesian positions from one or multiple RGB-D cameras suffers from a lower precision and major artifacts such as occlusions. Further experiments should study the impact of these artifacts on the full poses estimation for a more realistic context.

Moreover, the estimation of such information might suffer from a computation time that is critical in the design of real-time applications. Although *AvatarPoser* [9] exhibits interesting real-time performance in the avatar's full poses estimation from sparse inputs, extending this method with 2D or 3D pose estimation from RGB videos might induce a larger delay and make the proposed framework irrelevant for real-time animation. Moreover, the data acquisition from different sources might be performed at different framerate, inducing a synchronization issue for the articulated avatar's motion estimation. These parameters should also be studied in future works.

Although we focus our work on single avatar full-body pose estimation, we believe that multi-avatar animation from sparse inputs can also be augmented with full-body Cartesian positions. However, within this scenario, implementing a subject identification algorithm would be advantageous for the animation framework. This algorithm would facilitate the assignment of joints to specific subjects.

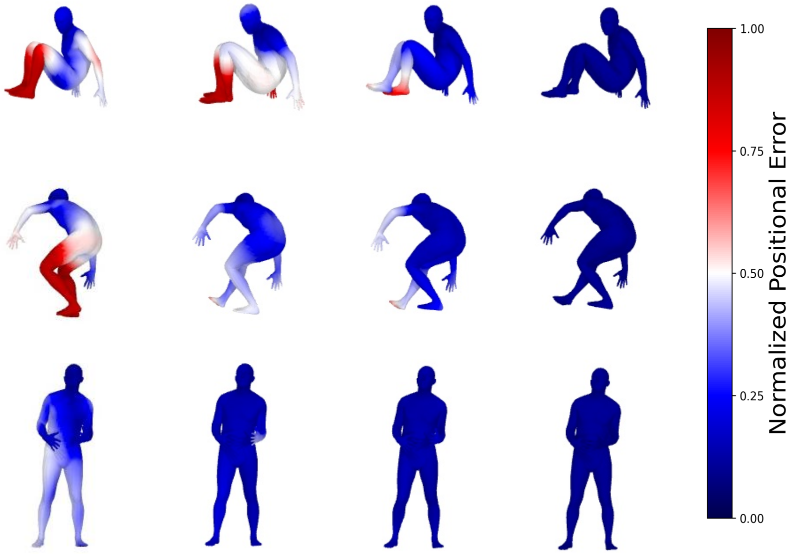

Figure 3: Examples of poses from (left to right): sparse 3 inputs, 3 inputs + 2D Cartesian positions projected on camera 4, 3 inputs + 3D Cartesian position and ground truth motion. Extending the sparse motion signals by the Cartesian position from an external source of motion capture improves the full-body motion reconstruction, especially in the reconstruction of the lower-body parts.

# 6 Conclusion

This work investigates the impact on extending orientations and positions of the headset and handheld controllers in the articulated self-avatar's full-body motion estimation. More precisely, we provide the 3D ground truth Cartesian positions as well as the 2D projection of these positions in the camera space. These features can be estimated from RGB-D cameras, which is suitable in the VR animation context, since we avoid to equipped users with external tracker devices which might hinder the user experience quality. We observe that these two set of additional features significantly improve the precision of the motion reconstruction, especially for the lower body. While this study presents promising outcomes within the realm of VR animation, we believe that future explorations focusing on potential motion artifacts inherent to pose estimation from RGB videos algorithms or delay between sources of motion could provide valuable insights in this domain.

# References

[1] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D. Wilson. Coolmoves: User motion accentuation in virtual reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(2), jun 2021. doi: 10.1145/3463499. URL https://doi.org/10.1145/3463499.

[2] Sheldon Andrews, Ivan Huerta, Taku Komura, Leonid Sigal, and Kenny Mitchell. Real-

time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European conference on visual media production (CVMP 2016)*, pages 1–10, 2016.

[3] Michael Büttner and Simon Clavet. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 1(2015):2, 2015.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] 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 *CVPR*, 2023.

[7] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003. ISBN 0521540518.

[8] 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.

[9] 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*. Springer, 2022.

[10] 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 Conference Papers, 2022. doi: 10.1145/3550469.3555428.

[11] Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Resolving position ambiguity of imu-based human pose with a single rgb camera. *Sensors*, 20(19), 2020. ISSN 1424-8220. doi: 10.3390/s20195453. URL `https://www.mdpi.com/1424-8220/20/19/5453`.

[12] Ning Kang, Junxuan Bai, Junjun Pan, and Hong Qin. Interactive animation generation of virtual characters using single rgb-d camera. *The Visual Computer*, 35:1–12, 06 2019. doi: 10.1007/s00371-019-01678-7.

[13] Ning Kang, Junxuan Bai, Junjun Pan, and Hong Qin. Real-time animation and motion retargeting of virtual characters based on single rgb-d camera. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1006–1007, 2019. doi: 10.1109/VR.2019.8797856.

[14] Meejin Kim and Sukwon Lee. Fusion poser: 3d human pose estimation using sparse imus and head trackers in real time. *Sensors*, 22(13), 2022. ISSN 1424-8220. doi: 10.3390/s22134846. URL `https://www.mdpi.com/1424-8220/22/13/4846`.

[15] CMU Graphics Lab. Cmu graphics lab motion capture database., 2000. URL http://mocap.cs.cmu.edu/.

[16] Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. *arXiv preprint arXiv:2306.05666*, 2023.

[17] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1539–1548, 2023.

[18] 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.

[19] 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, October 2019.

[20] Charles Malleson, John Collomosse, and Adrian Hilton. Real-time multi-person motion capture from multi-view video and imus. *International Journal of Computer Vision*, 128:1594–1611, 2020.

[21] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.*, 36(4), jul 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073596. URL https://doi.org/10.1145/3072959.3073596.

[22] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019.

[23] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database hdm05. *Computer Graphics Technical Report CG-2007-2, Universität Bonn*, 2007.

[24] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021.

[25] Jose Luis Ponton, Haoran Yun, Carlos Andujar, and Nuria Pelechano. Combining Motion Matching and Orientation Prediction to Animate Avatars for Consumer-Grade VR Devices. *Computer Graphics Forum*, 41(8):107–118, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14628.

[26] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6040–6049, 2020.

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

[28] Pedro Manuel Santos Ribeiro, Ana Clara Matos, Pedro Henrique Santos, and Jaime S. Cardoso. Machine learning improvements to human motion tracking with imus. *Sensors*, 20(21), 2020. ISSN 1424-8220. doi: 10.3390/s20216383. URL https://www.mdpi.com/1424-8220/20/21/6383.

[29] Soyong Shin, Zhixiong Li, and Eni Halilaj. Markerless motion tracking with noisy video and imu data. *IEEE transactions on bio-medical engineering*, PP, 05 2023. doi: 10.1109/TBME.2023.3275775.

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

[31] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 197–212. Springer, 2020.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[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.

[34] Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, and Robert W. Lindeman. Towards an articulated avatar in vr: Improving body and hand tracking using only depth cameras. *Entertainment Computing*, 31:100303, 2019. ISSN 1875-9521. doi: https://doi.org/10.1016/j.entcom.2019.100303. URL https://www.sciencedirect.com/science/article/pii/S1875952119300138.

[35] Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, and Robert W Lindeman. Towards an articulated avatar in vr: Improving body and hand tracking using only depth cameras. *Entertainment Computing*, 31:100303, 2019.

[36] Di Xia, Yeqing Zhu, and Heng Zhang. Faster deep inertial pose estimation with six inertial sensors. *Sensors*, 22(19), 2022. ISSN 1424-8220. doi: 10.3390/s22197144. URL https://www.mdpi.com/1424-8220/22/19/7144.

[37] 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.

[38] Jackie Yang, Tuochao Chen, Fang Qin, Monica S Lam, and James A Landay. Hybridtrak: adding full-body tracking to vr using an off-the-shelf webcam. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[39] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 40(4), jul 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459786. URL https://doi.org/10.1145/3450626.3459786.

[40] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[41] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *arXiv preprint arXiv:2305.01599*, 2023.

[42] Anastasios Yiannakides, Andreas Aristidou, and Yiorgos Chrysanthou. Real-time 3d human pose and motion reconstruction from monocular rgb videos. *Computer Animation and Virtual Worlds*, 30(3-4):e1887, 2019. doi: https://doi.org/10.1002/cav.1887. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.1887. e1887 cav.1887.

[43] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6), 2020.

[44] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

[45] Xinkang Zhang, Xinrong Chen, Xiaokun Dai, and Xinhan Di. Dual attention poser: Dual path body tracking based on attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2794–2803, 2023.