Impact of 3D Cartesian Positions and Occlusion on Self-Avatar Full-Body Animation in Virtual Reality

Antoine Maiorca*, Adrien Kinart*, George Fletcher†, Seyed Abolfazl Ghasemzadeh‡,
Thierry Ravet*, Christophe De Vleeschouwer‡, and Thierry Dutoit*
*Numediart Institute, ISIA Lab, University of Mons (UMONS), Mons, Belgium
† Department of Computer Science, University of Bath, Bath, United Kingdom
‡ICTEAM/ELEN, UCLouvain, Louvain-la-Neuve, Belgium
{antoine.maiorca, adrien.kinart, thierry.ravet, thierry.dutoit}@umons.ac.be
gf321@bath.ac.uk
{seyed.ghasemzadeh, christophe.devleeschouwer}@uclouvain.be

***Abstract*—This study explores methods for enhancing full-body self-avatar animation in virtual reality (VR), particularly by leveraging 3D Cartesian positions of additional body joints. Self-avatar animation is essential to foster embodiment in virtual environments. However, it lacks accuracy when tracking signals are limited to head and hand controllers. We evaluated three state-of-the-art open source Transformer-based models to take advantage of lower- and upper-body joint data and analyze their impact on motion-tracking accuracy. The results show that augmenting the sparse VR inputs with the lower body joints (*e.g.* feet, knees) significantly improves the accuracy of the avatar, even for upper body motion. However, occlusion of these joints critically degrades the animation quality in every tested configuration, particularly in terms of temporal continuity and smoothness. We show that effectively handling occlusion artifacts is a crucial challenge in this research field.**

***Index Terms*—Virtual Reality, Animation, Deep Learning.**

I. Introduction

In VR systems, a self-avatar represents the user digitally within the virtual environment. This avatar replicates the gestures and movements of the user, allowing them to perceive their presence in the virtual world. The primary goal of a self-avatar is to strengthen the sense of embodiment and immersion. Studies have shown that its various attributes, such as appearance [35, 36, 19, 43, 15] and size/shape/movement [6, 32], significantly influence virtual body ownership and presence. The sense of embodiment, how strongly a person feels that the virtual body is their own [20], can be negatively affected by mismatches between the user's physical movements and the avatar's actions [37, 36].

Hence, the development of robust and reliable animation systems is essential in the VR industry [48]. In practice, consumer-grade VR products often only provide three 6DoF tracking signals; the Head-Mounted Display (HMD) and hand tracking. Within these constraints, full-body animation of the articulated self-avatar is ill-posed. However, various deep learning models have considered turning those few cues into a full-body animation pattern [10, 44, 2, 17, 7, 12, 9, 50, 3, 8, 33, 13, 11].

To enhance the quality of the user's motion tracking, a different approach consists of extending the motion signals tracked by the three VR devices (HMD and handheld controllers) by information from the environment [21], IMUs [46, 31] or RGB(D) camera(s) [47, 28]. In the latter case, user's full-body 3D Cartesian positions can be extracted with pose estimation from multi-view or monocular videos [49]. A major benefit of a camera-based setup is that it eliminates the need for additional trackers on the user's body, ensuring a seamless experience without hindering comfort or mobility. Nevertheless, major drawbacks can hinder the deployment of this kind of solution, *e. g.*,the inability to handle occlusion– phenomenon where parts of a subject's body are partially or fully blocked from view– or inaccuracy of joint position estimation.

In this paper, we investigate the impact on the user's motion tracking of (1) incorporating 3D Cartesian positions of strategically chosen pairs of joints and (2) the occlusion artifacts. We evaluated this across three open-source state-of-the-art Transformer-based models designed to animate the full-body of an articulated avatar using sparse input signals: *AvatarPoser* [17], *AvatarJLM* [50] and *HMD-Poser* [8]. We show that the quality of the generated animation improves when 3D Cartesian positions of selected joints are available, but is critically impacted by occlusion artifacts.

II. Related Work

A. Self-avatar full-body estimation from sparse signals

Estimating full-body motion from sparse sensors has been widely researched, particularly for VR applications, where self-avatar movements are inferred from head and hand data. Traditional methods use Inverse Kinematics (IK) to compute these poses [29, 4], while machine learning techniques, like k-NN and motion matching, have advanced this with more realistic motion synthesis from sparse data. For instance, *CoolMoves* [1] and *MMVR* [30] offer nearest-neighbour-based solutions, though require diverse and seamless motion databases to handle complex upper body movements.

Deep learning, especially using recurrent and transformer models [42], enhance motion reconstruction by learning spatial-temporal relationships from sparse signals. Methods like *AvatarPoser*[17] and *Dual Attention Poser*[9] leverage

transformers to encode motion, while hybrid models [3, 8, 33] integrate RNNs with transformers to handle tracking interruptions effectively and reduce computational complexity.

Various architectures have been applied to this problem such as *VAE-HMD* [10] and *FLAG* [2] use VAEs and Normalizing Flows [34] to reconstruct full poses, with more recent diffusion models like *BoDiffusion* [7] improving quality, albeit with high inference times. Efficient diffusion methods, such as *AGrOL* [12], are emerging to address this issue. Lastly, reinforcement learning-based methods use physical simulation, as opposed to learning purely from data, for realistic full-body pose reconstruction [44, 21].

B. Multimodal full-body pose estimation

To address pose ambiguity, additional motion sources have been used to enhance VR tracking data from headsets and controllers. *EgoLocate* [47] merges 6 *Inertial Measurement Units* (IMUs) with Simultaneous Localization and Mapping techniques (SLAM) for integrated motion capture, localization, and mapping, while Tome et al [39] introduce a neural network that resolves perspective and self-occlusion issues in both egocentric and standard camera setups. Additionally *BodyTrak* [23] uses wrist-mounted cameras to infer full-body poses, pioneering a new approach to capture poses from body silhouette images.

Inspired by monocular 2D pose estimation, *HybridTrak* [45] combines inside-out tracking with uncalibrated RGB cameras to capture lower-body movement, providing accessible and low-cost tracking. Liao et al. advance 3D body reconstruction by fusing IMUs and RGB camera data with adaptive learning, resulting in robust and accurate joint tracking, even in occluded or challenging environments [22]. While the last method does not rely on any pose estimation methods, [39, 28] extract the 2D pose from a monocular video and fuse it with 6 IMUs sensors.

Intrinsically, models that rely solely upon VR head and hand tracking signals struggle with lower body motion due to inherent statistical ambiguity. With the goal of improving such models in the future, our work analyses the change in accuracy of state-of-the-art deep learning-based VR avatar animation models when they are augmented with the 3D positions of body joints. We also analyze the impact of applying occlusion to these 3D positions so as to better emulate real pose estimators (e.g. HybridTrak) [45]).

III. Experiments

Animating the full body of an articulated virtual character from sparse tracking sensors can be formulated as:

$$\tilde{y}_T = \Theta(x_t) \tag{1}$$

where $\tilde{y}_T$ represents the full-body motion features estimated at frame T, consisting of the global root translation and orientation and the remaining local joint rotations for a given topology (i.e. SMPL [24]), computed by the animation function Θ. Here, Θ is modeled by a deep learning model taking input sequence x_t for $t = 0, ..., T$, derived from VR tracking data (position, orientation, and corresponding velocities of tracked joints).

We propose to extend the motion signals from VR tracking devices with additional joint Cartesian positions and velocities. Often, pose estimators for RGB videos compute the Cartesian position of body joints such as knees or shoulders. To facilitate common pose estimation methods, we chose shoulders, elbows, knees, and feet as additional joints j, as these tend to be the most consistent across different skeletal topologies (*e.g.*,SMPL[24] as used in this work, COCO-WholeBody[18], MediaPipe[5]). We consider different combinations of these joints selected among J, in order to measure the impact of each pair of joints on the animation quality.

We employ three Transformer-based architectures that animate the self-avatar from VR sparse inputs: *AvatarPoser*[17], *AvatarJLM*[50] and *HMD-Poser*[8]. We train and evaluate these models on 3 subsets of the motion capture dataset AMASS [25] –CMU Motion Capture Database [40], BMLrub[41] and HDM05[27], as proposed in these works. The sequence x_t of head and hands motion features are extracted from these datasets. Then, this sequence feeds the model to animate the avatar.

Our proposed analysis augments the sequence x_t with additional spatial information, $x_t^{p_j}$ and $x_t^{v_j}$, which represent the ground-truth Cartesian positions and linear velocities of the defined body joints, respectively. $x_t^{p_j}$ and $x_t^{v_j}$ are concatenated to the position and linear velocity of the VR tracking data. We train the models from scratch for each combination of these joints to reconstruct the full-body pose. We keep the same hyperparameters and training procedure as defined in each respective released source codes[1] [2] [3]

Next, we assess the impact of occluding a specific joint j on the quality of the avatar animation. To mimic occlusion, we assume that the information related to the occluded joint j, *i.e.*, $x_t^{p_j}$ and $x_t^{v_j}$ are masked out, replacing the actual value by (1) zero or (2) the mean across the dataset. First, we occlude a portion σ_o of the whole sequence length N. We randomly select $\lfloor N\sigma_o \rfloor$ frames for which the joint $j \in J$ is occluded.

As evaluation metrics, we use the Mean Per Joint Position/Rotation/Velocity Error, respectively denoted as MPJPE, MPJRE, MPJVE as well as the Jitter ratio (*i.e.*, ratio between output and ground truth jitter effect), on the resulting animation. We also compute the positional error separately for the upper body (head, arms, and torso) and the lower body (feet, legs, and lower spine) denoted respectively Up and Low PE.

IV. Results

A. Additional body joint analysis

Table I presents evaluation metrics for the three different animation models using various configurations of joint inputs. The models are extended with 3D Cartesian positions of specific body joints $j \in J$ where

[1]https://github.com/Pico-AI-Team/HMD-Poser
[2]https://github.com/eth-siplab/AvatarPoser
[3]https://github.com/zxz267/AvatarJLM

TABLE I: Evaluation metrics: MPJRE (°), MPJVE (cm/s), Up PE (cm), Low PE (cm) and Jitter ratio in each configuration. Additional body joints are provided to extend the sparse motion signals (head and hand global positions and orientations). Even though providing the full set of joints leads to most accurate motion tracking, lower-body motion signals seem most effective, even for accurate upper-body motion. Moreover, providing only upper-body joints has a smaller effect on achieving accurate upper- and lower-body animations compared to providing only the lower-body joints.

Inputs	AvatarJLM					HMD-Poser					AvatarPoser				
	MPJRE$_\downarrow$	MPJVE$_\downarrow$	Up PE$_\downarrow$	Low PE$_\downarrow$	Jitter$_\downarrow$	MPJRE$_\downarrow$	MPJVE$_\downarrow$	Up PE$_\downarrow$	Low PE$_\downarrow$	Jitter$_\downarrow$	MPJRE$_\downarrow$	MPJVE$_\downarrow$	Up PE$_\downarrow$	Low PE$_\downarrow$	Jitter$_\downarrow$
VR baseline	5.96	19.78	1.27	5.50	1.75	2.28	17.34	1.86	5.4	1.47	5.97	27.16	1.65	6.79	3.66
F	5.02	**10.49**	1.09	**2.04**	**1.04**	1.92	**11.49**	1.67	**2.57**	**1.28**	**6.22**	**19.33**	1.9	**3.68**	**2.26**
K	**4.98**	13.05	0.97	2.42	1.38	**1.88**	11.95	**1.62**	2.78	1.3	6.24	21.35	1.75	4.22	2.56
E	5.61	18.69	0.95	5.55	1.64	2.17	16.92	1.78	5.25	1.55	6.85	29.02	1.75	8.49	3.55
S	5.38	18.06	**0.82**	5.03	1.85	2.18	16.63	1.66	5.12	1.52	6.85	28.56	**1.69**	8.23	3.55
FK	4.49	**8.43**	0.93	**1.19**	1.12	**1.7**	**9.87**	**1.53**	**1.88**	**1.24**	5.71	**17.13**	1.74	**2.84**	**1.95**
FS	4.50	9.97	0.72	1.90	1.29	1.8	10.77	1.67	2.42	1.26	5.73	18.2	1.61	3.46	2.14
FE	4.48	9.79	0.799	1.99	**1.1**	1.84	11.27	1.71	2.57	1.28	**5.7**	18.45	1.62	3.56	2.17
KS	**4.47**	11.59	**0.66**	2.37	1.31	1.76	11.28	1.61	2.62	1.28	5.77	20.5	1.51	4.08	2.53
KE	4.54	11.51	0.73	2.41	1.17	1.78	11.64	1.54	2.72	1.3	5.69	20.52	1.49	4.15	2.49
ES	5.30	19.01	0.76	5.18	2.05	2.1	16.67	1.58	5.12	1.52	6.56	28.15	1.57	8.18	3.49
KSE	4.29	13.47	**0.58**	2.42	1.93	1.7	10.97	1.45	2.58	1.28	5.42	20.02	**1.38**	4.04	2.44
FES	4.26	9.31	0.65	1.89	1.19	1.74	10.59	1.52	2.37	1.25	5.41	17.94	1.49	3.43	2.18
FKE	4.03	**7.15**	0.68	1.19	**1.02**	1.62	9.44	1.46	1.82	1.2	**5.17**	16.33	1.47	2.74	**1.99**
FKS	**3.99**	9.16	0.64	**1.18**	1.6	**1.6**	**9.15**	**1.45**	**1.81**	**1.19**	5.23	**16.13**	1.48	**2.71**	**1.99**
FKES	**3.78**	8.65	**0.54**	**1.18**	1.54	**1.53**	**8.79**	**1.42**	**1.78**	**1.16**	**4.91**	**15.81**	**1.38**	**2.71**	**1.94**

$$J = \left\{ \begin{array}{l} \text{right foot, left foot, right knee, left knee,} \\ \text{right elbow, left elbow, right shoulder, left shoulder} \end{array} \right\}$$

in addition to motion signals from VR devices (head and handheld controllers' motion features). Alongside the baseline VR motion signals from the head and handheld controllers, these models are evaluated using configurations that include each joint individually, all possible pairs, and select multi-joint combinations from J. The configurations cover a variety of combinations, such as F (feet), K (knees), E (elbows), and S (shoulders), as well as multi-joint combinations like FK (feet and knees), ES (elbows and shoulders), FE (feet and elbows), FKS (feet, knees, and shoulders), $FKES$ (feet, knees, elbows, and shoulders), and so on. These configurations enable us to assess the impact of each joint or set of joints on the animation accuracy. A few samples of generated and ground truth poses are represented in Figure 1 and qualitative results in video format are available here [4].

First, we observe that, in some configuration for *AvatarPoser*, extending the sparse inputs with Cartesian positions does not imply more accurate animation output. The MPJRE and Up PE metrics for F, K, E or S configurations are higher in comparison with the model trained with VR features only. This outcome likely arises because *AvatarPoser* lacks awareness of the spatial arrangement of the skeleton's topology. Unlike *AvatarJLM* and *HMD-Poser* that are equipped with spatial attention mechanisms, *AvatarPoser* handles motion features as simple concatenated inputs rather than incorporating them as structured joint information, which hinders its ability to capture spatial relationships between body parts effectively.

We also see that providing lower-body joints significantly improve errors, even for upper-body animation. Indeed, we observe that the combination of feet and knees alone yields notable improvement for *AvatarJLM* and *HMD-Poser*. While including lower-body joints enhances the animation quality of the upper-body in the avatar, configurations such as E, S, or even ES face challenges in accurately tracking lower-body poses.

[4]https://figshare.com/s/7957da90bbb4e7bd3dad

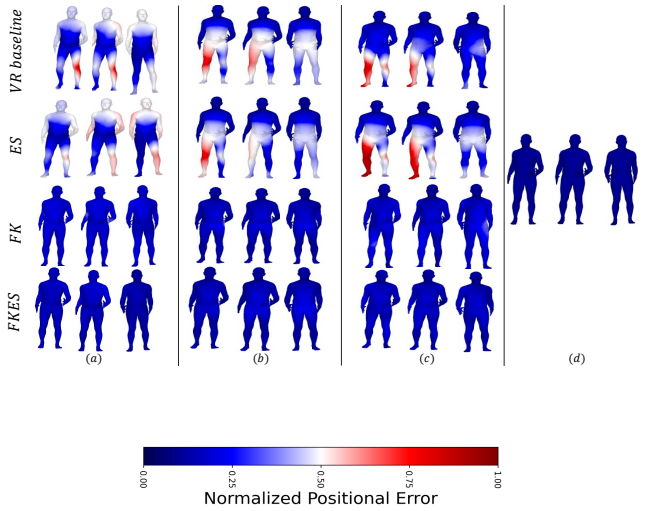

Fig. 1: Samples of predicted poses by (a) *HMD-Poser*, (b) *AvatarLM* and (c) *AvatarJLM* in baseline (VR), ES, FK and $FKES$ configurations. The positional error of vertices are highlighted in red on the body mesh. These qualitative results show that lower body joints are essential information to accurately animate user's avatar in virtual reality. In (d), the ground truth poses are shown as reference.

B. Impact of occlusion on animation

Fig. 2 shows the evolution of errors with σ_o when the occluded motion features are replaced by either zero or the

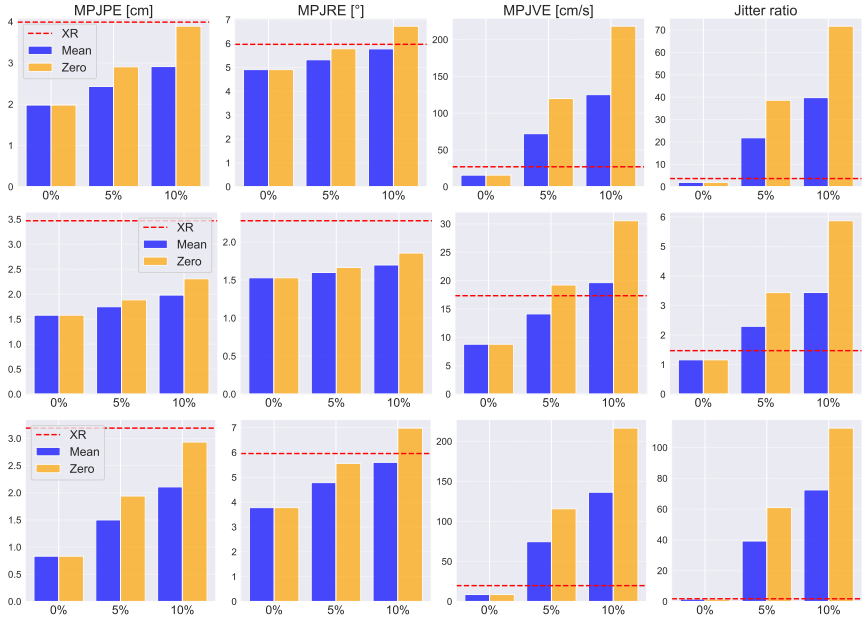

Fig. 2: Error metrics when $\lfloor N\sigma_o \rfloor$ randomly selected frames are masked, either by zero or by the mean value for $\sigma_o = 0$, 0.05 and 0.1. The displayed errors are the mean calculated across all errors obtained when each joint $j \in J$ is occluded individually. The animations are computed by **Top**: *AvatarPoser*, **Mid**: *HMD-Poser*, and **Bottom**: *AvatarJLM*. Occlusion significantly affects the velocity profile and jitter, reducing the efficiency of using 3D Cartesian positions compared to the baseline, which relies solely on motion signals from VR trackers.

mean value. This graph presents the average results when each joint $j \in J$ is occluded individually. First of all, across all models, increasing the percentage of masked frames leads to a rise in all error metrics indicating a decline in performance with more missing data. The MPJVE and jitter are impacted severely, likely due to the disruption of temporal continuity caused by masked frames, making it challenging for the animation model to maintain accurate velocity profiles and smooth transitions. As a result, this leads to increased MPJVE and jitter ratios. This impact is significant, even for a low σ_o and for both masking strategies, as it results in higher error than the VR trackers-only baseline, undermining the effectiveness of adding 3D Cartesian positions. Finally, the results indicate that *HMD-Poser* is more robust to occlusion compared to the other two models, since the ratios of MPJP/R/VE and the jitter effect between $\sigma_o = 0$ and $\sigma_o = 0.1$ are lower than for *AvatarJLM* and *AvatarPoser*.

V. Limitations and Perspectives

While we observe that 3D Cartesian positions, especially those from lower body joints, help to improve the accuracy of avatar animation, our analyses and experiments rely on ground-truth positions, which is a purely synthetic setup. Ground truth motion is typically captured using highly accurate motion capture systems that provide precise and complete information about human poses, including joint positions and orientations. However, when utilizing 3D data from monocular or multi-view pose estimation, the artifacts inherent to these methods must be carefully considered. Since the models are trained on ground truth 3D positions (captured by optical marker-based motion capture system) from subsets of the AMASS dataset, they lack robustness to such issues, limiting the effectiveness of the animation algorithms in real-world scenarios.

Managing occlusion can be performed via two approaches: by the pose estimator or the animation model. *HMD-Poser* or *ReliaAvatar* proposed protocols to handle loss of motion signals from IMU trackers. The training procedure involves masking out information from specific trackers, which enables *HMD-Poser* to handle multiple configurations of IMUs and makes *ReliaAvatar* robust to the short- or long-term loss of tracking signals. On the other hand, various pose estimation methods address occlusion more directly [38, 16].

Moreover, pose estimation from monocular or multi-view RGB cameras offers a less accurate solution than optical motion capture with markers to track the 3D body pose. We believe that the discrepancy between the accuracy of the training data from AMASS and the inference data from 3D reconstruction based on RGB videos would be critical in the deployment of such solutions; especially for monocular pose estimation, as a single RGB camera is often already available to a typical VR consumer. To address this problem, one solution is to apply pose estimation to AMASS renders using *BodyVisualizer*[5] and *Trimesh*[6], as proposed in various works[14, 26]. This would allow comparison of models trained

[5]https://github.com/nghorbani/body_visualizer
[6]https://trimesh.org/

with ground truth and estimated 3D Cartesian positions and further measure the impact of the pose estimation performance on the resulting full-body motion tracking.

Also, in this work, we consider that the 3D positions from feet, knees, elbows or shoulders are both synchronized with, and as frequent as, the HMD signals. In an online solution, this may not be the case as data acquisition rate from VR trackers and cameras may differ, and the inference time of an accurate 3D pose estimation will add some latency. Fortunately, promising methods are available to deal with asynchronous sparse observations in Transformers [38].

Finally, to validate the effectiveness of methods aimed at mitigating occlusion effects, further analysis across additional datasets is necessary; using more subsets of AMASS or employing custom datasets as in prior works [31, 50].

VI. Conclusion

This paper addresses the task of full-body articulated self-avatar animation for VR. We employ three state-of-the-art Deep Learning-based models that animate the avatar from motion data tracked by consumer VR devices. We extend their motion features by 3D Cartesian positions of several pairs of joints, that could be extracted by off-the-shelf pose estimators. Our analyses highlight the usefulness of providing lower-body information to models in the resulting avatar's animation. Finally, we show that effectively managing occlusion artifacts is crucial for ensuring high-quality animation, since each tested animation model failed to mitigate them. We hope that this work will encourage research into improving the quality of full-body self-avatar animation for VR devices via external cameras.

Acknowledgment

This research is partially supported and funded by TRAIL Institute. S.A. Ghasemzadeh is funded by FRIA/FNRS. G. Fletcher is funded by UKRI (EP/S023437/1).

References

[1] Karan Ahuja et al. "CoolMoves: User Motion Accentuation in Virtual Reality". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5.2 (June 2021). DOI: 10.1145/3463499. URL: https://doi.org/10.1145/3463499.

[2] Sadegh Aliakbarian et al. "FLAG: Flow-based 3D Avatar Generation from Sparse Observations". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13243–13252. URL: https://api.semanticscholar.org/CorpusID:247411442.

[3] Sadegh Aliakbarian et al. "HMD-NeMo: Online 3D Avatar Motion Generation From Sparse Observations". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 9588–9597. URL: https://api.semanticscholar.org/CorpusID:261064795.

[4] Andreas Aristidou and Joan Lasenby. "FABRIK: A fast, iterative solver for the Inverse Kinematics problem". In: *Graph. Models* 73.5 (Sept. 2011), pp. 243–260. ISSN: 1524-0703. DOI: 10.1016/j.gmod.2011.05.003. URL: http://dx.doi.org/10.1016/j.gmod.2011.05.003.

[5] Valentin Bazarevsky and Ivan Grishchenko. "On-device, real-time body pose tracking with mediapipe blazepose". In: *Google AI Blog* (2020).

[6] Lauren Buck, Soumyajit Chakraborty, and Bobby Bodenheimer. "The Impact of Embodiment and Avatar Sizing on Personal Space in Immersive Virtual Environments". In: *IEEE transactions on visualization and computer graphics* PP (Feb. 2022). DOI: 10.1109/TVCG.2022.3150483.

[7] Angela Castillo et al. "BoDiffusion: Diffusing Sparse Observations for Full-Body Human Motion Synthesis". In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2023), pp. 4223–4233. URL: https://api.semanticscholar.org/CorpusID:258291803.

[8] Peng Dai et al. "HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 874–884. URL: https://api.semanticscholar.org/CorpusID:268253060.

[9] Xinhan Di et al. "Dual attention poser: dual path body tracking based on attention". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2794–2803.

[10] Andrea Dittadi et al. "Full-Body Motion from a Single Head-Mounted Device: Generating SMPL Poses from Partial Observations". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 11667–11677. DOI: 10.1109/ICCV48922.2021.01148.

[11] Kun Dong et al. "Realistic Full-Body Motion Generation from Sparse Tracking with State Space Model". In: *ACM Multimedia 2024*. 2024. URL: https://openreview.net/forum?id=dA6yat7UNM.

[12] Yuming Du et al. "Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 481–490. URL: https://api.semanticscholar.org/CorpusID:258187221.

[13] Han Feng et al. "Stratified Avatar Generation from Sparse Observations". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 153–163. URL: https://api.semanticscholar.org/CorpusID:270199440.

[14] Seyed Abolfazl Ghasemzadeh, Alexandre Alahi, and Christophe De Vleeschouwer. "MPL: Lifting 3D Human Pose from Multi-view 2D Poses". In: *arXiv preprint arXiv:2408.10805* (2024).

[15] 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)*. 2017, pp. 233–234. DOI: 10.1109/3DUI.2017.7893357.

[16] Buzhen Huang et al. “Occluded human body capture with self-supervised spatial-temporal motion prior”. In: *arXiv preprint arXiv:2207.05375* (2022).

[17] Jiaxi Jiang et al. “AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing”. In: *European Conference on Computer Vision*. 2022. URL: https://api.semanticscholar.org/CorpusID:251135349.

[18] Sheng Jin et al. “Whole-Body Human Pose Estimation in the Wild”. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 196–214. ISBN: 978-3-030-58544-0. DOI: 10.1007/978-3-030-58545-7_12. URL: https://doi.org/10.1007/978-3-030-58545-7_12.

[19] Dongsik Jo et al. “The impact of avatar-owner visual similarity on body ownership in immersive virtual reality”. In: *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. VRST ’17. Gothenburg, Sweden: Association for Computing Machinery, 2017. ISBN: 9781450355483. DOI: 10.1145/3139131.3141214. URL: https://doi.org/10.1145/3139131.3141214.

[20] Konstantina Kilteni, Raphaela Groten, and Mel Slater. “The Sense of Embodiment in Virtual Reality”. In: *Presence Teleoperators &amp Virtual Environments* 21 (Nov. 2012). DOI: 10.1162/PRES_a_00124.

[21] Sunmin Lee et al. “QuestEnvSim: Environment-Aware Simulated Motion Tracking from Sparse Sensors”. In: *ACM SIGGRAPH 2023 Conference Proceedings* (2023). URL: https://api.semanticscholar.org/CorpusID:259129290.

[22] Xianhua Liao et al. “Reconstructing 3D human pose and shape from a single image and sparse IMUs”. In: *PeerJ Computer Science* 9 (2023), e1401.

[23] Hyunchul Lim et al. “BodyTrak: Inferring Full-body Poses from Body Silhouettes Using a Miniature Camera on a Wristband”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.3 (Sept. 2022). DOI: 10.1145/3552312. URL: https://doi.org/10.1145/3552312.

[24] Matthew Loper et al. “SMPL: a skinned multi-person linear model”. In: *ACM Trans. Graph.* 34.6 (Oct. 2015). ISSN: 0730-0301. DOI: 10.1145/2816795.2818013. URL: https://doi.org/10.1145/2816795.2818013.

[25] Naureen Mahmood et al. “AMASS: Archive of Motion Capture As Surface Shapes”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 5441–5450. URL: https://api.semanticscholar.org/CorpusID:102351100.

[26] Antoine Maiorca et al. “Self-Avatar Animation in Virtual Reality: Impact of Motion Signals Artifacts on the Full-Body Pose Reconstruction”. In: *arXiv preprint arXiv:2404.18628* (2024).

[27] M. Müller et al. *Documentation Mocap Database HDM05*. Tech. rep. CG-2007-2. Universität Bonn, June 2007.

[28] Shaohua Pan et al. “Fusing Monocular Images and Sparse IMU Signals for Real-time Human Motion Capture”. In: *SIGGRAPH Asia 2023 Conference Papers* (2023). URL: https://api.semanticscholar.org/CorpusID:261494051.

[29] Donald L. Peiper. *THE KINEMATICS OF MANIPULATORS UNDER COMPUTER CONTROL*. Tech. rep. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1968. URL: https://apps.dtic.mil/sti/citations/AD0680036.

[30] J. L. Ponton et al. “Combining Motion Matching and Orientation Prediction to Animate Avatars for Consumer-Grade VR Devices”. In: *Computer Graphics Forum* 41.8 (2022), pp. 107–118. DOI: https://doi.org/10.1111/cgf.14628. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14628. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14628.

[31] Jose Luis Ponton et al. “SparsePoser: Real-time Full-body Motion Reconstruction from Sparse Data”. In: *ACM Trans. Graph.* 43.1 (Oct. 2023). ISSN: 0730-0301. DOI: 10.1145/3625264. URL: https://doi.org/10.1145/3625264.

[32] Jose Luis Ponton et al. “Stretch your reach: Studying Self-Avatar and Controller Misalignment in Virtual Reality Interaction”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. New York, NY, USA: Association for Computing Machinery, May 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642268. URL: https://doi.org/10.1145/3613904.3642268.

[33] Bo Qian et al. “ReliaAvatar: A Robust Real-Time Avatar Animator with Integrated Motion Prediction”. In: *arXiv preprint arXiv:2407.02129* (2024).

[34] Danilo Rezende and Shakir Mohamed. “Variational inference with normalizing flows”. In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[35] Anca Salagean et al. “Meeting Your Virtual Twin: Effects of Photorealism and Personalization on Embodiment, Self-Identification and Perception of Self-Avatars in Virtual Reality”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581182. URL: https://doi.org/10.1145/3544548.3581182.

[36] Anca Salagean et al. “The Utilitarian Virtual Self – Using Embodied Personalized Avatars to Investigate Moral Decision-Making in Semi-Autonomous Vehicle Dilemmas”. In: *IEEE Transactions on Visualization and Computer Graphics* 30.5 (2024), pp. 2162–2172. DOI: 10.1109/TVCG.2024.3372121.

[37] Aubrieann Schettler et al. “Visual Self-Motion Feedback Affects the Sense of Self in Virtual Reality”. In: *Multisensory Research* 34.3 (2020), pp. 323–336. DOI:

10.1163/22134808-bja10043. URL: https://brill.com/view/journals/msr/34/3/article-p323_6.xml.

[38] Sindhu Tipirneni and Chandan K Reddy. "Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.6 (2022), pp. 1–17.

[39] Denis Tome et al. "Selfpose: 3d egocentric pose estimation from a headset mounted camera". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2020), pp. 6794–6806.

[40] Fernando De la Torre Frade et al. *Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database*. Tech. rep. CMU-RI-TR-08-22. Pittsburgh, PA, Apr. 2008.

[41] Nikolaus F. Troje. "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns". In: *Journal of Vision* 2.5 (Sept. 2002), pp. 2–2. ISSN: 1534-7362. DOI: 10.1167/2.5.2. eprint: https://arvojournals.org/arvo/content_public/journal/jov/933498/jov-2-5-2.pdf. URL: https://doi.org/10.1167/2.5.2.

[42] Ashish Vaswani et al. "Attention is All you Need". In: *Neural Information Processing Systems*. 2017. URL: https://api.semanticscholar.org/CorpusID:13756489.

[43] Thomas Waltemate et al. "The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response". In: *IEEE Transactions on Visualization and Computer Graphics* PP (Jan. 2018), pp. 1–1. DOI: 10.1109/TVCG.2018.2794629.

[44] Alexander W. Winkler, Jungdam Won, and Yuting Ye. "QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars". In: *SIGGRAPH Asia 2022 Conference Papers* (2022). URL: https://api.semanticscholar.org/CorpusID:252383239.

[45] Jackie (Junrui) Yang et al. "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*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502045. URL: https://doi.org/10.1145/3491102.3502045.

[46] Xinyu Yi, Yuxiao Zhou, and Feng Xu. "TransPose". In: *ACM Transactions on Graphics (TOG)* 40 (2021), pp. 1–13. URL: https://api.semanticscholar.org/CorpusID:234357531.

[47] Xinyu Yi et al. "EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors". In: *ACM Transactions on Graphics (TOG)* 42 (2023), pp. 1–17. URL: https://api.semanticscholar.org/CorpusID:258437292.

[48] Haoran Yun et al. "Animation Fidelity in Self-Avatars: Impact on User Performance and Sense of Agency". In: *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 2023, pp. 286–296. DOI: 10.1109/VR55154.2023.00044.

[49] Ce Zheng et al. "Deep Learning-based Human Pose Estimation: A Survey". In: *ACM Comput. Surv.* 56.1 (Aug. 2023). ISSN: 0360-0300. DOI: 10.1145/3603618. URL: https://doi.org/10.1145/3603618.

[50] Xiaozheng Zheng et al. "Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 14632–14642. URL: https://api.semanticscholar.org/CorpusID:261030328.