AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing

Jiaxi Jiang [1], Paul Streli [1], Huajian Qiu [1], Andreas Fender [1], Larissa Laich [2], Patrick Snape [2], Christian Holz [1]
[1] Department of Computer Science, ETH Zürich, Switzerland [2] Reality Labs at Meta, Switzerland

code available!

ECCV TEL AVIV 2022

abstract

We present AvatarPoser, the first learning-based method that predicts full-body poses in world coordinates using only motion input from the user's head and hands. Our method builds on a Transformer encoder to extract deep features from the input signals and decouples global motion from the learned local joint orientations to guide pose estimation. To obtain accurate full-body motions that resemble motion capture animations, we refine the arm joints' positions using an optimization routine with inverse kinematics to match the original tracking input. AvatarPoser achieved new state-of-the-art results in evaluations on the AMASS dataset, providing a practical interface to support the control and representation of a holistic avatar for Metaverse applications.

prior work

Most general-purpose applications use Inverse Kinematics (IK) to estimate full-body poses. This often generates human motion that appears static and unnatural, especially for those joints that are far away from the known joint locations in the kinematic chain.

Despite the goal of using input from only the head and hands, existing deep learning-based methods implicitly assume knowledge of the pelvis pose. However, pelvis tracking may never be available in most portable MR systems, which increases the difficulty of full-body estimation.

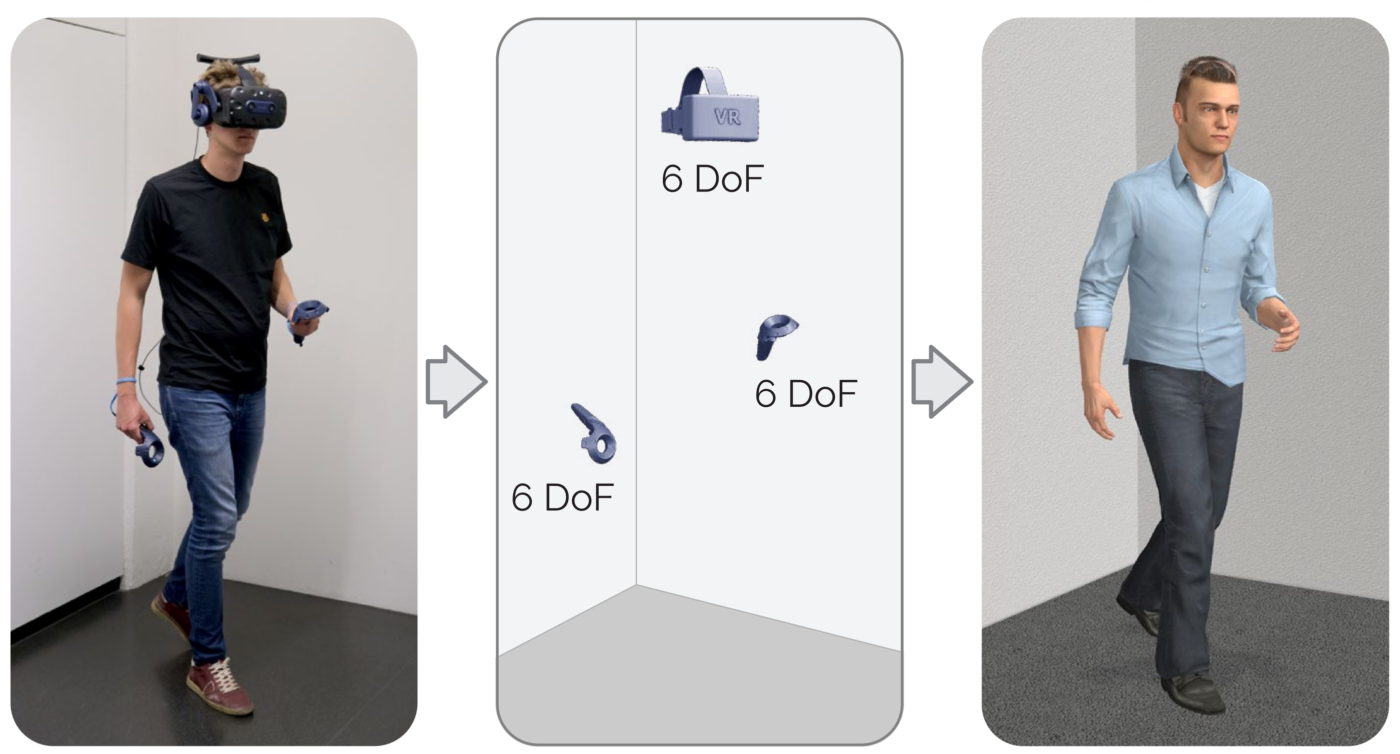
Even with a tracked pelvis joint, animations from estimated lower-body joints sometimes contain jitter and sliding artifacts. These tend to arise from unintended movements of the pelvis tracker, which is attached to the abdomen and thus moves differently from the actual pelvis joint.

problem formulation

We use the SMPL model to represent and animate human body poses. We use the first 22 joints defined in the kinematic tree of the SMPL skeleton.

Input: global positions, rotations, and corresponding velocities of the headset and the two controllers.

Output: local rotations of all joints, global position and pelvis rotation.



method

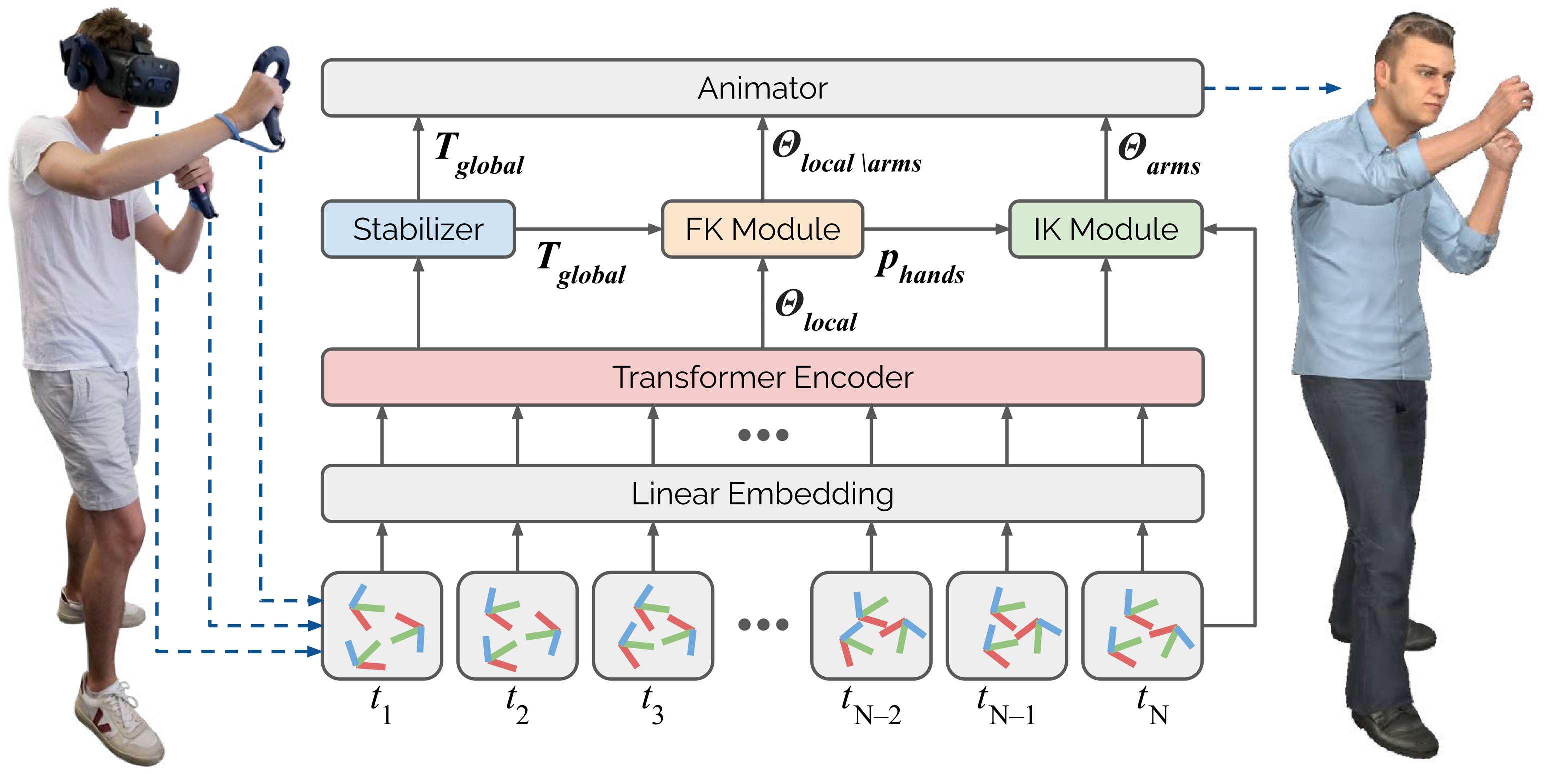
The framework of our proposed AvatarPoser for Mixed Reality avatar full-pose estimation integrates four parts: a Transformer Encoder, a Stabilizer, a Forward-Kinematics Module, and an Inverse-Kinematics Module.

Transformer Encoder: extracts deep pose features from previous time step signals from the headset and hands.

Stabilizer: responsible for global motion navigation by decoupling global orientation from pose features and estimating global translation from the head position through the body's kinematic chain.

FK (Forward Kinematics) Module: calculates joint positions from a human skeleton model and a predicted body pose.

IK (Inverse Kinematics) Module: adjusts the estimated rotation angles of the shoulder and elbow joints to reduce the hand position error.



numerical results

We evaluate state-of-the-art methods for the setup with three inputs (one headset, two controllers), and four inputs (with an additional tracker at the pelvis). AvatarPoser achieves new SOTA performance.

The evaluation metrics are MPJRE (mean per joint rotation error [°]), MPJPE (mean per joint positional error [cm]), and MPJVE (mean per joint velocity error [cm/s]).

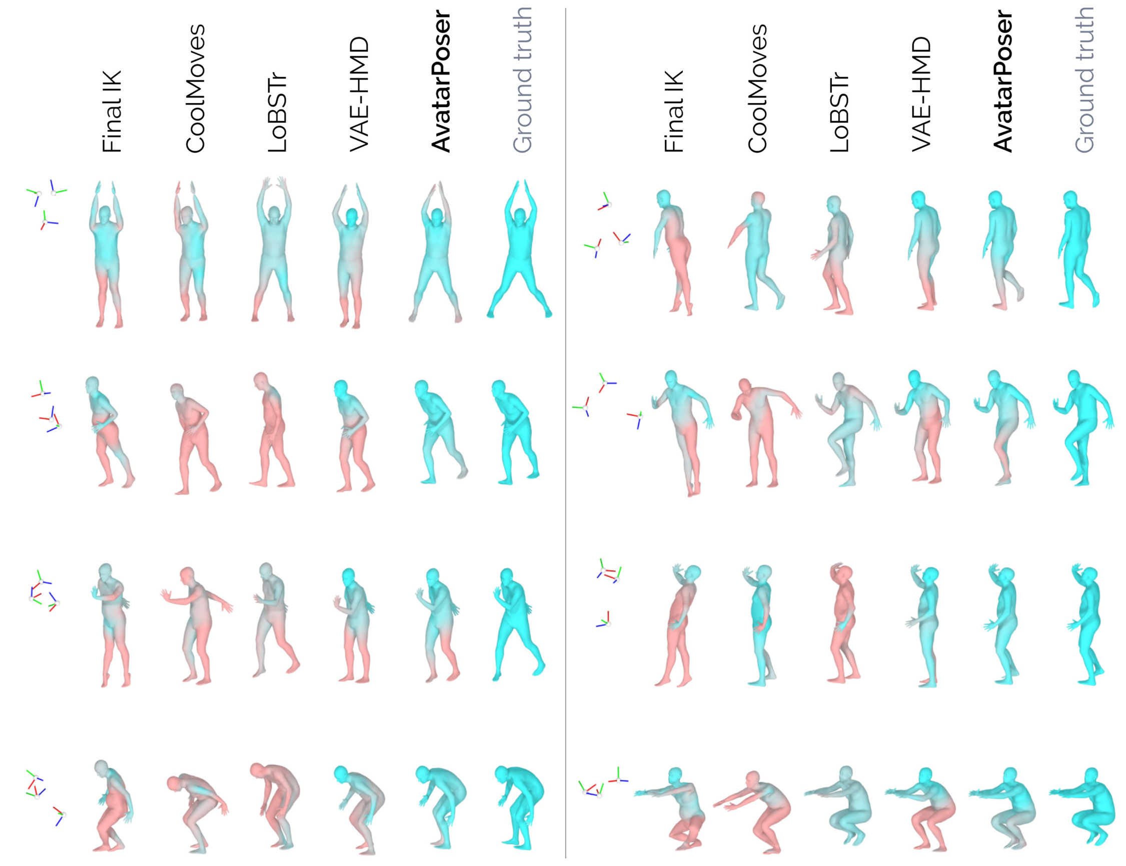
	Four Inputs			Three Inputs		
Methods	MPJRE	MPJPE	MPJVE	MPJRE	MPJPE	MPJVE
Final IK	12.39	9.54	36.73	16.77	18.09	59.24
CoolMoves	4.58	5.55	65.28	5.20	7.83	100.54
LoBSTr	8.09	5.56	30.12	10.69	9.02	44.97
VAE-HMD	3.12	3.51	28.23	4.11	6.83	37.99
AvatarPoser (Ours)	**2.59**	**2.61**	**22.16**	**3.21**	**4.18**	**29.40**

Ablation studies show the contribution of each module.

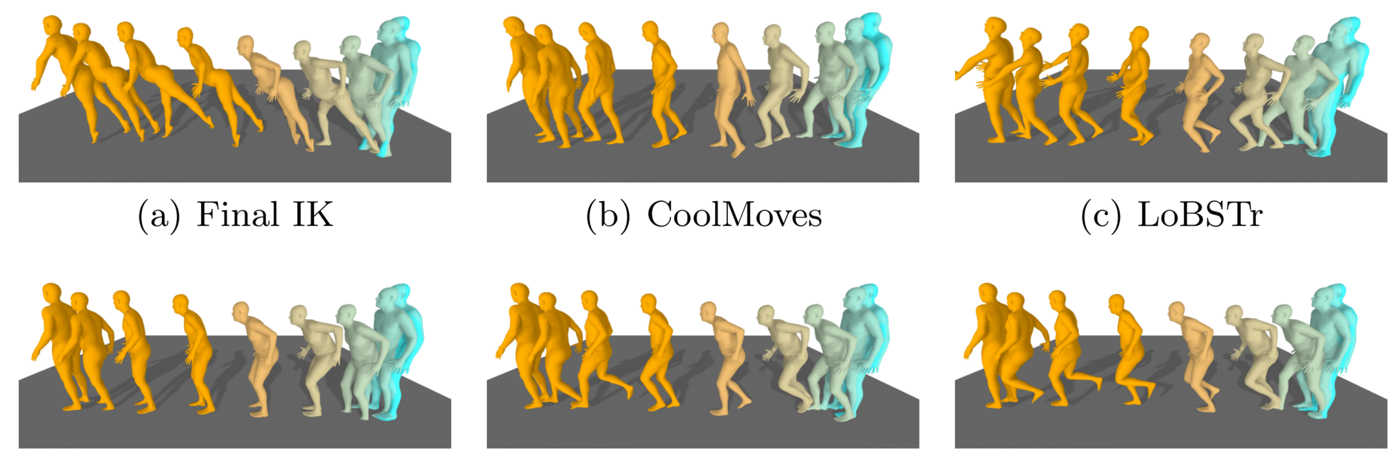
Configurations	MPJRE	MPJPE-Full Body	MPJPE-Hand
Default	6.39	**8.05**	**1.86**
No Stabilizer	6.39	9.29	2.15
Predict Pelvis Position	6.42	8.82	2.11
No FK Module	**6.24**	8.41	2.04
No IK Module	6.41	8.07	3.17

visual results

Visual results for the different methods based on given sparse inputs for various motions. Avatars are color-coded to show errors in red.



Visual results for the different methods for a running motion. Color illustrates different timestamps.



(a) Final IK (b) CoolMoves (c) LoBSTr
(d) VAE-HMD (e) AvatarPoser (Ours) (f) Ground Truth

test on VR device

AvatarPoser works on popular VR systems. We used an HTC VIVE HMD as well as two controllers, each providing real-time input with six degrees of freedom (rotation and translation).