MANIKIN: Biomechanically Accurate Neural Inverse Kinematics for Human Motion Estimation

Jiaxi Jiang, Paul Streli, Xuejing Luo, Christoph Gebhardt, Christian Holz

Department of Computer Science, ETH Zürich, Switzerland
{firstname.lastname}@inf.ethz.ch
https://github.com/eth-siplab/MANIKIN

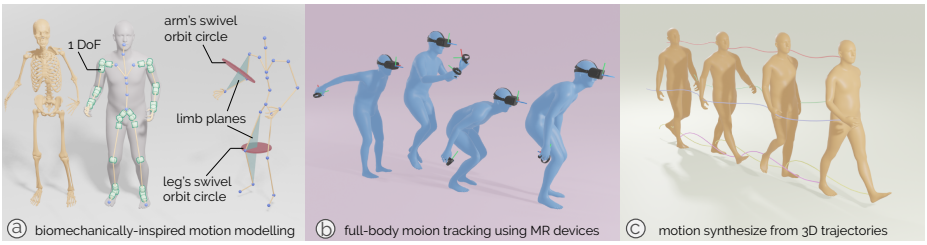


Fig. 1: We propose MANIKIN, a biomechanically accurate neural inverse kinematics solver for human motion estimation. Drawing inspiration from the biomechanical constraints of human movement, we adjust the SMPL models to reflect the appropriate degrees of freedom for human limbs (a). Building on it, we develop a neural-analytic inverse kinematics solver capable of tracking full-body motion using only head and hand poses in Mixed Reality environments (b). Additionally, our system can synthesize human movements based on the 3D trajectories of the head, hands, and feet (c).

Abstract. Mixed Reality systems aim to estimate a user's full-body joint configurations from just the pose of the end effectors, primarily head and hand poses. Existing methods often involve solving inverse kinematics (IK) to obtain the full skeleton from just these sparse observations, usually directly optimizing the joint angle parameters of a human skeleton. Since this accumulates error through the kinematic tree, predicted end effector poses fail to align with the provided input pose. This leads to discrepancies between the predicted and the actual hand positions or feet that penetrate the ground. In this paper, we first refine the commonly used SMPL parametric model by embedding anatomical constraints that reduce the degrees of freedom for specific parameters to more closely mirror human biomechanics. This ensures that our model produces physically plausible pose predictions. We then propose a bio**m**echanically **a**ccurate **n**eural **i**nverse **kin**ematics solver (MANIKIN) for full-body motion tracking. MANIKIN is based on swivel angle prediction and perfectly matches input poses while avoiding ground penetration. We evaluate MANIKIN in extensive experiments on motion capture datasets and demonstrate that our method surpasses the state of the art in quantitative and qualitative results at fast inference speed.

Keywords: Inverse Kinematics · Human Body Models · Motion Tracking · Character Animation · Mixed Reality

1 Introduction

In today's Mixed Reality environments, user interaction primarily relies on the user's head pose and hand for input. The problem of full human body pose estimation has since emerged as a challenging task using such sparse tracking cues as input. Recent work [3,4,11,13,21,64,67,74] has leveraged data-driven techniques to address this challenge by learning the mapping from sparse observations to full-body poses, trained on large-scale motion capture datasets.

Typically, models for representing full-body poses are parameterized with a kinematic chain and encode a pose as a vector of joint angles. While this model-based approach maintains realistic human body proportions with consistent bone lengths, it also accumulates positional errors throughout the kinematic tree. These errors, however, lead to discrepancies between predicted results and observed positions or poses.

To reconcile the mismatch between observed joint poses and those represented through human models, prior work has proposed several approaches. Methods have combined traditional IK solvers for the upper body, where hand poses are available as input, with neural networks to estimate the lower body [67]. Others integrate iterative optimization techniques that further process the values estimated by neural networks [9,21]. Finally, researchers have proposed specialized loss functions to enforce alignment during network training [74].

However, several limitations remain in existing methods that prevent accurate tracking from sparse input cues alone. First, commercial IK solvers (e.g., Final IK [2]) primarily focus on aligning predictions with observations without considering the naturalness and smoothness of motion, as shown and solved in previous work [3, 21]. Second, optimization-based methods [9, 21] are sensitive to initial values, prone to getting stuck in local minima, and they are time-consuming to execute, which is challenging in real-time applications. Third, purely learning-based methods [13,74] struggle to generalize well to unseen data.

Besides method limitations, previous approaches rely on a parametric human motion model, often SMPL [40], which assumes knowledge of full 3-DoF rotations during processing [39, 42]. Since multiple joint orientation configurations can result in the same final joint positions, however, unconstrained optimization may yield physically impossible joint angles. Optimization may also produce end-effector positions that diverge from those captured as input, resulting in a mismatch of hand positions that complicates precise interaction or feet locations that penetrate the ground.

In this paper, we introduce MANIKIN, a novel biomechanically accurate neural inverse kinematics solver that is designed to predict full-body poses from sparse tracking of the end-effectors (e.g., hand) while ensuring precise alignment of its position with the known observational data. Our neural solver is fully differentiable for efficient training, while providing fast inference for real-time applications. The key novelty of our solver is its prediction of limb *swivel*, which, in the case of the arm, is the angle between the vertical plane through the shoulder and wrist and the plane containing the entire arm [23, 29, 60]. Besides, our method refines SMPL-based models [40] with a lower degree-of-freedom represen-

tation that better matches human biomechanical constraints and, thus, ensures physically plausible predictions.

Our method surpasses the state-of-the-art tracking accuracy in two steps. Our neural network first forecasts the position of the base joint shoulder and the swivel angle. We then determine the joint configurations based on the swivel angle and the provided positions of the wrists through analytic geometry, guided by human motion constraints. Our approach reduces hand position errors from more than 5 cm [74] to a mere 0.1 cm on real VR data, which enables precise hand control in AR/VR environments. We validate our approach in more detail in extensive experiments on existing motion capture datasets. Our method consistently outperforms existing techniques, yielding more accurate and smoother motion results.

To summarize, the main contributions of this paper are:

1. MANIKIN, a novel IK solver for full-body pose estimation that builds on a biomechanically accurate body representation that matches human motion constraints. MANIKIN is fully differentiable and affords end-to-end training.
2. A neural-analytic solution for the joint angle configuration of human limbs based on swivel angle prediction. This enables precise end-effector control, accurate full-body pose tracking, and ensures physically plausible predictions.
3. New state-of-the-art performance for full-body pose estimation, surpassing previous optimization-based as well as pure learning-based methods while achieving considerably fast inference time. Our approach also generalizes well other IK tasks, thereby providing a new paradigm for IK problems and showing promising potential for real applications.

2 Related Work

Inverse Kinematics for Human Motions. Inverse kinematics (IK) aims to solve for the joint parameters of a kinematic chain given a desired end-effector pose. It has been extensively studied and applied in fields such as robotics [14, 43, 47, 53, 63] and computer animation [2, 6, 16, 46, 57, 73]. Previous research [3, 21] has highlighted the limitation as unrealistic animations of traditional full-body IK solutions (e.g., *Final IK* [2]). Recent works have combined IK with deep learning to improve the robustness and flexibility of the predictions [25,33–35,55,67,72]. Specifically, some methods employ optimization-based techniques to refine predictions obtained from neural networks [5, 8, 9, 21, 48]. However, iterative optimization is slow by design and often suffers from local minima. To address these challenges, HybrIK [34] integrates neural networks with analytic IK through twist-and-swing decomposition. However, HybrIK's effectiveness depends on precise full-body keypoint estimation from images, limiting its applicability when only end-effector poses are available. Motivated by this limitation, we have developed a framework that combines neural networks with analytic IK for pose estimation using only end-effector poses. We demonstrate that traditional analytic IK methods [23, 60] can still play a significant

role when effectively integrated with neural networks.

Model-based Human Pose Estimation. 3D human pose estimation is a well-established problem that has been addressed using various sensing modalities [7,8,26,56,62,68]. Many methods aim to locate the body's joints by directly regressing on their positions within the human skeleton [36,49,58,59,71,75]. To ensure bone consistency, SMPL-based pose estimation algorithms [8,24,40] have been proposed that estimate the parameters of a human body model. More recent works use Transformers to operate on sequential data [37,38,41,65,66]. Methods for estimating full-body pose using IMU sensors [19,62,68,69], or tracked headsets and hand controllers [4,13,20,21] have also employed the SMPL body representation. However, although SMPL offers flexibility in possible poses, allowing 3-DoF rotations for all joints is not biomechanically accurate [27] and leads to an overparameterized model. While some previous methods [8,48] train body priors to constrain body poses to be realistically looking, the internal joint configurations of these poses are not guaranteed to be biomechanically plausible.

Motion Capture from Sparse Observations. Motion capture using sparse wearable devices has gained significant attention due to its portability and cost-effectiveness. Existing methods can generally be categorized into IMU-based methods [7,19,22,44,62,68–70] and MR device-based methods [4,11,13,20,21,30,32,50,64,67]. In the context of virtual and augmented reality, the research advancements aimed at capturing full-body movements from tracked end-effector poses are also particularly significant, as inside-out tracking of users' heads and hands [17,18], or hand-held controllers, is now common in mixed reality devices. Our previous work AvatarPoser [21] was the first 3-point-tracking method that generalizes over different motion types. It uses a Transformer model to estimate the full-body pose and integrates it with IK optimization to reduce the hand position error. Following this, AGRoL [13] proposed a diffusion model for smoother predictions, and then EgoPoser [20] focuses on efficient egocentric pose and shape estimation in the wild with realistic field-of-view modelling. AvatarJLM [74] improved upon these results by modelling joint-level correlations and introducing a hand alignment loss to replace the IK optimization. However, these methods can not yet guarantee solutions that are globally optimal and anatomically plausible.

3 Proposed Method

3.1 Task Description

The aim of our paper is to accurately track the full-body pose based on the poses of the end effectors. Despite our proposed solution's applicability to general IK problems, we focus on a common scenario in Mixed Reality due to page constraints. Specifically, we address situations where the global 6 DoF poses of the head and hands are available, as formulated in our previous work AvatarPoser [21]. Each 6DoF pose encompasses 3D positions and orientations. Using these poses as input, our objective is to reconstruct the spatial positions of the

articulated joints comprising the user's entire body within the global space. This is a challenging and ill-posed IK task, as multiple full-body pose configurations can lead to the same hand poses.

3.2 Human Motion Modeling

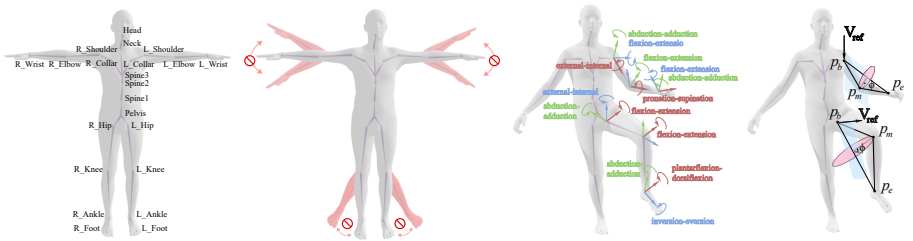


Fig. 2: Comparison between SMPL and biomechanically constraint limb motion: (a) the skeleton of the SMPL model, (b) unrealistic SMPL joint configurations in the AMASS dataset, (c) biomechanically plausible joint configurations, (d) swivel angle parametrization of arm and leg.

Biomechanically Plausible Motion Constraints. To describe human movement, previous methods [13,21,74] use the first 21 joints of the SMPL model [40] for full body human motion modeling (Fig. 2a). The finger joints are omitted, and the hand pose is denoted by the wrist pose. SMPL employs a standard skeletal structure to define a human body, and the pose of each joint is characterized by its relative rotation concerning its parent joint within the kinematic tree. The adopted SMPL rig encompasses 22 joints, resulting in the definition of a pose through $3 \times 22 = 66$ parameters represented as axis angles, supplemented by three parameters for root translation, summing up to 69 parameters in total.

While SMPL can adequately represent the skin of the human body and keep the bone length consistent, it is important to acknowledge that the 3 DoF assumption of the SMPL model does not fully account for the inherent constraints governing realistic human motion. For instance, once the pose of the shoulder of a human is fixed, the wrist position on the same arm can only be controlled by a constrained flexion-extension movement of the elbow joint. In contrast, the 3 DoF rotation of the SMPL model can make the wrist position move on a sphere, which is impossible for humans. This is even evident in the AMASS dataset [42], which contains instances of unrealistic SMPL joint configurations (Fig. 2b).

In key literature of biomechanics [10, 31, 51, 52, 54], the prevalent model for describing the human arms and legs involves three rigid segments connected by frictionless joints, constituting a total of seven degrees of freedom (Fig. 2c). Informed by this, we correct the degrees of freedom in the SMPL model's joints to

more accurately mirror the natural motion constraints of human arms and legs while maintaining three degrees of rotational freedom for other joints, including those in the spine and hip.

Swivel Angle. Informed by prior work [23, 29, 60], we base motion modelling of arms and legs on the *swivel angle*, which represents the extent to which the respective mid joint (elbow or knee) is rotated around the end-base axis (wrist-shoulder or ankle-hip). When the position of the base joint is fixed, the 7-DoF limb has one more DoF than the 6 constraints of the controlled end effector, which can lead to infinite possible joint angle configurations to reach the end effector pose. The swivel angle can parameterize the extra degree of freedom to give a unique solution. This becomes clear when noting that as the swivel angle ϕ of the arm changes, the elbow traces a circular arc situated on a plane whose normal is parallel to the wrist-to-shoulder axis (Fig. 2d, equally applies to legs).

3.3 Biomechanially Accurate Neural Inverse Kinematics

By utilizing realistic motion constraints, we propose a novel solution for full-body pose estimation from sparse tracking points at the end effectors. Fig. 3 shows an overview of our framework. Given the 6D poses of the head and hands as input, the *neural network* predicts the body's global orientation, local poses of the joints on the torso, the foot pose, and the swivel angle of arms and legs. Utilizing the torso angles, forward kinematics is applied to attain shoulder and hip positions (*Torso FK*). Our biomechanical constraints allow us then to analytically compute limb angles via inverse kinematics on the respective swivel angle and base joint position (*Analytic Arm/Leg Solver*). Finally, we render the full *body model*. We will introduce the details of each part in the following.

Input Representation. The inputs of our network are positions p and orientations Φ of the head and hands, as well as the corresponding linear and angular velocities to obtain a signal of temporal smoothness. For each input joint j, the linear velocity $\mathbf{v}$ is given by backward finite difference at each time stamp $\mathbf{v}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$. Similarly, the angular velocity ω is given by $\omega_t = \Phi_{t-1}^{-1}\Phi_t$. We convert orientation and angular velocity into their 6D representation $\Phi^{1\times6}$ and $\omega^{1\times6}$. As a result, the final input representation is a concatenated vector of position, linear velocity, rotation, and angular velocity from all given sparse inputs, which we write as $\mathbf{x_j} = \left[\mathbf{p}_j^{1\times3}, \mathbf{v}_j^{1\times3}, \Phi_j^{1\times6}, \omega_j^{1\times6}\right]$. To utilize the temporal information, we use a sliding window of N frames as the input. The input window is then fed into a neural network to predict the full-body pose at a current frame.

Neural Network. Given the 6D poses of the head and hands, we use the neural network to predict the poses of joints on the torso, the 6 DoF foot poses, and the swivel angle of arms and legs. Our framework is plug-and-play and supports any network backbone. To compare to state-of-the-art methods on full-body pose estimation from head and hand pose, we employ two kinds of backbone models:

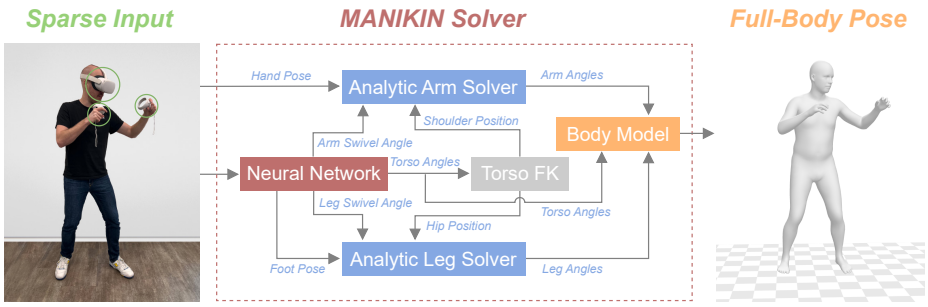


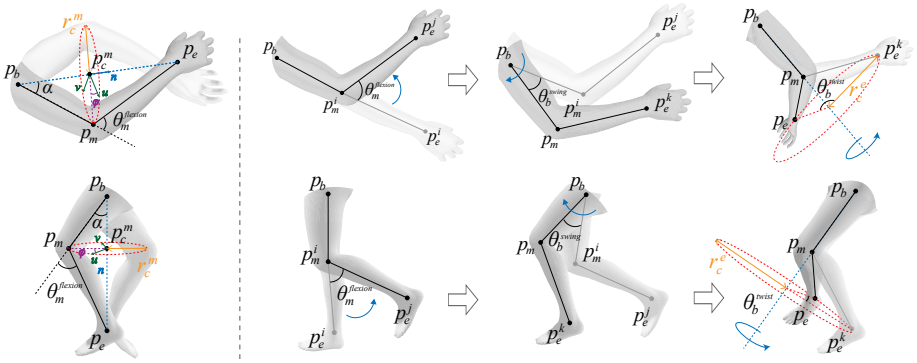
Fig. 3: Overview of MANIKIN for full-body pose estimation from sparse tracking signals. Given the 6 DoF of the head and two hands, the neural network predicts the foot pose, swivel angles for arms and legs, as well as joint angles in the torso. Next, the shoulder and hip position can be obtained through forward kinematics on the torso. Given the poses of an end effector (hand and foot), predicted swivel angles, and base joint (shoulder and hip) positions, we use analytic geometry to get all limb joint configurations. Combined with the predicted torso joints, we obtain a full-body pose.

(1) a three-layer lightweight Transformer [21], denoted as MANIKIN-S and (2) a larger temporal-spatial Transformer network [74], denoted as MANIKIN-L. For a fair comparison with previous methods, we maintain the identical architecture as the original model, with the exception of the output layer where we predict swivel angles ϕ for arms and legs, in addition to foot poses.

Torso FK. After the network predicts the torso joint angles, we employ forward kinematics on these angles to calculate the joint positions. The benefits of using Torso FK are two-fold. First, the joint positions are directly optimized to reduce the joint position error. Second, the obtained positions of base joints (*i.e.* shoulder and hip) are further used for swivel angle prediction and for analytic IK solution. Compared to previous methods that apply forward kinematics on the entire body, our method is more efficient and accurate, because the network can focus on the torso part. The loss function for the torso FK is the L1 loss between the predicted joint positions and the ground truth joint positions:

$$\mathcal{L}_{\text{FK}}^{\text{torso}} = \left\| FK(\theta_{\text{torso}}^{\text{pred}}) - FK(\theta_{\text{torso}}^{\text{gt}}) \right\|_1 \tag{1}$$

Swivel Angle Estimation. Existing datasets do not provide the swivel angle, so we calculate the ground truth swivel angle from the ground truth joint positions. By varying the swivel angle ϕ, different positions of the elbow joint are obtained along an orbiting circle of the elbow. Let $\mathbf{p}_b$, $\mathbf{p}_m$, and $\mathbf{p}_e$ be the position of the base joint, mid joint, and end joint of a human limb. l_{bm} represents the distance between the base joint and the mid joint, corresponding to the length of the upper arm in an arm or the length of the thigh in a leg. l_{me} denotes the distance between the mid joint and the end joint, equating to the length of the forearm for an arm or the length of the lower leg for a leg. The relationship be-



(a) swivel angel (b) mid flexion (c) base swing (d) base twist

Fig. 4: Illustrations of the triangular geometry of the human limbs. (a) shows the relationship between swivel angel and mid joint position. (b) to (d) show the procedure to rotate the limb from T-pose to desired positions.

tween the swivel angle, elbow joint, and the orbit circle is illustrated in Fig. 4a. We define the reference vector of the corresponding limb $\mathbf{v}_{\text{ref}}$ to align with the downward direction of the forward direction of the body's longitudinal axis for the arm or sagittal axis for the leg (Fig. 2d). Furthermore, u is defined as the projection of $\mathbf{v}_{\text{ref}}$ onto the orbit circle plane, and $\mathbf{v}$ is orthogonal to $\mathbf{u}$ and the normalized base-end vector $\mathbf{n}$:

$$\mathbf{n} = \frac{\mathbf{p}_e - \mathbf{p}_b}{\|\mathbf{p}_e - \mathbf{p}_b\|}, \quad \mathbf{u} = \frac{-\mathbf{v}_{\text{ref}} + (\mathbf{v}_{\text{ref}} \cdot \mathbf{n})\mathbf{n}}{\| - \mathbf{v}_{\text{ref}} + (\mathbf{v}_{\text{ref}} \cdot \mathbf{n})\mathbf{n}\|}, \quad \mathbf{v} = \mathbf{u} \times \mathbf{n} \tag{2}$$

In the triangle formed by the base joint, the mid joint, and the end joint, the angle α between the base-mid and base-end vectors can be calculated by

$$\alpha = \arccos\left(\frac{{l_{bm}}^2 + \|\mathbf{p}_e - \mathbf{p}_b\|^2 - {l_{me}}^2}{2l_{bm}\|\mathbf{p}_e - \mathbf{p}_b\|}\right) \tag{3}$$

Then the center of the mid-joint orbit circle $\mathbf{p}_c$ can be derived by:

$$\mathbf{p}_c^m = \mathbf{p}_b + l_{bm}\mathbf{n}\cos\alpha \tag{4}$$

The swivel angle ϕ can be calculated by

$$\cos\phi = \frac{\mathbf{u} \cdot (\mathbf{p}_m - \mathbf{p}_c^m)}{\|\mathbf{p}_m - \mathbf{p}_c^m\|}, \quad \sin\phi = \frac{\mathbf{v} \cdot (\mathbf{p}_m - \mathbf{p}_c^m)}{\|\mathbf{p}_m - \mathbf{p}_c^m\|} \tag{5}$$

We normalized the predicted $\cos\phi$ and $\sin\phi$ by their root sum squared so that $\cos^2\phi + \sin^2\phi = 1$. Since single angle representation ϕ is discontinuous (e.g., 0 and 2π are equivalent), it is hard for networks to fit [76]. Therefore, we optimize the swivel angle in its continuous representation $[\cos(\phi), \sin(\phi)]^T$:

$$\mathcal{L}_{\text{swivel}} = \left\|\sin(\phi) - \sin(\phi^{gt})\right\|_1 + \left\|\cos(\phi) - \cos(\phi^{gt})\right\|_1 \tag{6}$$

The mid-joint position is a function of the swivel angle. Once the swivel angle is predicted, we can also get the mid joint position by:

$$\mathbf{p}_{\text{m}}^{\text{pred}}(\phi) = \mathbf{p}_c^m + r_c^m \left(\mathbf{u}\cos\phi + \mathbf{v}\sin\phi\right) \tag{7}$$

where $\mathbf{p}_c^m$ can be calculated by Eq. 4 and $r_c^m = l_{bm}\sin\alpha$ is the radius of the mid joint orbit circle.

Since our goal is to find the best mid-joint position that depends on the swivel angle, we also optimize the middle joint position in the loss function.

$$\mathcal{L}_{\text{m}} = \left\|\mathbf{p}_{\text{m}}^{\text{pred}}(\phi) - \mathbf{p}_{\text{m}}^{\text{gt}}\right\|_1 \tag{8}$$

Analytic Limb Solver. Given the predicted position of the base joint and the swivel angle, there exists a closed-form solution when the 6 DoF pose of end effectors is provided. We will solve for the joint angles in the following.

1) Mid joint flexion rotation. The desired positions of the base, mid, and end joints define the triangle formed by these points. The flexion angle rotates around the vertical axis in the T pose and brings the end joint from the position from $\mathbf{p}_e^i$ to $\mathbf{p}_e^j$ (Fig. 4b). The flexion rotation angle of the mid-joint $\theta_{\text{m}}^{\text{flexion}}$ is calculated by the law of cosines:

$$\theta_{\text{m}}^{\text{flexion}} = \pi - \arccos\left(\frac{{l_{bm}}^2 + {l_{me}}^2 - \|\mathbf{p}_e - \mathbf{p}_b\|^2}{2 l_{bm} l_{me}}\right) \tag{9}$$

2) Base joint swing and twist rotation. The base joint is always a ball-and-socket joint with three DoFs and can be parameterized with the swing-and-twist decomposition [15]. The swing rotation is determined by the rotation needed to bring the mid joint from the position $\mathbf{p}_m^i$ in T-Pose to the desired position $\mathbf{p}_m$ (Fig. 4c), which can be given by:

$$\theta_{\text{b}}^{\text{swing}} = \frac{(\mathbf{p}_m^i - \mathbf{p}_b) \times (\mathbf{p}_m - \mathbf{p}_b)}{\|(\mathbf{p}_m^i - \mathbf{p}_b) \times (\mathbf{p}_m - \mathbf{p}_b)\|} \arccos\left(\frac{(\mathbf{p}_m^i - \mathbf{p}_b) \cdot (\mathbf{p}_m - \mathbf{p}_b)}{l_{bm}^2}\right) \tag{10}$$

Once the base swing rotation is applied to the base joint in T-pose, the end-joint position moves from $\mathbf{p}_e^j$ to $\mathbf{p}_e^k$. Next, the base twist will make the limb rotate in order for the end joint to move from $\mathbf{p}_e^k$ to reach the desired position $\mathbf{p}_e$ (Fig. 4d). The rotation axis is the vector $\overrightarrow{\mathbf{p}_b\mathbf{p}_m}$, and the twist rotation of the base joint can form an orbiting circle of the end joint. The radius of this orbit circle r_c^e can be calculated by:

$$r_c^e = l_{me}\sin(\theta_{\text{m}}^{\text{flexion}}) \tag{11}$$

Then the twist rotation of the base joint can be solved by:

$$\theta_{\text{b}}^{\text{twist}} = 2\arcsin\left(\frac{\|\mathbf{p}_e - \mathbf{p}_e^k\|}{2r_c^e}\right) = 2\arcsin\left(\frac{\|\mathbf{p}_e - \mathbf{p}_e^k\|}{2 l_{me}\sin(\theta_{\text{m}}^{\text{flexion}})}\right) \tag{12}$$

Since the foot pose is directly predicted by the neural network, we can now obtain the poses for all joints of the leg.

3) Elbow twist. The mid joint on the arm has one more DoF than the mid joint on the leg, which is the elbow twist that controls the pronation-supination of the forearm and hand. Although the elbow twist and the wrist swing rotations are applied to different joints, we observed that their combined effect is the same as a single 3-DOF rotation $\mathbf{q}^{\text{global}}_{\text{wrist}}$. The elbow twist quaternion is given by [12]:

$$\mathbf{q}^{\text{twist}}_{\text{elbow}} = \left(q_w \|\mathbf{v}^{\text{twist}}_{\text{elbow}}\|^2, v_x(\mathbf{v}^{\text{twist}}_{\text{elbow}} \cdot \mathbf{q}), v_y(\mathbf{v}^{\text{twist}}_{\text{elbow}} \cdot \mathbf{q}), v_z(\mathbf{v}^{\text{twist}}_{\text{elbow}} \cdot \mathbf{q})\right) \tag{13}$$

where $\mathbf{q} = (q_x, q_y, q_z)$ is the vector part of quaternion $\mathbf{q}^{\text{global}}_{\text{wrist}} = (q_w, q_x, q_y, q_z)$, and $\mathbf{v}^{\text{twist}}_{\text{elbow}} = (v_x, v_y, v_z) = \overrightarrow{\mathbf{p_e p_m}}$ is the twist axis vector.

4) Wrist joint swing. The swing quaternion of the wrist is calculated by multiplying the original global wrist quaternion with the conjugate of the twist quaternion:

$$\mathbf{q}^{\text{swing}}_{\text{wrist}} = \mathbf{q}^{\text{global}}_{\text{wrist}} \mathbf{q}^{\text{twist}*}_{\text{elbow}} \tag{14}$$

With these equations in place, we can now determine the poses of all body joints.

End-to-End Training. Our method is fully differentiable and supports end-to-end training. The final loss function is composed of a local rotational loss $\mathcal{L}_{\text{local}}$ to optimize the local body pose, a global orientation loss $\mathcal{L}_{\text{global}}$ to optimize the global root orientation, a foot pose loss $\mathcal{L}_{\text{Foot}}$ to optimize the foot position and orientation, a torso FK loss to optimize the torso joint positions $\mathcal{L}^{\text{torso}}_{\text{FK}}$, a swivel angle loss $\mathcal{L}_{\text{swivel}}$ together with a mid joint loss $\mathcal{L}_{\text{m}}$ to optimize the swivel angles.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}} + \lambda^{\text{torso}}_{\text{FK}}\mathcal{L}^{\text{torso}}_{\text{FK}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} \tag{15}$$

We set the weight of λ_{ori} to 0.05, λ_{swivel} and λ_{m} to 0.2, and all other weights to 1. To optimize the parameters of the network, we adopt the Adam solver [28] with batch size 256. We set the input window size as 40 frames. We train our model with PyTorch on NVIDIA GeForce GTX 4090 GPUs.

Joint Angles in SMPL-based dataset. Our method can also refine the SMPL-based dataset (*e.g.,* AMASS) to ensure biologically plausible arm- and leg configurations. To do so, we can just apply the same calculations (Equations 9-14) to analytically derive joint angles given the known base and mid joint positions of the dataset. This method accurately replicates SMPL model joint positions and improves joint angle representations based on real human motion constraints. This provides valuable resources for applications like healthcare.

4 Experiments

In this section, we will evaluate the effectiveness of MANIKIN to show its benefit over previous SMPL-based methods for full body tracking based on end-effectors.

4.1 Datasets and Evaluation Metrics

Following prior work [13, 21, 74], we use the subsets CMU [1], BMLrub [61] and MPI [45] of AMASS [42] for training and testing. For sim-to-real evaluation, we additionally use a real-world VR motion capture dataset [74].

The evaluation metrics employed in our paper are based on prior work [13, 21, 74]. Specifically, we use *MPJPE* (mean per-joint position error, in $[cm]$) to validate the accuracy of pose estimation, *MPJVE* (mean per-joint velocity error, in $[cm/s]$) to validate smoothness of motion and accuracy, *H-PE* to analyze hand position error, *U-PE* to analyze upper-body position error, *L-PE* to analyze lower-body position error, *Jitter* (measured as a jerk, in $[10^2 m/s^3]$) to validate smoothness of motion, *GP* to analyze ground penetration error.

4.2 Comparisons to State-of-the-Art Methods

We compare MANIKIN with state-of-the-art methods in MR-based full-body tracking: AvatarPoser [21], AGRoL [13], and AvatarJLM [74]. To ensure a fair comparison, we categorize the methods into two groups based on whether the model is *lightweight* or *heavyweight*. AGRoL originally was designed to predict the current frame using future frames during offline evaluation. To align it with the setting of the other compared methods, we adapted AGRoL (denoted with *) to predict the current frame based on past frames. AvatarPoser* denotes AvatarPoser without optimization-based IK, enabling real-time inference.

Evaluation Protocols. MANIKIN-S and MANIKIN-L follow the online evaluation protocol used in AvatarPoser and AvatarJLM. They utilize the past N=40 frames as input and only output the last frame as the prediction at timestamp t, with a step size of 1 for the sliding window. MANIKIN-LN is the seq2seq-variant of MANIKIN-L that inputs and outputs N=40 frames with the window shifting by a step size of N in evaluation. We introduced this variant for a fair comparison with the seq2seq model AGRoL (input & output N=196 frames with step size N), which leveraged future signals from its input window for predictions.

Results on AMASS Dataset. Following prior work [13, 21, 74], we use the subsets CMU [1], BMLrub [61], and MPI [45] of AMASS [42], as well as the same data splits, for training and testing. Tab. 1 shows the experimental results. Our default lightweight model significantly reduces position error compared to the very recent AvatarJLM [74], while achieving nearly 20× faster processing speed. Our model also has significantly fewer parameters.

Results of Cross-Dataset Evaluation. We follow [21, 74] to perform a 3-fold cross-dataset evaluation on the subsets CMU [1], BMLrub [61] and MPI [45]. Thus, we train on two subsets and test on the remaining one. Tab. 2 shows the results. Our method outperforms all other methods in all metrics. We provide visual comparisons in first-person view in Fig. 5 and third-person view in Fig. **??**.

Table 1: Results on AMASS dataset. The best result is highlighted in **boldface**.

Methods	MPJPE	H-PE	GP	U-PE	L-PE	MPJVE	Jitter	FPS	Param.	FLOPs
AvatarPoser* [21]	4.20	1.98	2.36	2.01	7.85	29.32	16.64	612	4.12M	0.33G
AGRoL* [13]	3.86	1.42	2.24	1.60	7.13	50.78	48.30	17.8	7.48M	1.00G
MANIKIN-S (Ours)	**3.36**	**0.02**	**0.75**	**1.32**	**6.72**	**23.18**	**14.08**	580	4.12M	0.33G
AvatarPoser [21]	4.18	1.15	2.36	1.76	7.85	29.40	16.80	12.4	4.12M	0.33G
AGRoL [13]	3.71	1.31	2.21	1.55	6.84	18.59	7.26	0.24	7.48M	1.00G
AvatarJLM [74]	3.35	1.24	0.81	1.53	6.54	20.79	10.08	31.1	63.8M	4.64G
MANIKIN-L (Ours)	3.19	0.01	0.69	1.43	6.27	20.10	9.97	30.5	63.8M	4.64G
MANIKIN-LN (Ours)	**2.73**	**0.01**	**0.52**	**1.30**	**5.13**	**13.55**	**7.95**	0.81	63.8M	4.64G

Table 2: Results of cross-dataset evaluation. The best results are in **boldface**.

Methods	CMU				BML				MPI			
	MPJPE	H-PE	GP	MPJVE	MPJPE	H-PE	GP	MPJVE	MPJPE	H-PE	GP	MPJVE
AvatarPoser* [21]	8.40	4.72	2.06	35.88	7.08	3.37	2.85	43.85	8.07	5.61	2.64	30.92
AGRoL* [13]	8.81	3.41	2.98	85.85	7.34	2.01	2.21	56.21	8.18	2.46	3.00	103.01
MANIKIN-S (Ours)	**7.21**	**0.04**	**0.72**	**29.06**	**6.31**	**0.03**	**1.99**	**40.29**	**6.41**	**0.04**	**0.70**	**27.03**
AvatarPoser [21]	8.37	2.08	2.06	35.76	7.04	1.54	2.85	43.70	8.05	2.56	2.64	30.85
AGRoL [13]	8.87	3.19	2.20	28.02	7.29	1.97	2.97	34.29	7.91	2.34	2.86	26.09
AvatarJLM [74]	7.75	1.37	1.11	26.54	6.49	0.85	2.24	36.96	6.60	1.17	1.65	23.57
MANIKIN-L (Ours)	6.92	0.04	0.62	25.38	6.02	0.02	1.97	33.74	6.19	0.03	0.66	23.28
MANIKIN-LN (Ours)	**6.90**	**0.04**	**0.60**	**22.49**	**5.96**	**0.01**	**1.68**	**30.33**	**6.08**	**0.01**	**0.65**	**21.74**

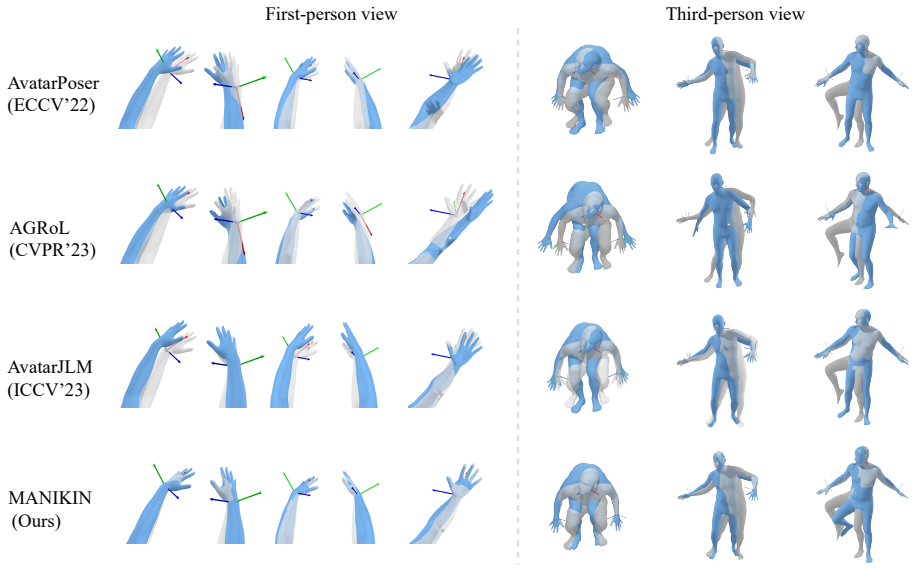


Fig. 5: Visual comparisons of different methods under the first and third person views. The ground truth pose is colored in transparent gray. Our method can perfectly match the hand observation and has better full-body prediction results than other methods.

Results on Real-Captured Data. To further evaluate our performance on the real-world headset-and-controller data of VR/AR applications, we test our

Table 3: Results on real-captured data. The best result is highlighted in **boldface**.

Methods	MPJPE	H-PE	GP	U-PE	L-PE	MPJVE	Jitter
AvatarPoser	11.22	6.60	2.20	5.79	20.73	31.67	12.87
AvatarJLM	9.72	5.27	0.17	5.32	17.43	27.59	13.10
MANIKIN-L	**8.85**	**0.12**	**0.01**	**4.43**	**16.57**	**26.12**	**12.25**

Table 4: Ablation studies for the effect of proposed IK solutions and swivel angle estimation. The evaluation is made on the MPI dataset.

Methods	MPJPE	H-PE	GP	U-PE	L-PE	MPJVE
Ours	**6.41**	**0.04**	**0.66**	**3.02**	**12.09**	**27.03**
No Upper IK	7.20	5.61	0.66	4.41	12.09	31.16
No Lower IK	7.06	0.04	2.64	3.02	14.13	27.05
Opt-based IK	7.65	2.56	1.89	4.18	13.75	31.02
Learning-based IK	7.96	4.35	2.08	4.41	14.13	31.27
No $\mathcal{L}_{mid}$	6.59	0.05	0.72	3.25	12.44	27.35
No $\mathcal{L}_{swivel}$	6.62	0.04	0.66	3.27	12.48	27.15
Swivel - Single	7.14	0.06	0.75	4.05	12.54	29.03
Swivel - Search	8.82	0.05	0.74	5.51	14.61	38.02

Table 5: Results estimating full body pose given the positions of the head, wrists, and ankles. The best result is highlighted in **boldface**.

Methods	MPJPE	MPJVE	U-PE	L-PE	H-PE	F-PE
Transformer-SMPL	5.34	22.74	4.37	6.73	5.65	6.48
Transformer-Ours	**3.74**	**14.98**	**3.31**	**4.36**	**0.11**	**0.12**
RNN-SMPL	7.93	43.84	5.64	11.24	7.94	12.86
RNN-Ours	**4.92**	**24.97**	**4.10**	**6.10**	**0.20**	**0.76**
MLP-SMPL	8.68	32.26	6.52	11.81	9.96	13.17
MLP-Ours	**5.37**	**20.10**	**4.52**	**6.60**	**0.24**	**0.78**

method on the real-captured data from [74]. Tab. 3 shows the results. We train our method using the same training data as AvatarJLM [74]. Our method outperforms all other methods in all metrics.

4.3 Ablation Studies

We ablate our method to illustrate the effectiveness and essentiality of each component. Accordingly, we assess the performance of our method across various IK methods and swivel angle estimation approaches.

Design Choices of IK Methods. The upper part of Table 4 shows a performance comparison of different IK approaches. Results show that our approach outperforms alternative options.

Design Choices of Swivel Angle Estimation. We assess different methods for determining the swivel angle in the lower section of Table 4. Experiments show removing the mid-joint loss(No $\mathcal{L}_{\text{mid}}$) will give worse results, indicating the mid-joint loss can regularize the swivel angle prediction to reach the best position. However, if we remove the swivel angle loss and only optimize the mid-joint position (No $\mathcal{L}_{\text{swivel}}$), it is still worse than our default setting because swivel also helps regularize the network optimization. Besides, predicting the swivel angle as a single value (Swivel-Single) is worse than predicting that as a continuous representation. We also tried searching the swivel angle using the traditional method outlined in [23] (Swivel-Search), yielding the least favorable results. This can be explained by the algorithm's limitation of merely verifying the swivel angle's validity without leveraging motion priors learned from data.

4.4 Generalization to Other Tasks

Our method is a general solution for human pose estimation from sparse tracking signals and can be applied to other tasks. To demonstrate its applicability, we evaluate the performance on the task of full-body pose estimation from the 3D positions of 5 end-effectors, i.e., the head, two wrists, two ankles (orientations are not provided). This task resembles motion capture- or motion editing settings where only end-effector positions are available. We employ three commonly utilized network backbones — Transformer, RNN, and MLP — and conduct direct comparisons of the same model's performance when trained using SMPL and our method. We train all methods on the CMU- and the BioMotionLab dataset and test them on the MPI dataset. Experiments show that our method achieves significantly better results on all metrics (Table 5). *F-PE* denotes foot error, measured on the ankle joint.

4.5 Limitations and Disscussion

Our method shows significant improvements over previous approaches for full-body IK tasks across multiple datasets, though some limitations still remain. First, a prerequisite for our solver is knowledge of the user's body dimensions, specifically the lengths of the skeleton bones. This requirement aligns with prior work that assumes specific dimensions of the body to be known [11, 13, 21, 74]. For a wider adoption of our method, the body dimensions of the user could be obtained by direct measurement through calibration or body shape estimation, as addressed in our recent work EgoPoser [20]. Second, when only head and hand signals are available, full-body reconstruction can still be challenging for complex motions, such as a person sitting and crossing their legs. Third, since this paper focuses on designing an effective neural-analytic IK framework for full-body motion estimation tasks, we utilized network architectures similar to those in previous studies [21, 74]. Based on our framework, future work could explore more advanced backbone models for this task.

5 Conclusion

We have presented MANIKIN, a novel IK paradigm based on realistic biomechanical human motion constraints and analytic geometry for full-body pose estimation from end-effector poses. MANIKIN is differentiable and allows end-to-end training. Our approach predicts the base joint position and swivel angle based on neural networks and provides a closed-form solution for the joint angle configuration of human limbs. This enables accurate end-effector control, full-body pose tracking, and physically plausible predictions. MANIKIN achieves new state-of-the-art performance on full-body pose estimation, surpassing previous optimization-based and pure learning-based methods while keeping an extremely fast inference time. Our solution provides a new paradigm for IK problems with considerable potential for real-world applications.

Acknowledgements:
We sincerely thank Changlin Qian for his help with figures and Xintong Liu for her help with videos.

References

1. CMU MoCap Dataset. http://mocap.cs.cmu.edu/ (2004) 11
2. RootMotion Final IK. https://assetstore.unity.com/packages/tools/animation/final-ik-14290 (2018) 2, 3
3. Ahuja, K., Ofek, E., Gonzalez-Franco, M., Holz, C., Wilson, A.D.: Coolmoves: User motion accentuation in virtual reality. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **5**(2), 1–23 (2021) 2, 3
4. Aliakbarian, S., Cameron, P., Bogo, F., Fitzgibbon, A., Cashman, T.J.: Flag: Flow-based 3d avatar generation from sparse observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13253–13262 (2022) 2, 4
5. Aliakbarian, S., Saleh, F., Collier, D., Cameron, P., Cosker, D.: Hmd-nemo: Online 3d avatar motion generation from sparse observations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9622–9631 (2023) 3
6. Aristidou, A., Lasenby, J.: Fabrik: A fast, iterative solver for the inverse kinematics problem. Graphical Models **73**(5), 243–260 (2011) 3
7. Armani, R., Qian, C., Jiang, J., Holz, C.: Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging. In: ACM SIGGRAPH 2024 Conference Papers. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (2024) 4
8. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016) 3, 4
9. Choutas, V., Bogo, F., Shen, J., Valentin, J.: Learning to fit morphable models. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 160–179. Springer (2022) 2, 3
10. Desmurget, M., Prablanc, C.: Postural control of three-dimensional prehension movements. Journal of neurophysiology **77**(1), 452–464 (1997) 5
11. Dittadi, A., Dziadzio, S., Cosker, D., Lundell, B., Cashman, T.J., Shotton, J.: 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. pp. 11687–11697 (2021) 2, 4, 14
12. Dobrowolski, P.: Swing-twist decomposition in clifford algebra (2015) 10
13. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: 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, 4, 5, 11, 12, 14
14. Goldenberg, A., Benhabib, B., Fenton, R.: A complete generalized solution to the inverse kinematics of robots. IEEE Journal on Robotics and Automation **1**(1), 14–20 (1985) 3

15. Grassia, F.S.: Practical parameterization of rotations using the exponential map. Journal of graphics tools **3**(3), 29–48 (1998) 9
16. Grochow, K., Martin, S.L., Hertzmann, A., Popović, Z.: Style-based inverse kinematics. In: ACM SIGGRAPH 2004 Papers, pp. 522–531 (2004) 3
17. Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T.H., Tai, C.J., Akbay, M., Wang, Z., et al.: Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics (ToG) **39**(4), 87–1 (2020) 4
18. Han, S., Wu, P.c., Zhang, Y., Liu, B., Zhang, L., Wang, Z., Si, W., Zhang, P., Cai, Y., Hodan, T., et al.: Umetrack: Unified multi-view end-to-end hand tracking for vr. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 4
19. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: 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) 4
20. Jiang, J., Streli, P., Meier, M., Fender, A., Holz, C.: EgoPoser: Robust Real-Time Ego-Body Pose Estimation in Large Scenes. arXiv preprint arXiv:2308.06493 (2023) 4, 14
21. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: 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. pp. 443–460. Springer (2022) 2, 3, 4, 5, 7, 11, 12, 14
22. Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 4
23. Kallmann, M.: Analytical inverse kinematics with body posture control. Computer animation and virtual worlds **19**(2), 79–91 (2008) 2, 3, 6, 13
24. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018) 4
25. Kang, M., Cho, Y., Yoon, S.E.: Rcik: Real-time collision-free inverse kinematics using a collision-cost prediction network. IEEE Robotics and Automation Letters **7**(1), 610–617 (2021) 3
26. Kaufmann, M., Zhao, Y., Tang, C., Tao, L., Twigg, C., Song, J., Wang, R., Hilliges, O.: Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11510–11520 (2021) 4
27. Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3d digital humans. ACM Transactions on Graphics (TOG) **42**(6), 1–12 (2023) 4
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015) 10
29. Korein, J.U.: A geometric investigation of reach. MIT press (1986) 2, 6
30. Lee, S., Starke, S., Ye, Y., Won, J., Winkler, A.: Questenvsim: Environment-aware simulated motion tracking from sparse sensors. arXiv preprint arXiv:2306.05666 (2023) 4
31. Lemay, M.A., Crago, P.E.: A dynamic model for simulating movements of the elbow, forearm, and wrist. Journal of biomechanics **29**(10), 1319–1330 (1996) 5

32. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023) 4
33. Li, J., Bian, S., Liu, Q., Tang, J., Wang, F., Lu, C.: Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12933–12942 (2023) 3
34. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3383–3393 (2021) 3
35. Li, S., Jiang, J., Ruppel, P., Liang, H., Ma, X., Hendrich, N., Sun, F., Zhang, J.: A mobile robot hand-arm teleoperation system by vision and imu. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10900–10906. IEEE (2020) 3
36. Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12. pp. 332–347. Springer (2015) 4
37. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13147–13156 (2022) 4
38. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1954–1963 (2021) 4
39. Loper, M., Mahmood, N., Black, M.J.: Mosh: motion and shape capture from sparse markers. ACM Trans. Graph. **33**(6), 220–1 (2014) 2
40. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 2, 4, 5
41. Ma, H., Chen, L., Kong, D., Wang, Z., Liu, X., Tang, H., Yan, X., Xie, Y., Lin, S.Y., Xie, X.: Transfusion: Cross-view fusion with transformer for 3d human pose estimation. arXiv preprint arXiv:2110.09554 (2021) 4
42. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019) 2, 5, 11
43. Marić, F., Giamou, M., Hall, A.W., Khoubyarian, S., Petrović, I., Kelly, J.: Riemannian optimization for distance-geometric inverse kinematics. IEEE Transactions on Robotics **38**(3), 1703–1722 (2021) 3
44. Mollyn, V., Arakawa, R., Goel, M., Harrison, C., Ahuja, K.: Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2023) 4
45. Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., Weber, A.: Documentation mocap database hdm05. Tech. Rep. CG-2007-2, Universität Bonn (June 2007) 11
46. Parger, M., Mueller, J.H., Schmalstieg, D., Steinberger, M.: Human upper-body inverse kinematics for increased embodiment in consumer-grade virtual reality. In: Proceedings of the 24th ACM symposium on virtual reality software and technology. pp. 1–10 (2018) 3

47. Parker, J.K., Khoogar, A.R., Goldberg, D.E.: Inverse kinematics of redundant robots using genetic algorithms. In: 1989 IEEE International Conference on Robotics and Automation. pp. 271–272. IEEE Computer Society (1989) 3
48. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019) 3, 4
49. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Computer Vision and Pattern Recognition (CVPR) (2017) 4
50. Ponton, J.L., Yun, H., Aristidou, A., Andujar, C., Pelechano, N.: Sparseposer: Real-time full-body motion reconstruction from sparse data. ACM Transactions on Graphics **43**(1), 1–14 (2023) 4
51. Prokopenko, R., Frolov, A., Biryukova, E., Roby-Brami, A.: Assessment of the accuracy of a human arm model with seven degrees of freedom. Journal of biomechanics **34**(2), 177–185 (2001) 5
52. Raikova, R.: A general approach for modelling and mathematical investigation of the human upper limb. Journal of biomechanics **25**(8), 857–867 (1992) 5
53. Ruppel, P., Hendrich, N., Starke, S., Zhang, J.: Cost functions to specify full-body motion and multi-goal manipulation tasks. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3152–3159. IEEE (2018) 3
54. Seireg, A., Arvikar, R.: Biomechanical analysis of the musculoskeletal structure for medicine and sports. (No Title) (1989) 5
55. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019) 3
56. Streli, P., Armani, R., Cheng, Y.F., Holz, C.: 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. pp. 1–16 (2023) 4
57. Sumner, R.W., Zwicker, M., Gotsman, C., Popović, J.: Mesh-based inverse kinematics. ACM transactions on graphics (TOG) **24**(3), 488–495 (2005) 3
58. Sun, X., Li, C., Lin, S.: An integral pose regression system for the eccv2018 posetrack challenge. arXiv preprint arXiv:1809.06079 (2018) 4
59. Sun, X., Xiao, B., Liang, S., Wei, Y.: Integral human pose regression. arXiv preprint arXiv:1711.08229 (2017) 4
60. Tolani, D., Goswami, A., Badler, N.I.: Real-time inverse kinematics techniques for anthropomorphic limbs. Graphical models **62**(5), 353–388 (2000) 2, 3, 6
61. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision **2**(5), 2–2 (2002) 11
62. Von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Computer graphics forum. vol. 36, pp. 349–360. Wiley Online Library (2017) 4
63. Wang, L.C., Chen, C.C.: A combined optimization method for solving the inverse kinematics problems of mechanical manipulators. IEEE Transactions on Robotics and Automation **7**(4), 489–499 (1991) 3
64. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022) 2, 4
65. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Visibility aware human-object interaction tracking from single rgb camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4757–4768 (2023) 4

66. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022) 4
67. Yang, D., Kim, D., Lee, S.H.: Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In: Computer Graphics Forum. vol. 40, pp. 265–275. Wiley Online Library (2021) 2, 3, 4
68. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13167–13178 (2022) 4
69. Yi, X., Zhou, Y., Xu, F.: Transpose: real-time 3d human translation and pose estimation with six inertial sensors. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021) 4
70. Yi, X., Zhou, Y., Xu, F.: Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024) 4
71. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J.Y., Lee, K.M., Molchanov, P., Kautz, J., Honari, S., Ge, L., et al.: 3d hand pose estimation: From current achievements to future goals. arXiv preprint arXiv:1712.03917 (2017) 4
72. Zhang, X., Bhatnagar, B.L., Guzov, V., Starke, S., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. In: European Conference on Computer Vision). Springer (October 2022) 3
73. Zhao, J., Badler, N.I.: Inverse kinematics positioning using nonlinear programming for highly articulated figures. ACM Transactions on Graphics (TOG) **13**(4), 313–336 (1994) 3
74. Zheng, X., Su, Z., Wen, C., Xue, Z., Jin, X.: Realistic full-body tracking from sparse observations via joint-level modeling. arXiv preprint arXiv:2308.08855 (2023) 2, 3, 4, 5, 7, 11, 12, 13, 14
75. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 186–201. Springer (2016) 4
76. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019) 8