PoseAugment: Generative Human Pose Data Augmentation with Physical Plausibility for IMU-based Motion Capture

Zhuojun Li[1,2*], Chun Yu[1,2,3**], Chen Liang[1,2], and Yuanchun Shi[1,2,3]

[1] Department of Computer Science and Technology, Tsinghua University, China
[2] Key Laboratory of Pervasive Computing, Ministry of Education, China
[3] Qinghai University, China

Abstract. The data scarcity problem is a crucial factor that hampers the model performance of IMU-based human motion capture. However, effective data augmentation for IMU-based motion capture is challenging, since it has to capture the physical relations and constraints of the human body, while maintaining the data distribution and quality. We propose PoseAugment, a novel pipeline incorporating VAE-based pose generation and physical optimization. Given a pose sequence, the VAE module generates infinite poses with both high fidelity and diversity, while keeping the data distribution. The physical module optimizes poses to satisfy physical constraints with minimal motion restrictions. High-quality IMU data are then synthesized from the augmented poses for training motion capture models. Experiments show that PoseAugment outperforms previous data augmentation and pose generation methods in terms of motion capture accuracy, revealing a strong potential of our method to alleviate the data collection burden for IMU-based motion capture and related tasks driven by human poses.

1 Introduction

Human motion capture (MoCap) with IMUs has become a rising topic in recent research [16,20,30,43,47,49,50], due to its advantages in terms of power efficiency, privacy preservation, and usability compared with CV-based methods. However, training MoCap models requires large IMU and corresponding pose data, which are normally collected via professional MoCap systems, such as OptiTrack [31] and Vicon [42], which are both expensive and time-consuming. Xsens [17] reduces the cost by using 17 IMUs, but it is still inconvenient for individuals. Besides, human motion distribution is also highly personalized and task-related. The IMU signals can also be affected by different mount positions and hardware errors, resulting in low data transferability across different tasks.

To avoid the data collection burden, we need effective data augmentation methods. Current methods synthesize IMU data directly from open-source pose

* Email: lizj23@mails.tsinghua.edu.cn
** Corresponding author: chunyu@mail.tsinghua.edu.cn

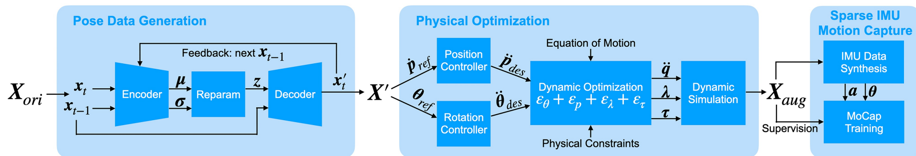


Fig. 1: Method overview. Given an original pose sequence $\boldsymbol{X}_{ori} = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_T\}$, we learn a VAE model to generate new poses $\boldsymbol{X}'$ frame-by-frame autoregressively. It captures the motion variance and can generate infinite poses within this distribution. Then, the motion jitter and artifacts are optimized by solving a quadratic optimization problem, which is based on a dual position ($\boldsymbol{p}$) and rotation ($\boldsymbol{\theta}$) PD controller and physical constraints on reaction forces $\boldsymbol{\lambda}$ and torques $\boldsymbol{\tau}$. The final augmented poses $\boldsymbol{X}_{aug}$ can be used to augment the dataset by synthesizing IMU data.

datasets [20, 47, 49, 50], such as AMASS [27]. However, it is unsuitable for specialized tasks where the training samples are hard to collect and have a unique distribution, like in disease/sports analysis [5, 39]. Researchers also leveraged noise-based methods to directly augment the IMU data [49,50], but these methods are unable to capture the physical constraints among multiple IMU nodes, thus only having marginal improvements. Recently, generative deep models (*e.g.* VAE, diffusion) demonstrated a huge potential in generating natural human motions for animations. However, these methods mainly use high-level texts or actions to condition pose generation, which will affect the data distribution and quality of the original dataset. We found generating poses that are close to the original data is more suitable for IMU-based motion capture in our evaluations.

We propose PoseAugment, a two-stage human pose data augmentation pipeline for IMU-based motion capture (Fig. 1). Given a reference pose, the first stage leverages a VAE [32] model to generate infinite new poses, covering the original data space comprehensively. The encoder encodes the difference between adjacent frames into a latent representation. Then, the decoder reconstructs the next frame based on the previous frame and the latent vector. The generation process is guided by the original data frame-by-frame autoregressively, which maintains the low-level motion distribution with the original data.

The second stage optimizes poses with physical constraints to improve motion naturalness, as the generated poses may have artifacts (jitter, unnatural motion). We apply a dual PD controller [49] to provide positional and rotational motion references. Different from prior work [38, 49], our method only has minimal physical constraints, which do not assume the motion to be on flat ground and do not need contact labeling, thus supporting a wider range of motions (*e.g.* climbing stairs, skating) and achieving a soft estimation of reaction forces.

After that, high-quality IMU data could be synthesized from the augmented poses for MoCap training. We compared our data augmentation method with adding random noise, and other generative methods (MotionAug [26], ACTOR [34], MDM [41]) on training MoCap models (TransPose [50]). Results show that PoseAugment outperformed these methods in pose accuracy. Future works can leverage PoseAugment as a pre-trained generative model to alleviate the data collection burden in IMU-based MoCap and other tasks driven by human poses.

Our main contribution in this paper is two-fold:

1. We propose PoseAugment, a novel pose data augmentation pipeline incorporating pose generation and physical optimization. Future works can benefit from our method to reduce the data collection burden.
2. We thoroughly evaluate PoseAugment with previous pose generation methods, which shows a significant improvement in IMU-based motion capture.

2 Related Work

We will first review the context of IMU-based motion capture, then introduce IMU data augmentation, pose data generation, and physical optimization, mainly from the perspective of their influences on the data quality.

2.1 IMU-based Motion Capture

To overcome the occlusion and privacy issues of CV-based motion capture methods, a series of works focus on reconstructing human poses solely from sparse IMU sensors, worn on key body joints.

SIP [43] first proposes a 6-IMU setting, and regards the motion capture task as a non-convex optimization problem. DIP [16] leverages bidirectional RNN to regress human poses, providing a learning-based approach. TransPose [50] proposes a three-stage network structure and achieves a better regression accuracy both in pose measurements and global translations. The follow-up study PIP [49] optimizes poses with physical constraints to fix motion artifacts, but only supports motions on the ground. TIP [20] and DiffIP [47] applied Transformer and diffusion models to this task, increasing the model robustness to different IMU configurations. IMU Poser [30] further reduces IMU numbers to daily settings (1 to 3), making inertial motion capture more pervasive. AvatarPoser [19], BoDiffusion [2], and AGRoL [7] aim to reconstruct full poses only from positions and rotations of VR devices, which could also be regarded as preprocessed IMU data.

All of these works require large IMU and pose data for model training. Apart from network improvements, our motivation is that augmenting data with high quality is equally crucial to model performance, especially for tasks with little available data to use. Therefore, human pose data augmentation in the context of IMU-based motion capture is worthy of exploring.

2.2 IMU Data Augmentation

Inspired by image augmentation in CV fields, numerous methods have been adapted to IMU signals, including noise-based methods (*e.g.* Jitter) [18, 45, 48] and generative methods (*e.g.* GAN) [3]. However, these methods only work in classification tasks, since the data labels need not change after data augmentation. But for IMU-based MoCap, the semantic changes in IMU signals after augmentation are ambiguous, making it hard to augment poses accordingly.

Due to this restriction, only TransPose [50] among works in Sec. 2.1 leverages Jitter to augment data. We demonstrate that direct augmentation on IMU signals with traditional methods only has marginal improvements to MoCap models (Sec. 4.3). Our method, on the other hand, augments poses and synthesizes the corresponding IMU signals, achieving better semantic control.

2.3 Pose Data Generation

Generating natural human poses is a basic problem in computer vision. From the model perspective, some works use autoencoders to reconstruct human poses, like VAE [12,23,26,33,34,40,46]. Recent works also leverage diffusion models to generate human poses [4,11,21,41,51,54], inspired by the success of AIGC. From the perspective of generation goals, some works generate similar poses based on a reference pose (M2M) [10,14,23,26,36], while a large amount of research converts data from other modalities to poses, like action-to-motion (A2M) [12, 34] and text-to-motion(T2M) [4, 11, 21, 41, 51, 54].

To augment poses for MoCap training, our goal is to generate more pose samples while maintaining the data distribution. To achieve the fine-grained motion control, we choose to augment poses in the M2M manner, instead of A2M and T2M, where the motions can not be controlled precisely. MotionAug [26] proposed a similar VAE/IK and optimization solution like ours. But it needs users to annotate IK keyframes and is only limited to 8 motion types. So, we do not regard it as an off-the-shelf general-purpose solution. MVAE [23] proposed a frame-to-frame generation design. A new frame is predicted using the previous frame and the difference between frames. This is more suitable for IMU data synthesis since the IMU signals are also frame-based when taking differentials on poses. Therefore, we adopted this design and improved the network structure for our task (Sec. 3.1).

2.4 Physical Optimization

Physics-based optimization is used to correct motion artifacts after motion capture. Some works use direct constraints (*e.g.* foot contacts) to restrict body motions [6,52,58]. Some works impose physical constraints as loss terms when training motion capture models [13, 29, 37]. Recent optimization-based methods estimate the physical properties of the human body (*e.g.* torques, reaction forces). They consider the human body as an articulated rigid body system, and calculate the dynamic properties based on the equation of motion [22,24,35,44,53]. Then, the rigid body system could be actuated by the optimized dynamic properties to generate motions [38, 49, 56].

PIP [49] proposes a dual PD controller to calculate desired motion accelerations. However, it only supports motions on the ground and needs contact point labeling. We adopted the dual PD controller with a looser assumption and no contact information, supporting a wider variety of motions (*e.g.* climbing stairs).

3 Method

The goal of our method is to augment human pose data for MoCap training. The input is a pose sequence, and the output is augmented poses and the corresponding IMU signals (including accelerations and rotations at key body joints). PoseAugment incorporates two stages: (1) **Pose Data Generation** generates pose sequences with both high fidelity and diversity (Sec. 3.1); (2) **Physical Optimization** corrects motion artifacts to improve motion naturalness (Sec. 3.2). Then, the IMU data are synthesized from poses by taking derivatives.

3.1 Pose Data Generation

We aim to generate new poses with minimal reconstruction errors while diverse enough to cover the motion space, which is a typical trade-off problem. All pose data are in SMPL [25] format, which considers the human body as a 24-joint articulated rigid body system. We will first introduce the data representation, and then detail the VAE model structure, training, and inferencing.

Motion Frame Representation. In this stage, we represent pose data as a motion frame sequence $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$, each frame

$$\boldsymbol{x} = [\boldsymbol{p}_{root}, \boldsymbol{v}_{root}, \boldsymbol{\theta}_{root}, \boldsymbol{p}_{joint}, \boldsymbol{v}_{joint}, \boldsymbol{\theta}_{joint}] \in \mathbb{R}^{240} \tag{1}$$

where $\boldsymbol{p}, \boldsymbol{v}, \boldsymbol{\theta}$ represent 3D position, 3D velocity and 6D rotation [57] respectively. The subscript *root* means the SMPL root joint (pelvis) in the global frame, while *joint* means other 19 SMPL joints (excluding L/R feet and L/R hands, which are simplified as identical rotations) in the root frame. So, the total dimension of a motion frame is $3 + 3 + 6 + (3 + 3 + 6) \times 19 = 240$. To reconstruct poses from the augmented motion frames, we only use the original $\boldsymbol{p}_{root}, \boldsymbol{\theta}_{root}$, and the augmented $\boldsymbol{\theta}_{joint}$. We added other $\boldsymbol{p}$ and $\boldsymbol{v}$ terms as an additional reconstruction task for the VAE model to better learn global motion features.

VAE Model Structure. When designing the VAE model, the key rule is to make it powerful enough to reconstruct accurate poses, while emphasizing the latent vector to ensure pose diversity, thus covering the motion space comprehensively. We leveraged several successful designs of MVAE [23], including the autoregressive prediction and the MoE architecture [28, 55]. We further propose several improvements to make the VAE model fit our goal better.

Given two frames $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$, the encoder first encodes them into latent mean and standard deviation vectors $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{40}$. After reparameterization [32], the latent vector $\boldsymbol{z} \in \mathbb{R}^{40}$, together with frame $\boldsymbol{x}_{t-1}$, will be decoded to predict frame $\boldsymbol{x}'_t$, which will be used as the next $\boldsymbol{x}_{t-1}$ to generate poses autoregressively.

Unlike MVAE, we incorporate two separate residual blocks in the encoder and one residual block in the decoder. They deepen the network while also ensuring that the motion and latent features would not be ignored during inferencing, which ensures both fidelity and diversity. Furthermore, we add a separate layer to expand the latent vector to $\boldsymbol{z}_{exp} \in \mathbb{R}^{240}$ (with the same length as a motion frame) to extract the compressed latent features. Then, $\boldsymbol{z}_{exp}$ is input to both the MoE gate network and decoding layers, together with $\boldsymbol{x}_{t-1}$. The VAE structure details are shown in Appendix A.1.

VAE Model Training. Following the former practice [49,50], we used AMASS [27] and DIP train split [16] to train the VAE model, and evaluated it on the DIP test split. The training dataset contains about 45.6h poses sampled at 60Hz.

To make pose prediction stable, we trained the VAE model with mini-batches and scheduled sampling, following MVAE [23]. The pose sequences are first cut into mini-batches with length L. Within each mini-batch, each pose prediction $\boldsymbol{x}'_t$ will be used as the next conditione frame $\boldsymbol{x}_{t-1}$ with probability $(1-p)$, (ground truth $\boldsymbol{x}_t$ with probability p). p will decrease gradually from 1 to 0 during training, transforming the training process from fully supervised to fully autoregressive. This method will make the VAE model robust to self-prediction errors.

In the β-VAE design [15], the training loss is $loss_{reconst} + \beta \cdot loss_{KL}$, where β is introduced to balance the reconstruction accuracy and the latent vector diversity. We found the pose diversity is highly sensitive to β. As a result, we fine-tuned β to be 3×10^{-3} empirically, which achieved a balance of pose fidelity and diversity (Sec. 4.1). More training details are provided in Appendix A.2.

VAE Model Inferencing. After training, the VAE model is used to augment poses autoregressively, during which the prediction errors may accumulate in motion frames, resulting in unrecoverable deviations from the ground truth. Since we need to augment poses with lengths far longer than the mini-batch, (200 frames for [49, 50]), how to ensure the temporal stability is a crucial problem. The general idea is to let the ground truth guide pose generation. We designed two techniques named **best sampling** and **motion refinement** to tackle it.

For *best sampling*, at each timestamp, we use $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$ to repeatedly generate N different predictions $\{\boldsymbol{x}'^{(1)}_t, \boldsymbol{x}'^{(2)}_t, \ldots, \boldsymbol{x}'^{(N)}_t\}$, from which we choose the closest frame to $\boldsymbol{x}_t$ in MSE error as the final prediction. It can discard predictions with large errors to make data augmentation more stable, but may harm pose diversity on the other hand. We choose $N = 2$ in our evaluations.

Furthermore, *motion refinement* restricts frame values in a reasonable range, to avoid error accumulation. For $\boldsymbol{p}_{root}$ and $\boldsymbol{p}_{joint}$, we let their distances to $\boldsymbol{p}_{GT}$ (ground truth) within $d_p = 15cm$. For $\boldsymbol{v}_{root}$ and $\boldsymbol{v}_{joint}$, we force their values to be no larger than $d_v \cdot \boldsymbol{v}_{GT}$ and no smaller than $(1/d_v) \cdot \boldsymbol{v}_{GT}$, where $d_v = 2.0$. For $\boldsymbol{\theta}_{root}$ and $\boldsymbol{\theta}_{joint}$, since each 6D rotation consists of two orthogonal unit vectors (but networks do not ensure that), we renormalized them after each prediction.

They significantly improve the prediction stability. Tuning these hyperparameters also helps to balance the pose fidelity and diversity after training.

3.2 Physical Optimization

The pose data augmented by the VAE model are purely kinematics-based, but do not consider how the human body is actuated by dynamic properties like forces and torques. In this stage, we perform physical optimization on pose data to improve motion naturalness and temporal consistency. We regard the human body as a 24-joint articulated rigid body system as introduced by [8], which is actuated by the internal torques and external reaction forces. Then, the optimal dynamic properties are calculated by solving a quadratic optimization problem, which satisfies the motion of equation [8], the desired accelerations given by a dual PD controller [49], and physical constraints on reaction forces and torques. Using these optimized dynamic properties, the final poses could be simulated.

Problem Definition. Similar to PIP [49], we use a floating-base rigid body system as the human body, which is controlled by reaction forces, torques, and non-linear effects like gravity and Coriolis forces. Pose data in this stage are represented in RBDL format [9], which has the same $J = 24$ joints as SMPL [25] with a different order (root indices are both 0). We denote joint positions as $\boldsymbol{p} \in \mathbb{R}^{3J}$ and joint rotations as $\boldsymbol{\theta} \in \mathbb{R}^{3J}$ (in Euler angles). The corresponding velocities and accelerations would be $\dot{\boldsymbol{p}}, \ddot{\boldsymbol{p}}, \dot{\boldsymbol{\theta}}, \ddot{\boldsymbol{\theta}}$ respectively. A motion frame is represented by the root position and joint rotations (in parent frames) as $\boldsymbol{q} = [\boldsymbol{p}_{root}, \boldsymbol{\theta}] \in \mathbb{R}^N$, with $N = 3 + 3J$. The system is actuated by internal torques $\boldsymbol{\tau} \in \mathbb{R}^N$ on each DoF and external reaction forces $\boldsymbol{\lambda} \in \mathbb{R}^{3(J-1)}$.

The system follows the equation of motion [8]

$$\boldsymbol{\tau} + \boldsymbol{J}(\boldsymbol{q})^T \boldsymbol{\lambda} = \boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{h}(\boldsymbol{q}, \ddot{\boldsymbol{q}}) \tag{2}$$

where $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ is the inertial matrix; $\boldsymbol{h} \in \mathbb{R}^N$ is the non-linear effect term; $\boldsymbol{J} \in \mathbb{R}^{3(J-1) \times N}$ is the Jacobians of $(J - 1)$ joints [8]. Each joint Jacobian $\boldsymbol{J}_i \in \mathbb{R}^{3 \times N}$ converts the motion frame velocity to the global joint velocity as

$$\dot{\boldsymbol{p}}_i = \boldsymbol{J}\dot{\boldsymbol{q}} \tag{3}$$

Derivations can be found from rigid body dynamics [8]. Unlike PIP [49], we do not assume $\boldsymbol{\lambda}$ all come from flat ground. The reaction forces can act on any non-root joint to support a wider range of motions, which we will detail later.

Dynamic Optimization. The optimization problem can be formalized as:

$$\begin{aligned}
\arg\min_{\ddot{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\tau}} \quad & (\varepsilon_\theta + \varepsilon_p + \varepsilon_\lambda + \varepsilon_\tau) && \text{(energy terms)} \\
s.t. \quad & \boldsymbol{\tau} + \boldsymbol{J}^T \boldsymbol{\lambda} = \boldsymbol{M}\ddot{\boldsymbol{q}} + \boldsymbol{h} && \text{(equation of motion)} \\
& |\dot{\boldsymbol{p}}^T \boldsymbol{\lambda}| \leq \delta && \text{(stationary support)} \\
& \boldsymbol{\lambda} \in \mathcal{F} && \text{(friction constraint)}
\end{aligned} \tag{4}$$

This optimization is a Quadratic Programming problem, which gives the best $\ddot{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\tau}$ that minimize the four energy terms and satisfy the equation of motion and physical constraints on reaction forces. The optimization problem is solved using the Augmented Lagrangian algorithm [1] with sparse matrices.

In this optimization problem, the dual PD controller terms ε_{θ} and ε_p and the friction constraint $\boldsymbol{\lambda} \in \mathcal{F}$ are similar to PIP [49]. Different from PIP, we have a more general assumption on the regularization and constraint of the reaction forces $\boldsymbol{\lambda}$, which supports a wider range of motions and does not need contact labeling. The energy terms and physical constraints are detailed as follows.

Dual PD Controller Terms $\varepsilon_{\boldsymbol{\theta}}$ and $\varepsilon_{\boldsymbol{p}}$. Given a reference motion, the dual PD controller is used to calculate the desired positional and rotational accelerations $\ddot{\boldsymbol{p}}_{des}$ and $\ddot{\boldsymbol{\theta}}_{des}$. To achieve this, we first need to calculate the reference joint velocities $\dot{\boldsymbol{p}}_{ref}$ and rotations $\boldsymbol{\theta}_{ref}$ of the next timestamp. Given the current motion status of the optimized system, $\ddot{\boldsymbol{\theta}}_{des}$ can be written as

$$\ddot{\boldsymbol{\theta}}_{des} = k_{p_{\theta}}(\boldsymbol{\theta}_{ref} - \boldsymbol{\theta}_{cur}) - k_{d_{\theta}}\dot{\boldsymbol{\theta}}_{cur} \tag{5}$$

and $\ddot{\boldsymbol{p}}_{des}$ can be written as

$$\ddot{\boldsymbol{p}}_{des} = k_{p_p}(\dot{\boldsymbol{p}}_{ref}\varDelta t) - k_{d_p}\dot{\boldsymbol{p}}_{cur} \tag{6}$$

where $\varDelta t = (1/60)s$. $k_{p_{\theta}}, k_{d_{\theta}}, k_{p_p}, k_{d_p}$ are the gain parameters, which are empirically set to $1800, 60, 2400, 60$ (related to how the motion kinematics are updated).

Then, the energy terms give the best $\ddot{\boldsymbol{q}}$ that minimizes the distance of joint accelerations to $\ddot{\boldsymbol{\theta}}_{des}$ and $\ddot{\boldsymbol{p}}_{des}$:

$$\begin{aligned} \varepsilon_{\theta} &= ||\ddot{\boldsymbol{q}}_{3:} - \ddot{\boldsymbol{\theta}}_{des}||^2 \\ \varepsilon_p &= ||\ddot{\boldsymbol{p}} - \ddot{\boldsymbol{p}}_{des}||^2, \quad \ddot{\boldsymbol{p}} = \dot{\boldsymbol{J}}\dot{\boldsymbol{q}} + \boldsymbol{J}\ddot{\boldsymbol{q}} \end{aligned} \tag{7}$$

where $\ddot{\boldsymbol{p}}$ is the derivative of Eq. (3). Readers can see [8, 49] for more details.

Regularization Terms $\varepsilon_{\boldsymbol{\lambda}}$ and $\varepsilon_{\boldsymbol{\tau}}$. $\boldsymbol{\lambda}$ represents the external reaction forces on non-root body joints. Previous works only consider ground reaction forces (GRF). They assume the character moves on flat ground with no sliding. To calculate GRF, these methods either need foot contact prediction [37, 38] or labeling with threshold-based methods [49]. These approaches cannot deal with other common motions like climbing stairs, holding the handrail with hands, sitting on a chair, *etc.*, where the reaction forces may act on all body joints.

Instead, we propose a novel approach based on the **dis-to-root principle**. We allow the reaction forces to act on any non-root joint to support general human motions. Noticing that the reaction forces largely come from the ground, handrails, walls and act on body joints that are relatively far away from the center of mass of the human body, we will give larger penalties to joints closer to the root. We first calculate the distance between root and joint i as d_i, then regularize $\boldsymbol{\lambda}$ as

$$\varepsilon_{\lambda} = k_{\lambda}\sum_{i=1}^{23}||\frac{1}{d_i}\boldsymbol{\lambda}_i||^2 \tag{8}$$

where $k_\lambda = 0.02$ is the weight coefficient. For motions on flat ground, the foot joints will have the largest reaction forces, which is in accordance with previous work [37, 38, 49]. Note that this is only one requisite on $\boldsymbol{\lambda}$, together with other constraints in Eq. (4), we will make the optimization for $\boldsymbol{\lambda}$ more reasonable.

For $\boldsymbol{\tau}$, the regularization term is written as

$$\varepsilon_\tau = k_{root}||\boldsymbol{\tau}_{:6}||^2 + k_{joint}||\boldsymbol{\tau}_{6:}||^2 \tag{9}$$

where $k_{root} = 0.05$ is the weight coefficient of the root joint (the root position and rotation), while $k_{joint} = 0.02$ is for other joints [49].

Stationary Support Constraint. To further restrict reaction forces, we propose the **stationary support** constraint, which assumes all the objects that provide reaction forces must be still. Mathematically, it states

$$|\dot{\boldsymbol{p}}_i^T \boldsymbol{\lambda}_i| \leq \delta, \quad (1 \leq i \leq 23) \tag{10}$$

where $\dot{\boldsymbol{p}}_i$ and $\boldsymbol{\lambda}_i$ are the velocity of and the reaction force on joint i. $\delta = 10$ is a pre-defined threshold.

This equation has two connotations. First, it prevents the joint velocity and the reaction force to be in the same direction, which means the joint cannot move towards, or be pushed by the object that exerts the reaction force. Second, it allows the reaction force to be perpendicular to the velocity, which is common in motions like skating or running on a treadmill. This constraint cancels a large number of unreasonable reaction forces, while still supporting any environment with stationary objects (including but not limited to a ground plane).

Friction Constraint. We assume the friction force on a joint cannot exceed the maximum static friction force, while the joint can still slide on the object, as

$$|\lambda_{x_i}| \leq \mu\lambda_{y_i}, |\lambda_{z_i}| \leq \mu\lambda_{y_i}, \quad (1 \leq i \leq 23) \tag{11}$$

where $\mu = 0.6$ is the friction coefficient, and x, y, z denote the three dimensions.

3.3 IMU Data Synthesis

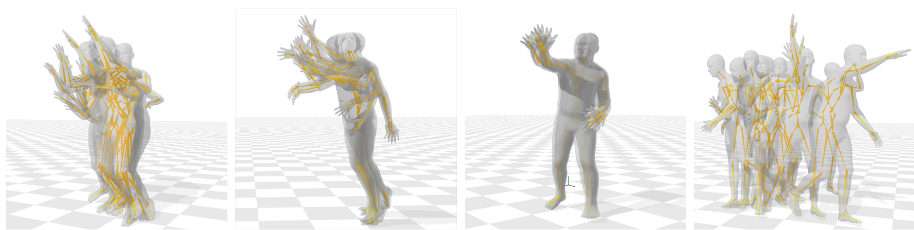
To train sparse IMU MoCap models, we need to synthesize the virtual IMU data, including global accelerations and rotations [4] at 6 body positions (2 wrists, 2 knees, head and pelvis). Given a pose sequence $[\boldsymbol{p}_{root}, \boldsymbol{\theta}_{root}, \boldsymbol{\theta}_{joint}]$, we first use Forward Kinematics [8] to calculate the global joint and mesh positions of the human body. Then, the accelerations and rotations on the 6 body positions could be simulated by taking differentials.

[4] Note that the raw IMU signals (in device local frames) would first be processed into the global frame before motion capture in practice.

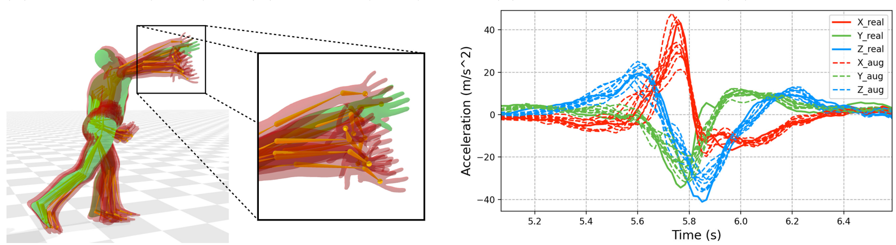
4 Experiments

In this section, we will first show our qualitative and quantitative evaluations on pose fidelity and diversity (Sec. 4.1), as well as on physical plausibility (Sec. 4.2). Then, the data augmentation performance is tested on training MoCap models, comparing with previous methods (Sec. 4.3).

4.1 Fidelity and Diversity



(a) MotionAug(M2M). **(b)** ACTOR(A2M). **(c)** MDM-M2M. **(d)** MDM-T2M.



(e) M2M pose generation by PoseAugment. The green pose is the ground truth, while the red ones are augmented. **(f)** Augmented accelerations of the right wrist by PoseAugment. The solid lines are the ground truth, while the dotted lines are augmented.

Fig. 2: Visualization of the motion *throwing a handball.* 10 motion sequences are generated by MotionAug, ACTOR, MDM-M2M, MDM-T2M, and our method.

A huge burden in pose data collection is to let the subjects repeat the same motion multiple times to cover the possible motion space. PoseAugment simulates this process by generating poses that satisfy: (1) **Fidelity**: the augmented poses should be close enough to the ground truth pose; (2) **Diversity**: the deviations within augmented poses should be large enough to cover the motion space. Therefore, with PoseAugment, researchers can focus more on the diversity of motion types, while reducing the motion repetition during data collection.

Qualitatively, we visualized the augmented poses by our baselines MotionAug, ACTOR, MDM-M2M, and MDM-T2M (Sec. 4.3), and our method respectively. As shown in Fig. 2, ACTOR and MDM-T2M are purely generated from high-level action labels or texts. The motion distribution is hard to control precisely, thus may violate the original data distribution. For other M2M methods, MotionAug uses a sequence-to-sequence model design, thus the pose diversity is also hard to control at the frame level. MDM-M2M is a variant of MDM, which denoises partially noised motions to generate similar data, but suffers from motion diversity. PoseAugment, on the other hand, generates poses frame-by-frame guided by the ground truth poses, which best simulates the repetitions of the same motion. The augmented IMU signals also achieve a high fidelity with the original data. We found this M2M way performed the best in our experiments (Sec. 4.3). More qualitative results of our method are provided in Appendix B.

Quantitatively, we define e_{pos} and e_{rot}, which are the positional and rotational joint errors between ground truth pose and the corresponding augmented poses, to reflect the reconstruction accuracy. For diversity, we also defined d_{pos} and d_{rot}, which are the mean positional and rotational standard deviations within each joint of the augmented poses. We augmented 4 times of new pose data on the AMASS dataset and got $e_{pos} = 4.54cm, d_{pos} = 0.77cm, e_{rot} = 10.86°, d_{rot} = 1.29°$ over 24 joints.

4.2 Physical Plausibility

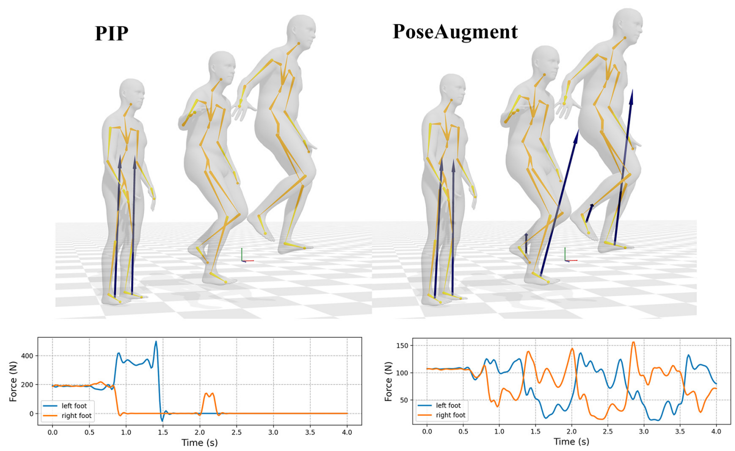


Fig. 3: The reaction force estimation of climbing stairs. We visualized the force vectors on two feet by the blue arrows, and the time sequences of vertical reaction forces. PIP fails when the subject is off-ground, while our method does not have this limitation.

As discussed in Sec. 3.2, we generalize the dynamic optimization by discarding the ground plane assumption and the requirement of contact labeling. We leverage the *distance-to-root* principle and the *stationary support* assumption to achieve a soft estimation of the reaction forces, broadening the supported motions compared with PIP [49].

Fig. 3 demonstrates an example of climbing stairs. The subject first stands at the bottom of the staircase and then climbs upward along the steps using two feet alternatively. Using previous threshold-based methods, the contact labels would be lost when the subject is off-ground. So, no reaction forces would be estimated. PoseAugment optimizes poses purely based on dynamic properties without contact labels. Though a small residual force may appear on unsuspended joints, it achieves a softer and more pervasive pose dynamic estimation.

Furthermore, we quantitatively show that the physical optimization module also improves the motion naturalness. We measured the jitter J (the 3rd derivative) of joint positions, including the original data, poses generated by VAE without optimization, and poses generated by VAE with optimization on the AMASS dataset. We got $J_{ori} = 2.11(100m/s^3)(SD = 2.19)$, $J_{VAE} = 5.15(SD = 4.38)$, and $J_{ours} = 2.32(SD = 2.27)$. These results show that the physical optimization significantly lowers the motion jitter introduced by the VAE noises, bringing the motion naturalness to a similar level to the original data.

4.3 Quantitative Evaluations

Method	$e_{SIP}(°)$	$e_{rot}(°)$	$e_{pos}(cm)$	$e_{mesh}(cm)$	Method	$e_{SIP}(°)$	$e_{rot}(°)$	$e_{pos}(cm)$	$e_{mesh}(cm)$
NoAug	36.38	17.12	11.07	12.42	NoAug	26.20	11.77	7.40	8.53
Jitter	−2.2%	−3.0%	−4.9%	−4.2%	Jitter	−1.2%	0.2%	0.7%	0.3%
MotionAug	−5.4%	**−13.5%**	−5.7%	**−6.7%**	MDM-M2M	−0.5%	−0.7%	−0.1%	−0.2%
Ours	**−14.5%**	−9.8%	**−8.0%**	−5.4%	Ours	**−7.6%**	**−9.1%**	**−7.8%**	**−8.6%**
NoAug	30.64	16.84	8.62	9.97	NoAug	25.76	11.77	7.23	8.37
Jitter	−2.0%	−1.5%	+0.7%	+0.4%	Jitter	1.4%	−0.7%	+1.0%	+0.2%
ACTOR	−2.7%	−4.9%	−2.6%	−2.1%	MDM-T2M	−1.0%	−2.5%	+0.9%	+0.3%
Ours	**−17.4%**	**−20.7%**	**−10.7%**	**−12.2%**	Ours	**−4.3%**	**−8.3%**	**−5.8%**	**−6.8%**

Table 1: Comparisons of our method with Jitter, MotionAug, ACTOR, MDM-M2M, and MDM-T2M. Jitter and PoseAugment are tested on all datasets, while other methods are tested on their own dataset. The performance of the basic datasets are shown in absolute errors, while the performance of augmented datasets are shown in relative improvements compared with using the basic datasets.

To quantitatively evaluate PoseAugment compared with previous data augmentation methods, we reproduced an IMU-based MoCap model TransPose [50] and trained it with datasets augmented by different methods. We use the performance of MoCap models to reflect the data augmentation quality.

Baseline Methods. We compare PoseAugment with 5 other data augmentation methods: (1)**Jitter** [50]: adding random noises $\sim \mathcal{N}(0, \sigma^2)$ directly to IMU data. (2)**MotionAug** [26]: a VAE/IK-based motion-to-motion augmentation method. (3)**ACTOR** [34]: a VAE-based action-to-motion generative model. (4)**MDM-T2M** [41]: a diffusion-based text-to-motion generative model. (5)**MDM-M2M** [41]: a modification of the MDM model by denoising noised GT motions, such that it can augment data in the motion-to-motion manner like ours.

To our best knowledge, though noise-based data augmentation methods have been widely adopted in classification tasks, they are rarely explored in IMU MoCap. Recent generative models have also been applied to pose generation but have not been evaluated in IMU-based MoCap either. Therefore, we selected Jitter, MotionAug, ACTOR, and MDM as the most representative methods to evaluate, covering the M2M, A2M, and T2M generation tasks. The algorithm implementation details are shown in Appendix A.3.

Datasets. Since MotionAug, ACTOR, and MDM are designed for different tasks (M2M, A2M, and T2M), which need GT motions, action labels, and text descriptions respectively, we followed the original papers and used them on different datasets (HDM05 [26], HumanAct12 [12], and HumanML3D [11]) respectively. We also added MDM-M2M as a variant of MDM, since the diffusion model in MDM can also generate motions in M2M manner.

To build the training datasets for the MoCap model, we first used these methods to generate 1× of their corresponding datasets, denoted as the *basic datasets*. They are treated as the unaugmented datasets in the following experiments. Then, up to 4× of motions are further generated using these methods, denoted as the *augmented datasets*. For Jitter and PoseAugment, we directly applied them on the basic datasets to augment data. We designed this setting for the reason that, the data distribution of baseline generative models are largely different from their original training datasets. We aim to compare the data quality of the generated data only, so the original data are discarded.

To simulate a common practice of using different amount of augmented data, we use the basic datasets(1×), together with 1× to 4× of the augmented data, to evaluate Jitter, MotionAug, ACTOR, MDM-M2M, MDM-T2M, and PoseAugment respectively. To be consistent with previous practices [16, 49, 50], we evaluate all MoCap models on the real IMU data of the DIP test split. The training details are described in Appendix A.4.

Metrics. We evaluate MoCap models on: (1)**SIP Error** e_{SIP}: the rotation error of upper arms and upper legs. (2)**Rotation Error** e_{rot}: the rotation error of all body joints. (3)**Position Error** e_{pos}: the position error of all body joints. (4)**Mesh Error** e_{mesh}: the position error of all vertices on the SMPL mesh. They are the same as the metrics used in [49, 50].

Results. Tab. 1 lists the data augmentation performance of PoseAugment compared with all baselines, tested by training the TransPose model. The data augmentation performance is measured by the relative error reduction compared to using the basic datasets only. We found that Jitter has almost no impact on the model performance, while ACTOR, MDM-M2M, and MDM-T2M are slightly better than Jitter, reducing the motion capture errors marginally. PoseAugment outperformed the above three methods in all metrics, showing a huge potential in improving the data quality for training MoCap models. MotionAug reveals a comparable performance with PoseAugment. However, as mentioned in Sec. 3.1, MotionAug needs users to annotate IK keyframes, and is limited to 8 motion types, while PoseAugment does not have these limitations.

Notably, we further investigated how many training data are needed to achieve the same level of model performance. We trained TransPose on all the original AMASS [27] data (45.61h), and got $e_{rot} = 10.2°$ and $e_{pos} = 6.4cm$. Using only the CMU dataset (a subdataset of AMASS, 9.06h) augmented with our method, we achieved a comparable performance with $e_{rot} = 10.2°$ and $e_{pos} = 6.5cm$, but only using 19.9% of training data. It indicates a huge potential of PoseAugment to reduce the data collection burden in practice.

Method	$e_{rot}(°)$			$e_{pos}(cm)$		
	MVAE+opt.	Ours-opt.	Ours	MVAE+opt.	Ours-opt.	Ours
MotionAug	**−13.6**%	−8.6%	−9.8%	**−16.7**%	−4.7%	−8.0%
ACTOR	−20.3%	**−21.7**%	−20.7%	−1.8%	**−11.4**%	−10.7%
MDM-M2M	−8.7%	−8.0%	**−9.1**%	−7.2%	−7.0%	**−7.8**%
MDM-T2M	−7.3%	−8.1%	**−8.3**%	−5.3%	−5.6%	**−5.8**%

Table 2: The ablation study on VAE model and physical optimization. Method indicates the datasets generated by the four techniques. Values are shown in the relative errors compared with using the basic datasets, the same as in Tab. 1.

Ablation Study. We further performed a two-part ablation study on the VAE model and the physical optimization. For the VAE part, our model has several network improvements over MVAE [23] as introduced in Sec. 3.1. We reproduced the original MVAE network, but kept the same motion frame representaion and inferencing techniques, and the same physical optimization as our method, denoted as *MVAE+opt.* in Tab. 2. For the physical optimization part, we removed it from PoseAugment, denoted as *Ours-opt.* in Tab. 2. As a result, PoseAugment outperformed MVAE+opt. in ACTOR, MDM-M2M, and MDM-T2M datasets, and outperformed Ours-opt. in MotionAug, MDM-M2M, and MDM-T2M datasets.

5 Conclusion and Limitations

We propose PoseAugment, a novel human pose data augmentation method that incorporates VAE-based pose generation and physical optimization. Experiments demonstrate a significant improvement of PoseAugment over previous pose augmentation methods, revealing a strong potential of our method to alleviate the data collection burden in human pose-related tasks.

However, our method requires more computational cost compared with traditional methods. Besides, the data augmentation performance is generally higher on smaller datasets. It will benefit tasks with fewer available data or tasks that involve personalization and few-shot learning the most.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant No. 62132010, Beijing Key Lab of Networked Multimedia, Institute for Artificial Intelligence, Tsinghua University (THUAI), Beijing National Research Center for Information Science and Technology (BNRist), 2025 Key Technological Innovation Program of Ningbo City under Grant No.2022Z080, Beijing Municipal Science and Technology Commission, Administrative Commission of Zhongguancun Science Park No.Z221100006722018, and Science and Technology Innovation Key R&D Program of Chongqing.

References

1. Bambade, A., El-Kazdadi, S., Taylor, A., Carpentier, J.: Prox-qp: Yet another quadratic programming solver for robotics and beyond. In: RSS 2022-Robotics: Science and Systems (2022)
2. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4221–4231 (2023)
3. Chen, W.H., Cho, P.C.: A gan-based data augmentation approach for sensor-based human activity recognition. Int'l J. Comp. and Comm. Engr **10**(4), 75–84 (2021)
4. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
5. Das, S., Trutoiu, L., Murai, A., Alcindor, D., Oh, M., De la Torre, F., Hodgins, J.: Quantitative measurement of motor symptoms in parkinson's disease: A study with full-body motion capture data. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 6789–6792. IEEE (2011)
6. Du, X., Vasudevan, R., Johnson-Roberson, M.: Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction. IEEE Robotics and Automation Letters **4**(2), 1501–1508 (2019). `https://doi.org/10.1109/LRA.2019.2895266`
7. 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. pp. 481–490 (2023)
8. Featherstone, R.: Rigid body dynamics algorithms. Springer (2014)
9. Felis, M.L.: Rbdl: an efficient rigid-body dynamics library using recursive algorithms. Autonomous Robots pp. 1–17 (2016). `https://doi.org/10.1007/s10514-016-9574-0`, `http://dx.doi.org/10.1007/s10514-016-9574-0`
10. Gong, K., Zhang, J., Feng, J.: Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8575–8584 (2021)
11. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)

12. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
13. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Trans. Graph. **38**(2) (mar 2019). `https://doi.org/10.1145/3311970`, `https://doi.org/10.1145/3311970`
14. Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Trans. Graph. **39**(6) (nov 2020). `https://doi.org/10.1145/3414685.3417836`, `https://doi.org/10.1145/3414685.3417836`
15. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International conference on learning representations (2016)
16. 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 Trans. Graph. **37**(6) (dec 2018). `https://doi.org/10.1145/3272127.3275108`, `https://doi.org/10.1145/3272127.3275108`
17. Inc., M.: Xsens (2024), `https://www.movella.com/products/xsens`
18. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. PLOS ONE **16**(7), 1–32 (07 2021). `https://doi.org/10.1371/journal.pone.0254841`, `https://doi.org/10.1371/journal.pone.0254841`
19. 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: European conference on computer vision. pp. 443–460. Springer (2022)
20. 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. SA '22, Association for Computing Machinery, New York, NY, USA (2022). `https://doi.org/10.1145/3550469.3555428`, `https://doi.org/10.1145/3550469.3555428`
21. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2151–2162 (October 2023)
22. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3d motion and forces of person-object interactions from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
23. Ling, H.Y., Zinno, F., Cheng, G., Van De Panne, M.: Character controllers using motion vaes. ACM Trans. Graph. **39**(4) (aug 2020). `https://doi.org/10.1145/3386569.3392422`, `https://doi.org/10.1145/3386569.3392422`
24. Liu, L., Yin, K., van de Panne, M., Shao, T., Xu, W.: Sampling-based contact-rich motion control. In: ACM SIGGRAPH 2010 Papers. SIGGRAPH '10, Association for Computing Machinery, New York, NY, USA (2010). `https://doi.org/10.1145/1833349.1778865`, `https://doi.org/10.1145/1833349.1778865`
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. **34**(6) (oct 2015). `https://doi.org/10.1145/2816795.2818013`, `https://doi.org/10.1145/2816795.2818013`

26. Maeda, T., Ukita, N.: Motionaug: Augmentation with physical correction for human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6427–6436 (June 2022)
27. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5441–5450 (Oct 2019). https://doi.org/10.1109/ICCV.2019.00554
28. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review **42**, 275–293 (2014)
29. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Trans. Graph. **36**(4) (jul 2017). https://doi.org/10.1145/3072959.3073596, https://doi.org/10.1145/3072959.3073596
30. 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. CHI '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581392, https://doi.org/10.1145/3544548.3581392
31. NaturalPoint, I.: Optitrack (2023), https://optitrack.com
32. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
33. Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. ACM Trans. Graph. **41**(4) (jul 2022). https://doi.org/10.1145/3528223.3530110, https://doi.org/10.1145/3528223.3530110
34. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10985–10995 (October 2021)
35. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video (2020)
36. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. Advances in neural information processing systems **29** (2016)
37. Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., Chen, B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Trans. Graph. **40**(1) (sep 2020). https://doi.org/10.1145/3407659, https://doi.org/10.1145/3407659
38. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: Physcap: Physically plausible monocular 3d motion capture in real time. ACM Trans. Graph. **39**(6) (nov 2020). https://doi.org/10.1145/3414685.3417877, https://doi.org/10.1145/3414685.3417877
39. Supej, M.: 3d measurements of alpine skiing with an inertial sensor motion capture suit and gnss rtk system. Journal of sports sciences **28**(7), 759–769 (2010)
40. Tessler, C., Kasten, Y., Guo, Y., Mannor, S., Chechik, G., Peng, X.B.: Calm: Conditional adversarial latent models for directable virtual characters. In: ACM SIGGRAPH 2023 Conference Proceedings. SIGGRAPH '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3588432.3591541, https://doi.org/10.1145/3588432.3591541

41. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
42. UK, V.M.S.L.: Vicon (2023), https://www.vicon.com
43. 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)
44. Wei, X., Chai, J.: Videomocap: Modeling physically realistic human motion from monocular video sequences. In: ACM SIGGRAPH 2010 Papers. SIGGRAPH '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1833349.1778779, https://doi.org/10.1145/1833349.1778779
45. Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., Xu, H.: Time series data augmentation for deep learning: A survey. pp. 4653–4660 (08 2021). https://doi.org/10.24963/ijcai.2021/631
46. Won, J., Gopinath, D., Hodgins, J.: Physics-based character controllers using conditional vaes. ACM Trans. Graph. **41**(4) (jul 2022). https://doi.org/10.1145/3528223.3530067, https://doi.org/10.1145/3528223.3530067
47. Wouwe, T.V., Lee, S., Falisse, A., Delp, S., Liu, C.K.: Diffusion inertial poser: Human motion reconstruction from arbitrary sparse imu configurations (2023)
48. Xu, X., Gong, J., Brum, C., Liang, L., Suh, B., Gupta, S.K., Agarwal, Y., Lindsey, L., Kang, R., Shahsavari, B., Nguyen, T., Nieto, H., Hudson, S.E., Maalouf, C., Mousavi, J.S., Laput, G.: Enabling hand gesture customization on wrist-worn devices. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3491102.3501904, https://doi.org/10.1145/3491102.3501904
49. 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 (CVPR). pp. 13167–13178 (June 2022)
50. Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. ACM Trans. Graph. **40**(4) (jul 2021). https://doi.org/10.1145/3450626.3459786, https://doi.org/10.1145/3450626.3459786
51. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16010–16021 (October 2023)
52. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2148–2157 (June 2018). https://doi.org/10.1109/CVPR.2018.00229
53. Zell, P., Wandt, B., Rosenhahn, B.: Joint 3d human motion capture and physical analysis from monocular videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
54. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
55. Zhao, L., Song, S., Wang, P., Wang, C., Wang, J., Guo, M.: A mlp-mixer and mixture of expert model for remaining useful life prediction of lithium-ion batteries. Frontiers of Computer Science **18**(5), 185329 (2024)

56. Zheng, Y., Yamane, K.: Human motion tracking control with strict contact force constraints for floating-base humanoid robots. In: 2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids). pp. 34–41 (Oct 2013). https://doi.org/10.1109/HUMANOIDS.2013.7029952
57. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5738–5746 (2019). https://doi.org/10.1109/CVPR.2019.00589
58. Zou, Y., Yang, J., Ceylan, D., Zhang, J., Perazzi, F., Huang, J.B.: Reducing footskate in human motion reconstruction with ground contact constraints. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (March 2020)