3D Human Pose Estimation with Muscles

Kevin Zhu **AliAsghar MohammadiNasrabadi** **Alexander Wong** **John McPhee**
University of Waterloo
{k79zhu, aa27moha, a28wong, mcphee}@uwaterloo.ca

Abstract

We introduce MusclePose as an end-to-end learnable physics-infused 3D human pose estimator that incorporates muscle-dynamics modeling to infer human dynamics from monocular video. Current physics pose estimators aim to predict physically plausible poses by enforcing the underlying dynamics equations that govern motion. Since this is an underconstrained problem without force-annotated data, methods often estimate kinetics with external physics optimizers that may not be compatible with existing learning frameworks, or are too slow for real-time inference. While more recent methods use a regression-based approach to overcome these issues, the estimated kinetics can be seen as auxiliary predictions, and may not be physically plausible. To this end, we build on existing regression-based approaches, and aim to improve the biofidelity of kinetic inference with a multihypothesis approach — by inferring joint torques via Lagrange's equations and via muscle dynamics modeling with muscle torque generators. Furthermore, MusclePose predicts detailed human anthropometrics based on values from biomechanics studies, in contrast to existing physics pose estimators that construct their human models with shape primitives. We show that MusclePose is competitive with existing 3D pose estimators in positional accuracy, while also able to infer plausible human kinetics and muscle signals consistent with values from biomechanics studies, without requiring an external physics engine.

1 Introduction

3D human pose estimation (HPE) is a fundamental task in computer vision that involves the localization of 3D human joints from images, which allows the user to track human movement from videos, leading to a plethora of potential downstream applications. However, since many pose estimators are purely data-driven, the inferred motion is modeled implicitly, which may lead to physically impossible poses and movements.

Physics-based human pose estimation (PHPE) methods aim to mitigate these artifacts by enforcing the underlying dynamics equations that govern the kinematic state $\mathcal{K} = \{\boldsymbol{q}, \dot{\boldsymbol{q}}, \ddot{\boldsymbol{q}}\}$,

$$\mathfrak{M}(\boldsymbol{q}, \mathcal{A}) \cdot \ddot{\boldsymbol{q}} + \mathfrak{C}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \mathcal{A}) = \boldsymbol{\tau}_q + \mathfrak{F} \tag{1}$$

where $\boldsymbol{q}$ are generalized coordinates that describe motion, often in terms of translational (*e.g.* 3D position of the root) and rotational (*e.g.* joint rotations) degrees of freedom (DoF). We denote "dot" ($\dot{}$) as the time derivative and "double dot" ($\ddot{}$) as the 2nd time derivative of a variable. $\mathfrak{M}$ is the mass matrix and $\mathfrak{C}$ contains the Coriolis, centrifugal, and gravitational forces, for a human with anthropometric features $\mathcal{A}$ at a given state $\mathcal{K}$. Here, we loosely lump together a human's dimensions, mass and inertia properties, and other intrinsic and mobility features using the anthropometrics term $\mathcal{A}$. On the right hand side, $\boldsymbol{\tau}_q$ describes the human joint torques generated by each DoF, and $\mathfrak{F}$ are external forces, both in the generalized space.

In this paper, we deal with monocular pose estimation, where our only input source is a monocular video, without force sensors. When only one unknown external force $\mathfrak{F}$ is applied on the human,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

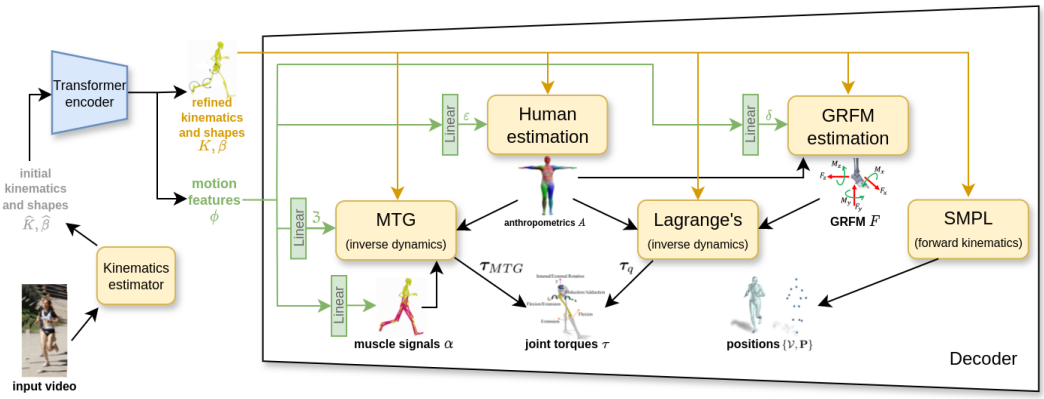


Figure 1: Overall framework of MusclePose.

we can solve for τ_q by enforcing the entries of τ_q that corresponds to the root link to be zero. However, when there are more than one external force applied to the human simultaneously at different locations, which is often the case (*e.g.* when both feet are in contact with the ground), Eq. (1) becomes underconstrained.

Without large-scale force-annotated video datasets, many methods estimate the corresponding kinetics via optimization with an external physics engine [20, 19, 77, 41]. However, these physics engines are often either non-differentiable and cannot be trained end-to-end, or are too slow for real-time inference. Furthermore, as discussed in [85], these methods are often combined with reinforcement learning to reach a desired outcome, but the effects of changing inputs on the outputs are unknown. Since joint torques can be hard to agree on in the biomechanics community, as they are often computed from different models, with different assumptions and post-processing, a more flexible learning framework may be preferred. More recently, PHPE methods have began regressing kinetics directly with neural networks [85, 37, 63]. While the regression-based approach improves the kinematic reliability of the predicted motion, the inferred kinetics can be seen as auxiliary predictions, which may not be directly constrained and may be physically implausible. Although these kinetic predictions are not the main focus of these pose estimators, they may still be of interest for downstream applications. In sports for example, in addition to kinematics, practitioners and researchers are often interested in analyzing the whole-body musculoskeletal dynamics of athletes. To do so, a multibody model of skeletal dynamics is commonly used in combination with an optimal control algorithm to generate predictive simulations of athlete movements [5, 28, 49]. However, these optimal control algorithms can take hours or days to produce results.

To this end, we build on existing regression-based PHPE approaches, to infer human kinetics simultaneously with kinematics, without a physics engine, and propose MusclePose (Fig. 1) to improve the plausibility of the predicted kinetics. To mitigate the underconstrained problem of regressing kinetics, we use a multihypothesis approach, and compute torques via Lagrange's equations, and also via muscle dynamics modeling with muscle torque generators (MTGs) [51, 25].

To maintain fidelity when modeling human movement, classical muscle models often represent muscles as linear actuators, and capture the nonlinear dependence of muscle tension on muscle length and the rate of lengthening [66, 32] using various Hill-type muscle models [23]. However, incorporating detailed muscles requires solving the actuator redundancy problem [3] and computing complex and varying musculoskeletal geometries [60, 12]. To overcome these drawbacks, parametric MTG models were proposed to mimic the behavior of muscles crossing a given joint to directly approximate joint torque by modeling kinematic dependence on active torque generation and passive impedance (Eq. (11)). Essentially, MTGs infer net joint torques from a joint's kinematics and activation levels, which is what we ultimately want, as we are not interested in isolated muscle tensions or granular joint contact forces. And since MTGs consist of differentiable equations, we are able to incorporate them into our learning framework, and train our pose estimator end-to-end.

Moreover, for computational efficiency, existing PHPE methods rely on human models with anthropometrics estimated from the predicted human dimensions, or use the intrinsics properties (*e.g.* inertia

and mass properties) of primitive shapes (*e.g.* spheres and simple rods), as proxies. From, Eq. (1), we see that, even if the kinematics state $\mathcal{K}$ and external forces $\mathfrak{F}$ are accurate, but $\mathcal{A}$ is not, the inferred torques $\boldsymbol{\tau}_q$ may not correspond to the actual human performing the motion. For example, existing pose estimators may infer the center of mass (CoM) of body parts by taking the mean of the predicted surface mesh, assuming constant density [85]. However, since the composition of bones, muscles, internal organs, etc. is different, the human body's density is not uniform [14]. For example, the CoM of the upper torso is slightly towards the left side [16], whereas taking the mean vertices will be in the center. As such, we further predict detailed anthropometrics for each human, and keep them close to values taken from biomechanics studies.

In summary, we introduce MusclePose to comprehensively predict human kinematics, kinetics, muscle signals, and detailed anthropometrics from monocular video. Specifically, we want a pose estimator with (i) a flexible learning framework easily adaptable for different scenarios, (ii) a reasonable degree of biofidelity, (iii) inference speed and (iv) positional accuracy both on par with purely kinematic pose estimators. To satisfy (i) and (iii), MusclePose is regression-based, consists of customizable and swappable components, can be trained end-to-end, and does not require an external physics engine. For (ii), MusclePose is the first pose estimator to incorporate muscle dynamics modeling and predict detailed human anthropometrics. We demonstrate improvements in the inferred kinetics on actions including *walking* from the H36M dataset [27] and *baseball pitching* and *golf swings* from PennAction [84]. Also, the use of MTGs allows us to further assess human motion at a musculoskeletal level, and we show that our inferred muscle signals are comparative to those from biomechanics studies, as well as to EMG data of pertinent muscle groups. Lastly, for (iv), we evaluate our method on benchmark 3D HPE datasets, H36M [27] and 3DPWoc, [71], to show that MusclePose is kinematically competitive with state-of-the-art (SOTA) pose estimators.

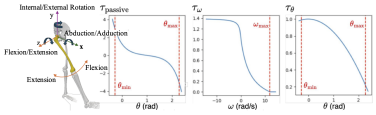


Figure 2: Examples of MTG curves for hip flexion. $\tau_{passive}$ models the passive torque [79] as a double exponential function. τ_ω models the active-torque–angular-speed relationship [69, 67] as a piecewise function. τ_θ models the active-torque-angle relationship [21, 33] as the non-negative portion of a polynomial.

2 Related work

Monocular 3D human pose estimation. Early deep learning 3D HPE approaches use convolutional neural networks to directly estimate human 3D keypoint positions from images, with intermediate values represented by 3D heatmaps [57], location maps [50], or 2D heatmaps with depth regression [87]. The more recent and popular approach lifts 2D keypoints to 3D, essentially forming a monocular sparse-depth estimation task. The lifting network can be fully-connected layers [47], temporal convolution networks [10, 58], graph convolution networks [74, 7, 11], or transformers [89, 39, 86]. Human pose and shape estimation (HPSE) refers to predicting a 3D surface mesh of humans. The popular model-based HPSE approach [38, 35, 31] predicts input parameters of a parametric human model, such as SMPL, which infers a 3D mesh from rotation and shape parameters. Non-parametric approaches [42, 52, 36] directly regress 3D coordinates of mesh vertices. Other methods combine both approaches, such as [70], which predicts a volumetric representation before fitting a SMPL model, or [45], which calibrates model-based mesh predictions with 3D keypoints.

Physics based pose estimation. To achieve more physically plausible human motion, PHPE methods apply dynamic constraints to encourage contact and penalize motion jitter, ground penetration, and unbalanced postures. [77, 63, 61, 64] model contact forces between the foot and ground, [20, 19, 82] include contact points between the full body and ground, while [41] also models human interaction with stick-like hand tools. Optimization-based frameworks [20, 19, 77, 64, 41] simulate physically plausible human motion from a physics engine and minimize an objective function to keep the simulated motion close to the detections obtained from a kinematic pose estimator. These frameworks are also combined with reinforcement learning [82, 59]. Recently, to overcome the need of an external physics engine, regression-based frameworks [85, 37, 63] directly estimate human kinetics using neural networks. As discussed in Sec. 1, this is an underconstrained problem, for which we hope to mitigate, and further inject biofidelity. We incorporate MTGs to do so.

Muscle torque generators (MTGs). Due to their simplicity, MTGs have been increasingly popular in multibody dynamics simulations as they reduce computational cost while maintaining a reasonable degree of biofidelity. Recently, MTGs have been incorporated to simulate human movement post hip and knee replacement surgeries [13], human interactions with exoskeletons [22, 26], and manual wheelchair propulsion [5]. For more dynamic movements, such as in sports, examples of MTG-driven simulations of athlete motor control include golf [49] and cycling [28].

3 MusclePose

We propose MusclePose (Fig. 1) as a physics-based pose estimator to directly regress comprehensive human dynamics from a monocular video of length T. We use a transformer encoder to refine initial pose estimates, and produce latent motion features $\phi^{\{1:T\}}$, which are used as inputs for 5 customizable modules to infer human anthropometrics $\mathcal{A}$, kinematics $\mathcal{K}$, external forces $\mathcal{F}$, joint torques via Lagrange's equations $\boldsymbol{\tau}_q$, and joint torques via MTGs $\boldsymbol{\tau}_{MTG}$. We describe our prototype in the following subsections, where we extract muscle signals $\boldsymbol{\alpha}^{\{1:T\}}$ and residual terms $\mathcal{E}$, $\delta^{\{1:T\}}$, $\mathfrak{Z}$ from $\phi^{\{1:T\}}$ as inputs for the 5 modules. All variables described in this section are sequences of length T, except $\mathcal{E}, \mathfrak{Z}$ and shape parameters $\boldsymbol{\beta}$, and we drop the superscript $(^{\{1:T\}})$.

Compared to the most recent regression-based PHPE method, PhysPT [85], we do not use a transformer decoder to directly regress joint torques, and instead, regress the input parameters of MTG models. Our motivation stems from the popular HPSE approach that regresses SMPL parameters instead of the human mesh directly, which not only reduces the computation complexity but also geometrically constrains the predicted human, as SMPL infers the mesh via forward kinematics. In parallel, we use MTGs to avoid estimating complex musculoskeletal geometries and granular joint contact forces, while enforcing a constraint on the inferred torques from Lagrange's equations.

3.1 Kinematics estimation

We follow the common approach in PHPE and clean initial kinematic estimates $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{T}}\}$ generated by some existing kinematic pose estimator, to obtain the refined $\{\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}, \mathbf{c}\}$ as our prediction. Here, $\mathbf{T} \in \mathbb{R}^3$ represents 3D pelvis **translation** in the world frame, and $\mathbf{c}$ are binary **contact** labels. **Rotation** parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_{23}\}$ represents local rotations of the 24 SMPL keypoints, relative to their parents in the SMPL kinematic tree, with $\boldsymbol{\theta}_0$ being the pelvis orientation in the world frame. We follow prior work [34] and predict the 6D continuous rotation representation [88] for each $\boldsymbol{\theta}_k \in \mathbb{R}^6$. **Shape** parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ denotes the first 10 principal components of SMPL's shape space. Since our inputs lack the shape information that RGB images provide, we follow the hybrid approach in [89] to regress shape residuals that are combined with initial predictions. We also use the same approach and regress anthropometric residuals $\mathcal{E}$ later on in Sec. 3.2 and force residuals δ in Sec. 3.3.

The parametric human model, SMPL, then uses a collection of linear functions to map these parameters to a triangulated mesh $\mathcal{V}$ of 6890 vertices that represents the surface of the human body, and 24 SMPL keypoint positions $\mathbf{P}$:

$$\{\mathcal{V}, \mathbf{P}\} = \text{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta}) + \mathbf{T} \tag{2}$$

We define the kinematic loss $\mathcal{L}_{kin}$ with weights $\boldsymbol{\lambda}_{kin}$ as

$$\mathcal{L}_{kin} = \boldsymbol{\lambda}_{kin} \cdot [\mathcal{L}_p \quad \mathcal{L}_v \quad \mathcal{L}_\theta \quad \mathcal{L}_\beta \quad \mathcal{L}_{norm} \quad \mathcal{L}_c]^\intercal \tag{3}$$

where the first five losses are from [89] which penalize joint position, linear velocity, SMPL parameter prediction L1 errors, and minimize the L2 norms of the SMPL parameters; and $\mathcal{L}_c$ is the binary contact loss from [83].

To facilitate multibody dynamics modeling in the following sections, we convert the predicted coordinates to **generalized coordinates**, $\boldsymbol{q} = [\boldsymbol{X}_0, \boldsymbol{q}_0, \boldsymbol{q}_1, ..., \boldsymbol{q}_{N_k}]^\intercal \in \mathbb{R}^{N_{DoF}}$, where $\boldsymbol{X}_0 \in \mathbb{R}^3$ is the global root translation, and each $\boldsymbol{q}_k$ describes the joint's rotational DoFs. Specifically, each $q_i \in \boldsymbol{q}_k$ are ZXY euler angles converted from the predicted $\boldsymbol{\theta}_k$, to match the International Society of Biomechanics (ISB) format, where a joint's local z-direction corresponds to flexion/extension, x for abduction/adduction and y for internal/external rotation. We denote the predicted kinematics as $\mathcal{K} = \{\boldsymbol{q}, \dot{\boldsymbol{q}}, \ddot{\boldsymbol{q}}\}$, with the velocity and acceleration terms estimated via finite differences.

3.2 Human estimation

Human model. We assume a rigid multibody dynamics model of a human with $N_k = 18$ segments and $N_{DoF} = 47$ total degrees of freedom (DoF). The 3D positions of the 18 joints, each corresponding to a segment, are the 24 SMPL keypoint positions minus the 5 end-effectors and the *spine3* SMPL keypoint. The wrists, elbows, scapulas, each contain 2 rotational DoFs, knees each with 1 rotational DoF, root with 3 rotational and 3 translational DoFs, and 3 rotational DoFs for each of the remaining joints, for a total of 47 DoFs. We selected this configuration as it aligns best with biomechanics studies with anthropometric measurements that we use in the remaining sections.

Anthropometrics prediction. To predict the human's anthropometrics $\mathcal{A} = \cup_k \{m_k, I_{0,k}, CoM_k\}$, specifically the **mass** m_k, **inertia tensor** at zero rotation $I_{0,k}$, and **CoM** of all segments, we scale literature values $\bar{\mathcal{A}}$ from [16] based on the predicted human shapes β and add the predicted offsets $\mathcal{E}$,

$$\mathcal{A} = s_\beta \bar{\mathcal{A}} + \mathcal{E} \tag{4}$$

The scaling term s_β is computed from the predicted β, with details in the supplementary material.

3.3 Kinetics estimation

Ground reaction forces and moments (GRFM) prediction. Let $\mathcal{F}_k = [\mathbf{F}_k, \mathbf{M}_k]^\intercal$ be the GRFM applied on the CoM of each segment k. We infer $\mathcal{F} = \sum_k \mathcal{F}_k$ from our previous predictions and our regressed force residuals δ,

$$\mathcal{F} = \text{GRFM model}(\mathcal{K}, \mathcal{A}, \delta) \tag{5}$$

Since we trained our model on the AMASS dataset [46] and feet-ground contact labels from RoHM [83], we assume feet-ground contact only for simplicity, as with many PHPE methods [37, 20, 82]. Omitting subscript k, let $\mathbf{F} = [F_X, F_Y, F_Z]^\intercal$ be the force in world cartesian coordinates where Y is the vertical direction, and let $\mathbf{z} = [z_x, z_y, z_z]^\intercal$ be the center of pressure (CoP) in the foot's local coordinates where x is along the length of the foot (*i.e.* $\mathbf{M} = R^0_{ankle}\mathbf{z} \times \mathbf{F}$ where R^0_{ankle} is the ankle's world orientation). From the regressed residuals $\delta_{\{Y,l\}} \subset \delta$ and the kinematics of each foot $\mathcal{K}_{foot}$, we estimate the vertical force applied on the foot scaled by bodyweight $F^W_Y = F_Y/W$, and the CoP along the foot scaled by foot length $z^l_x = z_x/l_{foot}$,

$$\{F^W_Y, z^l_x\} = \boldsymbol{\eta}\mathcal{K}_{foot} + \delta_{\{Y,l\}} \tag{6}$$

where linear coefficients $\boldsymbol{\eta}$ were fitted on the forceplate data in [72]. The remaining δ terms are scaling factors between -1 and 1 to ensure the values in the other directions are physically possible (*i.e.* $F^2_X + F^2_Z \leq \delta^2_\mu F^2_Y$ and $\boldsymbol{z}$ is within the foot's dimensions)

$$F_X = \delta_X \delta_\mu F_Y, \qquad F_Z = \delta_Z \sqrt{\delta^2_\mu F^2_Y - F^2_X} \tag{7}$$

$$z_y = -|\delta_h l_h|, \qquad z_z = \delta_s (l_w/2) \tag{8}$$

where l_w, l_h are the foot's width and height, respectively. Additional details of our GRFM model can be found in the supplementary material.

Inverse dynamics via Lagrange's. From here, we can analytically compute the mass matrix $\mathfrak{M}$, Coriolis term $\mathfrak{C}$, external forces in the generalized space $\mathfrak{F}$, and infer joint torques in the generalized space $\boldsymbol{\tau}_q$ from the equations of motion:

$$\boldsymbol{\tau}_q = \text{Lagrange's}(\mathcal{K}, \mathcal{A}, \mathcal{F}) = \mathfrak{M}\ddot{\boldsymbol{q}} + \mathfrak{C} - \mathfrak{F} \tag{9}$$

We include the calculations of these terms in the supplementary material. We define a residual force loss $\mathcal{L}_{res}$ to minimize the resulting forces and torques at the root, which correspond to the first 6 entries of $\boldsymbol{\tau}_q$

$$\mathcal{L}_{res} = |\boldsymbol{\tau}_{q_{[:6]}}| \tag{10}$$

Inverse dynamics via MTGs. Simultaneously, we use parametric MTG models [51] to infer joint torques $\boldsymbol{\tau}_{MTG}$ from the predicted kinematics $\mathcal{K}$ and muscle activations $\boldsymbol{\alpha}$. Specifically, this kinematic dependence is separated into active torque generation τ_{active} and passive impedance $\tau_{passive}$. For each joint rotational DoF $q \in \boldsymbol{q}_{[6:]}$ with angular velocity $\dot{q}$, let muscle signal $\alpha \in [0, 1]$ represent the joint's corresponding activation level for this DoF, we compute the corresponding torque as

$$\tau_{MTG} = \text{MTG}(\mathcal{K}, \mathcal{A}, \alpha, \mathfrak{Z}) = \tau_{active} + \tau_{passive} \tag{11}$$

The **active torque** is further broken down into

$$\tau_{active} = \alpha \cdot \tau_{\omega}(\dot{q}) \cdot \tau_{\theta}(q) \cdot \tau_0(\mathcal{A}, \mathfrak{Z}) \tag{12}$$

where $\tau_{\omega}(\dot{q}; \boldsymbol{\gamma}_{\omega})$ models the active-torque–angular-speed relationship [69, 67] and $\tau_{\theta}(q; \boldsymbol{\gamma}_{\theta})$ models the active-torque-angle relationship [21, 33], as shown in Fig. 2. These relationships are parameterized by the $\boldsymbol{\gamma}$ coefficients, which are unique for each joint's DoF and direction, and are identified via dynamometry. This joint-dependent parameterization preserves physiological realism (*e.g.*, hip flexion and knee extension should exhibit different peak torque and passive stiffness profiles), unlike uniform torque models that assume identical properties across the body. For this paper, we use the set of $\boldsymbol{\gamma}$ values summarized in [54, 53].

$\tau_0(\mathcal{A}; \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_e)$ is the peak isokinetic torque that controls peak MTG output at zero joint velocity, which can be measured with a dynamometer. τ_0 is estimated in [54, 53] as a linear approximation of the human's intrinsic, scaled by certain external factors such as the human's fitness or activity level. Since these external factors (and some intrinsic properties) are not readily known, we take the mean effects $\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_e$ from [54, 53], and add regressed offsets $\mathfrak{Z}$. We compute τ_0 as

$$\tau_0 = (\boldsymbol{\gamma}_i \mathcal{A} + \mathfrak{Z}_i)(\boldsymbol{\gamma}_e + \mathfrak{Z}_e) \tag{13}$$

Furthermore, to account for stability, we assume each joint is driven by a pair of agonist-antagonist MTGs — a flexor (+) and an extensor (-), that corresponds to the movement direction. Hence, for each joint rotational DoF, we regress 2 muscle signals $\{\alpha^{flex}, \alpha^{ext}\}$, and the active torque becomes:

$$\tau_{active} = \alpha^{flex} \tau_{\omega}^{flex} \tau_{\theta}^{flex} \tau_0^{flex} + \alpha^{ext} \tau_{\omega}^{ext} \tau_{\theta}^{ext} \tau_0^{ext} \tag{14}$$

$\tau_{passive}(q; \boldsymbol{\gamma}_p)$ is the **passive torque** [1] of a joint that arises when the surrounding muscles, tendons, and ligaments are strained and intensifies near anatomical joint limits [1, 81]. The joint's viscous damping and nonlinear stiffness are parameterized by $\boldsymbol{\gamma}_p$, which encourages the joint to move within its range of motion, as a large restoring torque is produced otherwise, as shown in Fig. 2. Equations to compute $\tau_{\omega}, \tau_{\theta}, \tau_{passive}$ can be found in the supplementary material.

We define the torque loss $\mathcal{L}_{\tau}$ as the absolute difference between the two sets of predicted joint torques, and another regularizing term $\mathcal{L}_{\epsilon}$ for all regressed residuals:

$$\mathcal{L}_{\tau} = |\boldsymbol{\tau}_{q[6:]} - \boldsymbol{\tau}_{MTG}| \tag{15}$$

$$\mathcal{L}_{\epsilon} = ||\mathcal{E}||_2 + ||\delta||_2 + ||\mathfrak{Z}||_2 \tag{16}$$

Finally, we have dynamic loss with weights $\boldsymbol{\lambda}_{dyn}$

$$\mathcal{L}_{dyn} = \boldsymbol{\lambda}_{dyn} \cdot [\mathcal{L}_{\tau} \quad \mathcal{L}_{res} \quad \mathcal{L}_{\epsilon}]^{\intercal} \tag{17}$$

4 Experiments

4.1 Implementation and datasets

For training, we used the AMASS dataset [46], with feet-ground contact labels from [83]. As such, we trained and evaluated on sequences with feet-ground contact only (denoted †), which is also the case for many PHPE experiments [37, 20, 82]. We trained MusclePose end-to-end with a sequence input length of 16 frames, using total loss $\mathcal{L}_{total} = \mathcal{L}_{kin} + \mathcal{L}_{dyn}$ for 25 epochs, using the AdamW optimizer [44] with a weight decay of 10^{-4} and an initial learning rate of 10^{-4} that decreases by 20% every 5 epochs. Following common curriculum learning [2] practices, we split the training into two phases — for the first 20 epochs, we trained using the ground truth as input, followed by 5 epochs using the model's predictions as inputs.

For evaluation, we assessed positional accuracy on the inference results of the H36M test set [27] and object-occlusion subset of 3DPW (3DPWoc) [71]. As with training, we removed input sequences containing non feet-ground contact, the *sitting* and *sitting down* actions in H36M, and *courtyard laceshoe, flat guitar, outdoors climbing, outdoors freestyle, outdoors parcours, downtown stairs* in 3DPWoc. We further assessed kinetic biofidelity from 3 actions — *walking* from H36M, and *baseball pitch* and *golf swing* from the PennAction dataset (PA) [84]. As with existing large-scale human video datasets, since neither datasets include force-annotations, we compared our inference results with existing biomechanics studies of these movements, and commented on overall trends and plausibility. We selected *walking* because human gait is heavily studied in biomechanics [72, 18, 76, 8, 29, 75], and is a relatively consistent and cyclic movement. We included the latter two actions to evaluate faster movements, for which we were able to find published lab measurements [55, 80, 62]. During inference, to promote a closer comparison with the SOTA regression-based physics pose estimator PhysPT [85], we used the same kinematic estimator, CLIFF [40], to extract initial kinematic estimates, and the global trajectory predictor in [85] to extract initial root DoFs. The rationale for using CLIFF in [85] is that it produces competitive positional accuracy but lacks in physical plausibility.

4.2 Positional accuracy

We followed standard evaluation protocol and reported the mean per-joint positional error (MJE) and procrustes-aligned MJE (PJE) in Tab. 1, for the 14 LSP [30] keypoints in millimetres. MJE is the root-aligned mean Euclidean distance in millimeters between the predicted and ground truth 3D keypoints. PJE is the MJE after aligning the predicted pose with the ground truth in translation, rotation, and scale using the Procrustes method. For 3DPWoc, since the data is captured using a moving camera with unknown extrinsics, and our method predicts the global root DoFs directly, we reported PJE only, with additional ablations in Tab. 3 to show consistency of results.

We see that MusclePose outperforms other PHPE methods on H36M but is slightly worse than PhysPT on 3DPWoc. This, along with the overall worse positional accuracy of PHPE compared to purely kinematic methods, could be due to a kinematics-kinetics trade-off, as our method produced a lower residual force (Tab. 4). Specifically, the root DoFs in 3DPW may be harder to estimate due to the moving camera, leading to higher residual forces, which the model may try to reduce (lower kinetic error) by estimating a set of local joint kinematics slightly different from the original motion (higher kinematic error).

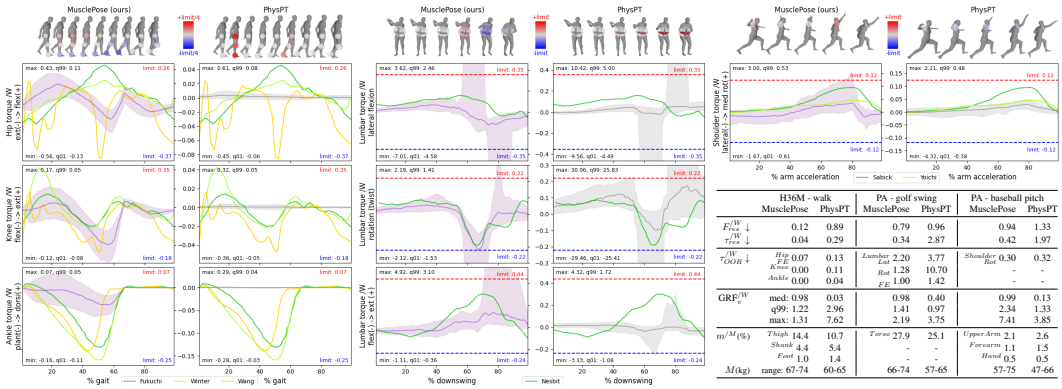
Table 1: Positional accuracy on H36M and 3DPWoc. “opti" denotes kinetics obtained from an external physics optimizer, and “regr" denotes regressed by a neural network. †denotes sequences with feet-ground contact only.

			H36M		3DPWoc
		Kinetics	MJE↓	PJE↓	PJE↓
kinematic	HybrIK [38]	-	55.4	33.6	-
	HybrIK [38]	-	†56.4	†36.7	-
	CLIFF [40]	-	52.2	36.8	-
	CLIFF [40]	-	†**46.5**	†**32.4**	†**24.0**
PHPE	SimPoE [82]	opti.	†56.7	†41.6	-
	DiffPhy [20]	opti.	†81.7	†55.6	-
	D&D [37]	regr.	†52.5	†35.5	-
	PhysPT [85]	regr.	†50.6	†35.5	†25.9
	MusclePose(ours)	regr.	†48.4	†33.5	†27.6

4.3 Biofidelity

Since most PHPE methods do not report kinetics results, we mainly compared ours with values we reproduced from PhysPT. For even comparison, for both pose estimators, we computed joint torques in the generalized space τ_q via Lagrange's equations (9) from the predicted motion, anthropometrics, and GRFM. Unlike the biomechanics studies, we did not apply additional signal post-processing or smoothing to τ_q. Hence, we see more noise and spikes in the pose estimators' results, which could also be amplified by the low frame rate of PennAction.

Qualitative. Since the different biomechanics studies computed torques differently from different datasets, we comment on general trends and evaluate qualitatively. In Fig. 3, we plotted median torques scaled by predicted body weight, with a 25-75% quantile band, for select joints of the 3 actions, for which we found reference values. Overall, compared to PhysPT (gray), we see that MusclePose (ours, purple) more closely follows the trends and magnitudes of the reference values (greens and yellow).



	H36M - walk		PA - golf swing		PA - baseball pitch	
	MusclePose	PhysPT	MusclePose	PhysPT	MusclePose	PhysPT
$F_{res}^{/W} \downarrow$	0.12	0.89	0.79	0.96	0.94	1.33
$\tau_{res}^{/W} \downarrow$	0.04	0.29	0.34	2.87	0.42	1.97
$\tau_{OOR}^{/W} \downarrow$	$_{Hip}$ 0.07	0.13	$_{Lumbar}$ 2.20	3.77	$_{Shoulder}$ 0.30	0.32
	$_{FE}$ 0.00	0.11	$_{Rot}$ 1.28	10.70	$_{Rot}$ -	-
	$_{Ankle}$ 0.00	0.04	$_{FE}$ 1.00	1.42	-	-
$GRF_v^{/W}$	med: 0.98	0.03	0.98	0.40	0.99	0.13
	q99: 1.22	2.96	1.41	0.97	2.34	1.33
	max: 1.31	7.62	2.19	3.75	7.41	3.85
$m^{/M}$(%)	$_{Thigh}$ 14.4	10.7	$_{Torso}$ 27.9	25.1	$_{UpperArm}$ 2.1	2.6
	$_{Shank}$ 4.4	5.4	-	-	$_{Forearm}$ 1.1	1.5
	$_{Foot}$ 1.0	1.4	-	-	$_{Hand}$ 0.5	0.5
M(kg)	range: 67-74	60-65	66-74	57-65	57-75	47-66

Figure 3: Median predicted joint torques scaled by bodyweight, with a 25-75% quantile band, for gait cycles, the downswing phase of golf drives, and the arm acceleration phase of baseball pitches, compared to values from biomechanics studies.

The 2 leftmost columns of Fig. 3 correspond to flexion/extension of the hip, knee, and ankle torques for walking, scaled such that the toe-off occurs at 60% of the gait cycle, and compared to reference torques from [73, 18, 75]. We see that MusclePose produced more reasonable trends overall, whereas PhysPT produced torques with low magnitudes but with higher extreme values. For hip flexion, our results resembled more of the yellow curve that was computed from a wearable system in [73], where the authors attributed their errors to a lack of shear force measurement, leading to more noticeable errors in the hips than more distal joints due to the increase in moment arms. Since the subjects in H36M walk in a small circle with slower and varying speeds, a discrepancy can arise in the generated shear force, leading to the discrepancy in hip torque, as the subjects in the biomechanics studies walk in a straight line at a consistent pace. This could also contribute to the low magnitude of ankle torque, where different timings (toe-off/heel-off/touch-down) could affect ankle power generation, as explained in [6].

The middle 2 columns of Fig. 3 correspond to lumbar torques during the downswing phase of golf drives, scaled such that the maximum lumbar rotation occurs at 2/3 of the motion, and compared to reference torques from [55]. While we see noticeable spikes from both pose estimators, the extreme values for lumbar lateral bending and axial rotation were much higher for PhysPT.

The 2 rightmost columns of Fig. 3 correspond to shoulder lateral/medial rotation torque during the arm acceleration phase of baseball pitches, scaled such that the maximum moment occurs at 80% of the motion, and compared to reference torques from professional pitchers (darker green) in [62], as well as amateurs (lighter green) in [80]. Although the peak of our 75% quantile slightly exceeds the shoulder medial rotation limit reported in [62], our band was able to cover values from both skill groups, whereas PhysPT's band was below the amateurs, even though the pitchers in PA range from teenage amateurs to adult professionals.

Quantitative. In the bottom right of Fig. 3, we reported the mean residual forces $\{F_{res}, \tau_{res}\}$, mean out of range joint torques τ_{OOR}, and the median value of the sum of GRFs in the direction opposite of gravity GRF_v. Residual F_{res} and τ_{res} were computed as the mean L2 norms of the entries of $\boldsymbol{\tau}_q$ that correspond to the translational and rotational DoFs of the root. τ_{OOR} was computed as the mean absolute amount outside of joint torque limits (red and blue values in Fig. 3) reported in [1, 43, 62]. These values are further scaled by predicted body weight and denoted $(\,^{/W})$. In the last column, we also reported the mean predicted segment mass $m^{/M}$ as a percentage of body mass M.

Overall, we see that MusclePose inferred more reasonable kinetics, indicated by the lower residual forces, less extreme joint torques, and a median GRF_v closer to body weight. Similar to its joint torques, PhysPT's GRF_v values were overall lower in magnitude, with occasional spikes. While MusclePose's maximum GRF_v were very large for the golf swing and baseball pitch, our 99% quantiles were more comparable to the maximum values of about 1.3 times body weight for golf reported in [55], and about 2.3 for pitching in [9].

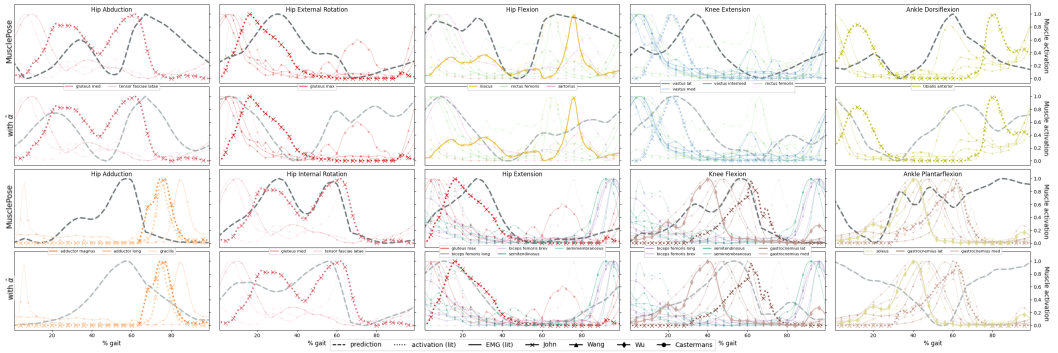


Figure 4: Mean predicted muscle activations (dashed) for gait cycles compared to EMG data and predictions of pertinent muscles from literature. Min-max scaling applied to all values.

4.4 Ablations and additional evaluation

We reported ablations results in Tab. 2. Row 2 includes results without custom anthropometrics, and instead computed directly from the SMPL mesh, assuming constant density. Row 3 includes results without MTGs. Row 4 includes results using kinematic-based muscle activations, with details in the next paragraph. We see that, overall, MusclePose has better positional accuracy, along with lower residual forces ($\mathbf{F}_{res} = (F_{res} + \tau_{res})/2$), and median GRF$_v$ closest to bodyweight.

Table 2: Ablations results.

	†H36M		†3DPWoc		H36M - Walk		PA - Golf swing		PA - Pitch	
	MJE↓	$\mathbf{F}_{res}^{/W}\downarrow$	PJE↓	$\mathbf{F}_{res}^{/W}\downarrow$	$\mathbf{F}_{res}^{/W}\downarrow$	med GRF$_v^{/W}$	$\mathbf{F}_{res}^{/W}\downarrow$	med GRF$_v^{/W}$	$\mathbf{F}_{res}^{/W}\downarrow$	med GRF$_v^{/W}$
MusclePose	**48.4**	**0.08**	27.6	**0.25**	**0.08**	**0.98**	**0.56**	**0.98**	0.68	**0.99**
w/o $\mathcal{A}$	49.5	0.26	27.9	0.35	0.22	0.74	0.61	0.64	**0.67**	0.39
w/o τ_{MTG}	49.7	0.14	**27.2**	0.30	0.15	0.94	0.66	0.88	0.74	0.90
with $\hat{\alpha}$	50.9	0.16	28.4	0.26	0.13	0.93	0.59	0.97	0.69	0.90

Furthermore, due to the lack of extrinsics information in 3DPWoc, we reported results from using different kinematic estimators in Tab. 3 to show consistency.

Table 3: †3DPWoc results with different kinematic estimators.

	Ours+CLIFF	CLIFF [40]	Ours+WHAM	WHAM [65]	Ours+CoMotion	CoMotion [56]
PJE↓	27.6	24.0	28.5	22.7	32.6	30.7
$\mathbf{F}_{res}^{/W}\downarrow$	0.3	-	0.3	-	0.2	-

Muscle activations. Since we were not able to find public video datasets with corresponding MTG activation signals to directly compare with, we followed the evaluation procedure in [29] to assess general trends, and overlayed our mean muscle activation predictions (black, dashed) for gait cycles from H36M with EMG data of pertinent muscles from other gait studies [72, 76, 8, 29] in rows 1 and 3 of Fig. 4, with min-max scaling applied to all values. We also included the predicted activations (dotted) from [29]; however, they use a different muscle model. While muscle activations can be seen as surrogate representations of EMGs, the two are not exactly the same. Hence, mismatches in timing and magnitude will exist, and peaks and valleys may be further amplified by the min-max scaling applied. In general, raw EMG values can vary widely due to electrode placement.

To mimic methods that regress joint torques as a linear combination of joint kinematics, we experimented with estimating the muscle activation as a "kinematic effort term" (denoted $\hat{\alpha}$), specifically as a joint's angular velocity relative to its limit plus an additionally regressed offset term:

$$\hat{\alpha}^d = \dot{q}^d / \dot{q}^d_{max} + \mathfrak{Z}^d_\alpha, \qquad d \in \{flex, ext\} \tag{18}$$

We plotted $\hat{\alpha}$ results (gray, dashed) in rows 2 and 4 of Fig. 4. In comparison, MusclePose (rows 1 and 3) seems to better follow literature trends overall, such as having a more noticeable hitch (or second peak) for hip adduction, external rotation, knee flexion, etc. Furthermore, for the $\hat{\alpha}$ case, ankle

plantarflexion seems to be deactivated during the middle of the gait cycle, when it should peak. Row 4 of Tab. 2 also shows that MusclePose quantitatively outperforms the $\hat{\alpha}$ case.

Kinematic plausibility. In addition to the joint positional errors in Sec. 4.2, metrics such as acceleration loss (ACC), foot skating (FS), and ground penetration (GP) were introduced to further evaluate kinematic plausibility. We computed these values for H36M and 3DPWoc in Tab. 4, where ACC is the mean L2 norm in mm/frame^2 between the predicted and ground truth keypoint accelerations to access jitter. We also included mean torque variation (MTV) as the mean absolute change in joint torques over consecutive frames (in Newton*metres/frame) to assess torque continuity. FS is the average displacement in mm of vertices in contact with the ground in consecutive frames. GP is the average vertical distance to the ground in mm of vertices below the ground.

Table 4: Plausibility metrics. ^{P} indicates Procrustes aligned. (%f) indicates % of frames.

		Pos.	*Kinematic plausibility*			Float (%f)			*Kinetic plausibility*		GRF$_v^{/W}$			GRF$_v^{/W}$ (%f)		
		MJE↓	ACC↓	FS	GP	$\mathcal{H}_{min} > \{1, 10, 20\}$mm			$\mathbf{F}_{res}^{/W}$↓	MTV	{med	q99	max}	< {0.01, 0.1, 0.5}		
†H36M	CLIFF	**46.5**	26.3	-	-		-		-	-	-			-		
	PhysPT	50.6	13.7	34.7	6.8	{59.0	31.6	8.5}	0.4	5.3	{0.4	2.4	10.0}	{7.2	20.8	60.0}
	MusclePose	48.4	**12.9**	37.2	26.0	{8.0	3.0	1.3}	**0.1**	2.5	{1.0	1.2	3.0}	{3.0	3.0	5.2}
†3DPWoc	CLIFF	**24.0**^{P}	13.8^P	-	-		-		-	-		-			-	
	PhysPT	25.9^P	**3.0**^{P}	7.8	11.2	{82.9	73.9	57.2}	0.9	27.0	{0.5	1.2	3.9}	{5.3	11.8	52.4}
	MusclePose	27.6^P	4.3^P	12.8	30.8	{6.0	4.7	3.7}	**0.3**	12.1	{1.0	1.6	4.3}	{5.3	5.3	6.3}

While we see an improvement in jitter from both physics pose estimators, as indicated by a lower ACC compared to CLIFF, a lower FS or GP may not be strictly better. For one, foot sliding may occur naturally. And in terms of the latter, while human bodies deform under pressure and contact, the SMPL mesh does not model this deformation, and will instead penetrate the object it is in contact with [68]. During walking for example, minimizing GP while assuming this rigidity may restrict natural ankle rotation, potentially leading to the smaller ankle torques compared to literature values in Fig. 3. On the other hand, we should also check for floating. We reported the percentage of frames (%f) when the minimum vertex height $\mathcal{H}_{min}$ is above certain thresholds (1, 10, 20mm). Since we removed the non-feet-ground contact sequences, there are very minimal frames where "floating" occurs. We see that while PhysPT has less GP, it also includes more floating.

Ground reaction force. We can also characterize floating as when GRF$_v$ is small. As such, we also reported the percentage of frames when GRF$_v$ is below certain thresholds (1%, 10%, 50% of body weight) in Tab. 4. The results are consistent with $\mathcal{H}_{min}$, both indicating less floating (lower %f) for MusclePose. In Fig. 5, we plotted the predicted median vertical GRF of a foot, divided by body weight, for gait cycles in H36M. Compared to PhysPT (gray), we see that MusclePose's predictions (purple) are closer to literature values (greens) from [18, 75].

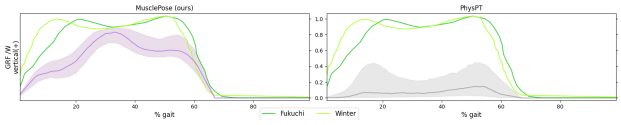


Figure 5: Median vertical GRF divided by body weight, of a foot for gait cycles in H36M, with a 25-75% quantile band.

5 Conclusion

In conclusion, we introduced MusclePose as the first PHPE method to simultaneously predict human kinematics, kinetics, muscle signals, and detailed anthropometrics from monocular video. In Sec. 4.3 and 4.4, we showed how the additions of muscle-dynamics modeling and detailed anthropometrics predictions improve the kinetic plausibility of regression-based PHPE, while being competitive with purely-kinematic pose estimators in positional accuracy in Sec. 4.2. Our framework consists of customizable components, does not require an external physics engine, and can be trained end-to-end.

Acknowledgements. We acknowledge financial support from the Canada Research Chairs Program, Canadian Sports Institute Ontario, and a Mitacs grant.

References

[1] D. E. Anderson, M. L. Madigan, and M. A. Nussbaum. Maximum voluntary joint torque as a function of joint angle and angular velocity: Model development and application to the lower limb. *Journal of Biomechanics*, 40(14):3105–3113, 2007.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, 2009.

[3] N. A. Bernshtein. The co-ordination and regulation of movements. 1967.

[4] C. Brown, W. McNally, and J. McPhee. Optimal control of joint torques using direct collocation to maximize ball carry distance in a golf swing. *Multibody System Dynamics*, 50(3), 2020.

[5] C. Brown and J. J. McPhee. Predictive forward dynamic simulation of manual wheelchair propulsion on a rolling dynamometer. *Journal of biomechanical engineering*, 2020.

[6] A. Buchmann, S. Wenzler, L. Welte, and D. Renjewski. The effect of including a mobile arch, toe joint, and joint coupling on predictive neuromuscular simulations of human walking. *Scientific Reports*, 2024.

[7] Y. Cai, L. Ge, J. Liu, J. Cai, T. J. Cham, J. Yuan, and N. M. Thalmann. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, 2019.

[8] T. Castermans, M. Duvinage, G. Cheron, and T. Dutoit. Towards effective non-invasive brain-computer interfaces dedicated to gait rehabilitation systems. *Brain Sciences*, 4(1):1–48, 2014.

[9] S.-W. Chen, W.-T. Tang, J.-T. Kung, T.-Y. Hung, W.-H. Lin, Y.-L. Chen, and D. J. Burgee. Comparison of ground reaction force among stride types in baseball pitching. *Sports Biomechanics*, 0(0):1–14, 2024.

[10] Y. Cheng, B. Yang, B. Wang, and R. T. Tan. 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020.

[11] H. Ci, C. Wang, X. Ma, and Y. Wang. Optimizing network structure for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, 2019.

[12] D. J. Cleather and A. M. J. Bull. Lower-extremity musculoskeletal geometry affects the calculation of patellofemoral forces in vertical jumping and weightlifting. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 224:1073 – 1083, 2010.

[13] B. Danaei and J. J. McPhee. Model-based acetabular cup orientation optimization based on minimizing the risk of edge-loading and implant impingement following total hip arthroplasty. *Journal of biomechanical engineering*, 2022.

[14] R. Drillis, R. Contini, and M. Bluestein. Body segment parameters. *Artificial limbs*, 8(1):44–66, 1964.

[15] G. A. Dudley, R. T. Harris, M. R. Duvoisin, B. M. Hather, and P. Buchanan. Effect of voluntary vs. artificial activation on the relationship of muscle torque to speed. *Journal of Applied Physiology*, 69(6), 1990.

[16] R. Dumas, L. Chèze, and J. P. Verriest. Adjustments to mcconville et al. and young et al. body segment inertial parameters. *Journal of biomechanics*, 40 3:543–53, 2007.

[17] R. Featherstone. *Rigid Body Dynamics Algorithms*. 2008.

[18] C. A. Fukuchi, R. K. Fukuchi, and M. Duarte. A public dataset of overground and treadmill walking kinematics and kinetics in healthy individuals. *PeerJ*, 6, 2018.

[19] E. Gartner, M. Andriluka, H. Xu, and C. Sminchisescu. Trajectory Optimization for Physics-Based Reconstruction of 3d Human Pose from Monocular Video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2022-June, 2022.

[20] E. Gärtner, M. Andriluka, E. Coumans, and C. Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022.

[21] D. Haering, C. Pontonnier, N. Bideau, G. Nicolas, and G. Dumont. Using Torque-Angle and Torque-Velocity Models to Characterize Elbow Mechanical Function: Modeling and Applied Aspects. *Journal of Biomechanical Engineering*, 141(8), 2019.

[22] N. Haraguchi, A. Nasr, K. A. Inkol, K. Hase, and J. McPhee. Human and passive lower-limb exoskeleton interaction analysis: Computational study with dynamics simulation using nonlinear model predictive control. *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, pages 844–849, 2023.

[23] A. V. Hill. The heat of shortening and the dynamic constants of muscle. *Proceedings of The Royal Society B: Biological Sciences*, 126:136–195, 1938.

[24] P. D. Hoang, R. B. Gorman, G. Todd, S. C. Gandevia, and R. D. Herbert. A new method for measuring passive length-tension properties of human gastrocnemius muscle in vivo. *Journal of Biomechanics*, 38(6):1333–1341, 2005.

[25] K. A. Inkol, C. Brown, W. McNally, C. Jansen, and J. McPhee. Muscle torque generators in multibody dynamic simulations of optimal sports performance. *Multibody System Dynamics*, 50(4), 2020.

[26] K. A. Inkol and J. J. McPhee. Using dynamic simulations to estimate the feasible stability region of feet-in-place balance recovery for lower-limb exoskeleton users. *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pages 1–6, 2022.

[27] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 2014.

[28] C. Jansen and J. J. McPhee. Predictive dynamic simulation of olympic track cycling standing start using direct collocation optimal control. *Multibody System Dynamics*, 49:53–70, 2020.

[29] C. T. John, F. C. Anderson, J. S. Higginson, and S. L. D. and. Stabilisation of walking by intrinsic muscle properties revealed in a three-dimensional muscle-driven simulation. *Computer Methods in Biomechanics and Biomedical Engineering*, 16(4):451–462, 2013.

[30] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[31] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[32] B. Katz. The relation between force and speed in muscular contraction. *The Journal of Physiology*, 96, 1939.

[33] M. A. King, C. Wilson, and M. R. Yeadon. Evaluation of a torque-driven model of jumping for height. *Journal of Applied Biomechanics*, 22(4), 2006.

[34] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[35] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the International Conference on Computer Vision*, 2019.

[36] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4496–4505, 2019.

[37] J. Li, S. Bian, C. Xu, G. Liu, G. Yu, and C. Lu. D &D: Learning Human Dynamics from Dynamic Camera. In *ECCV*, 2022.

[38] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021.

[39] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2022-June, 2022.

[40] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, 2022.

[41] Z. Li, J. Sedlar, J. Carpentier, I. Laptev, N. Mansard, and J. Sivic. Estimating 3D motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, 2019.

[42] K. Lin, L. Wang, and Z. Liu. Mesh graphormer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12919–12928, 2021.

[43] D. M. Lindsay and J. F. Horton. Trunk rotation strength and endurance in healthy normals and elite male golfers with and without low back pain. *North American journal of sports physical therapy : NAJSPT*, 1 2:80–9, 2006.

[44] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[45] T. Luan, Y. Wang, J. Zhang, Z. Wang, Z. Zhou, and Y. Qiao. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI Conference on Artificial Intelligence*, 2021.

[46] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, 2019.

[47] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, 2017.

[48] W. McNally and J. McPhee. Dynamic Optimization of the Golf Swing Using a Six Degree-of-Freedom Biomechanical Model. 2018.

[49] W. J. McNally and J. J. McPhee. Dynamic optimization of the golf swing using a six degree-of-freedom biomechanical model. 2018.

[50] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, 2018.

[51] M. Millard, T. K. Uchida, A. Seth, and S. L. Delp. Flexing computational muscle: modeling and simulation of musculotendon dynamics. *Journal of biomechanical engineering*, 135 2:021005, 2013.

[52] G. Moon and K. M. Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *ArXiv*, abs/2008.03713, 2020.

[53] A. Nasr, A. Hashemi, and J. McPhee. Scalable musculoskeletal model for dynamic simulations of upper body movement. *Computer Methods in Biomechanics and Biomedical Engineering*, 2023.

[54] A. Nasr and J. McPhee. Scalable musculoskeletal model for dynamic simulations of lower body movement. *Computer methods in biomechanics and biomedical engineering*, pages 1–27, 2024.

[55] S. Nesbit. Development of a full-body biomechanical model of the golf swing. *International Journal of Modelling and Simulation*, 27(4):392–404, 2007.

[56] A. Newell, P. Hu, L. Lipson, S. R. Richter, and V. Koltun. Comotion: Concurrent multi-person 3d motion. In *ICLR*, 2025.

[57] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, 2017.

[58] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, 2019.

[59] X. B. Peng, P. Abbeel, S. Levine, and M. Van De Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4), 2018.

[60] E. D. Pieri, M. Lund, A. Gopalakrishnan, K. P. Rasmussen, D. E. Lunn, and S. J. Ferguson. Refining muscle geometry and wrapping in the tlem 2 model for improved hip contact force prediction. *PLoS ONE*, 13, 2018.

[61] D. Rempe, L. J. Guibas, A. Hertzmann, B. Russell, R. Villegas, and J. Yang. Contact and Human Dynamics from Monocular Video. In *19th ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2020, SCA 2020 - Showcases*, 2020.

[62] M. Sabick, M. Torry, Y.-K. Kim, and R. Hawkins. Humeral torque in professional baseball pitchers. *The American journal of sports medicine*, 32:892–8, 07 2004.

[63] S. Shimada, V. Golyanik, W. Xu, P. Pérez, and C. Theobalt. Neural monocular 3D human motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), 2021.

[64] S. Shimada, V. Golyanik, W. Xu, and C. Theobalt. Phys Cap. *ACM Transactions on Graphics*, 39(6), 2020.

[65] S. Shin, J. Kim, E. Halilaj, and M. J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024.

[66] T. Siebert, C. Rode, W. Herzog, O. Till, and R. Blickhan. Nonlinearities make a difference: comparison of two common hill-type models with real muscle. *Biological Cybernetics*, 98:133–143, 2008.

[67] E. J. Sprigings. Simulation of the force enhancement phenomenon in muscle. *Computers in Biology and Medicine*, 16(6), 1986.

[68] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas. 3d human pose estimation via intuitive physics. *ArXiv*, abs/2303.18246, 2023.

[69] A. J. van Soest and M. F. Bobbert. The contribution of muscle properties in the control of explosive movements. *Biological Cybernetics*, 69(3), 1993.

[70] G. Varol, D. Ceylan, B. C. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. *ArXiv*, abs/1804.04875, 2018.

[71] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[72] H. Wang, A. Basu, G. Durandau, and M. Sartori. Comprehensive Kinetic and EMG Dataset of Daily Locomotion with 6 types of Sensors, May 2022.

[73] H. Wang, A. Basu, G. Durandau, and M. Sartori. A wearable real-time kinetic measurement sensor setup for human locomotion. *Wearable Technologies*, Feb. 2023.

[74] J. Wang, S. Yan, Y. Xiong, and D. Lin. Motion Guided 3D Pose Estimation from Videos. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12358 LNCS, 2020.

[75] D. A. Winter. *Biomechanics and motor control of human movement*. John Wiley & Sons, Hoboken, NJ, USA, 4th edition, 2009.

[76] A. R. Wu, F. Dzeladini, T. J. H. Brug, F. Tamburella, N. L. Tagliamonte, E. H. F. van Asseldonk, H. van der Kooij, and A. J. Ijspeert. An adaptive neuromuscular controller for assistive lower-limb exoskeletons: A preliminary study on subjects with spinal cord injury. *Frontiers in Neurorobotics*, Volume 11 - 2017, 2017.

[77] K. Xie, T. Wang, U. Iqbal, Y. Guo, S. Fidler, and F. Shkurti. Physics-based Human Motion Estimation and Synthesis from Videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[78] G. T. Yamaguchi. *Dynamic Modeling of Musculoskeletal Motion*. 2001.

[79] G. T. Yamaguchi. *Dynamic modeling of musculoskeletal motion: A vectorized approach for biomechanical analysis in three dimensions*. Springer, Boston, MA, USA, 1 edition, 2006.

[80] K. Yoichi, E. Sato, and T. Yamaji. Biomechanical analysis of the pitching characteristics of adult amateur baseball pitchers throwing standard and lightweight balls. *Journal of Physical Therapy Science*, 32:816–822, 12 2020.

[81] Y. S. Yoon and J. M. Mansour. The passive elastic moment at the hip. *Journal of Biomechanics*, 15(12):905–910, 1982.

[82] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021.

[83] S. Zhang, B. L. Bhatnagar, Y. Xu, A. Winkler, P. Kadlecek, S. Tang, and F. Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024.

[84] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *2013 IEEE International Conference on Computer Vision*, pages 2248–2255, 2013.

[85] Y. Zhang, J. O. Kephart, Z. Cui, and Q. Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2305–2317, June 2024.

[86] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding. 3D Human Pose Estimation with Spatial and Temporal Transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[87] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, 2017.

[88] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[89] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Learning human motion representations: A unified perspective. In *Proceedings of the International Conference on Computer Vision*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: our contributions are summarized in the last paragraph of Sec. 1, with corresponding results in Sec. 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: limitations are discussed in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: all formulas used can be found in Sec. 3, with details in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: all models implemented can be found in Sec. 3, with details, such as model coefficients or hyperparameters, listed in the supplementary material. Experiment implementation information, for both training and evaluation, can be found in Sec. 4.1, with additional details in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: while our code is not yet ready for open source, we included in the supplementary material, all necessary details required to reproduce our model and results, including implementation details of all novel components, as well as all the publicly available code repositories we borrowed from to implement and train our model, and to produce and evaluate our results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: training and test information can be found in Sec. 4.1, with additional details in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: we include quantile bands in Fig. 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: computing information is included in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: societal impacts are discussed in the supplementary material.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: assets used in this paper are cited in the main text, and the license and terms are respected and mentioned in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: the paper introduces MusclePose as a novel human pose estimation framework, with details on implementation and application.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: this paper does not involve crowdsourcing, and uses human data from publicly available datasets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: this paper does not involve crowdsourcing, and uses human data from publicly available datasets.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used for research development nor any part of this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Supplementary material for 3D Human Pose Estimation with Muscles

A Technical appendix

A.1 Human model

We assume a rigid multibody dynamics model of a human with $N_k = 18$ joints – *pelvis*, *lumbar joint*, *thoracic joint*, *neck*, *scapulas*, *shoulders*, *elbows*, *wrists*, *hips*, *knees*, and *ankles*. The pelvis is set as the root, with 3 rotational and 3 translational degrees of freedom (DoFs). The scapulas have 2 DoFs, corresponding to depression/elevation and protraction/retraction. The elbows have 2 DoFs, corresponding to flexion/extension and forearm pronation/supination. The wrists have 2 DoFs, corresponding to flexion/extension and ulnar/radial deviation. The knees have 1 DoF, corresponding to flexion/extension. All remaining joints have 3 DoFs, for a total of 47 DoFs. We selected this configuration as it aligns best with existing biomechanics models that we implemented, such as for anthropometrics estimation [16] and MTGs [51, 25].

Anthropometrics estimation. We predict the human's anthropometrics by combining our regressed residuals $\mathcal{E}$ with initial estimates based on literature values $\bar{\mathcal{A}}$ scaled by our predicted human dimensions. Specifically, we want to predict $\mathcal{A} = \cup_k \{m_k, I_{0,k}, \chi_k\}$, where m_k is the **mass** of segment k, with $I_{0,k}$ as its **inertia** tensor at zero rotation with scaling matrix Λ_k, and χ_k as its local **CoM** position relative to its segment length. For the remainder of the section, we assume relevant units to be in seconds, radians, meters, kilograms, Newtons.

From the predicted $\boldsymbol{\beta}$ and Eq. (2), we can compute the human's volume and all segment lengths L_k. We further compute the human's initial bodymass estimate $\hat{M}$ as its volume multiplied by a constant density of 985 kg/m^2. For segment k, let $s_{L,k} = L_k/H$ be its segment length relative to height, and $s_{m,k} = m_k/M$ be its mass relative to bodymass. Let "bar" $(\bar{})$ denote the human values measured by Dumas *et al.* in [16]. We set our initial estimates as $\bar{\mathcal{A}}$ scaled by $s_{L,k}/\bar{s}_{L,k}$:

$$\{M, s'_{m,k}, \Lambda_k, \chi_k\} = \{\hat{M}, \bar{s}_{m,k}\frac{s_{L,k}}{\bar{s}_{L,k}}, \bar{\Lambda}_k, \bar{\chi}_k\} + \mathcal{E} \tag{19}$$

$$m_k = s_{m,k}M, \text{ where } s_{m,k} = \frac{s'_{m,k}}{\sum_j s'_{m,j}} \tag{20}$$

$$I_{0,k} = m_k L_k^2 \Lambda_k \tag{21}$$

Lastly, we compute the **body weight** W of the human in Newtons, with $g = 9.8m/s^2$ as

$$W = g\sum_k m_k \tag{22}$$

A.2 GRFM Model

Let $\mathcal{F} = [\mathbf{F}, \mathbf{M}]^\intercal$ be the ground reaction forces and moments (GRFM) applied at the CoM of a foot in global cartesian coordinates. Let $\mathbf{F} = [F_X, F_Y, F_Z]^\intercal$ where Y is the vertical direction, and $\mathbf{z} = [z_x, z_y, z_z]^\intercal$ be the center of pressure (CoP) in the foot's local coordinates where x is along the length of the foot, such that

$$\mathbf{M} = R^0_{ankle}\mathbf{z} \times \mathbf{F} \tag{23}$$

where R_k is joint k's local rotation matrix, and $R^0_k = R^0_{p(k)}R_k$ describes the chain of rotational transformations from the world frame to its local frame. We use lower case x, y, z to denote the foot's local coordinates, and its dimensions $\{l_l, l_w, l_h\}$ as shown in Fig. 6.

We predict the force in the vertical direction scaled by body weight $F^W_Y = F_Y/W$, and CoP along the foot scaled by foot length $z^l_x = z_x/l_l$, from initial estimates based on the foot's kinematics Ψ and linear coefficients $\boldsymbol{\eta}$. Furthermore, let μ be the **coefficient of friction**, initialized at 0.8. With our regressed residuals δ, and binary contact $\mathbf{c}$, we infer:

$$\{F^W_Y, z^l_x, \mu\} = \{\boldsymbol{\eta}_{FY}\Psi,\ \boldsymbol{\eta}_{zx}\Psi,\ 0.8\} + \delta_{\{Y,l,\mu\}} \tag{24}$$

$$F_Y = F^W_Y \cdot \text{body weight} \tag{25}$$

$$z_x = z^l_x \cdot l_l \tag{26}$$

Specifically, $\Psi = [1, P_{ankle,Y}, P_{oppAnkle,Y}, \dot{P}_{ankle,Y}, \ddot{P}_{ankle,Y}, q_{ankle,z}, \dot{q}_{ankle,z}, \ddot{q}_{ankle,z},]$ includes the ankle's linear kinematics in the direction opposite of gravity, and angular kinematics corresponding to plantar/dorsiflexion. Linear coefficients were fitted on the forceplate data in [72], with $\boldsymbol{\eta}_{FY} =$ [0.3116, 3.1785, -2.2963, 0.4151, 0.0088, 0.3374, -0.1206, -0.0089] and $\boldsymbol{\eta}_{zx} =$ [0.68996, -3.1508, 0.5925, 0.21997, 0.0035, 0.18502, -0.03311, -0.00212].

The remaining δ terms are scaling factors between -1 and 1 to ensure the values in the other directions are physically possible (*i.e.* $F_X^2 + F_Z^2 \leq \mu^2 F_Y^2$ and $\boldsymbol{z}$ is within the foot's dimensions)

$$F_X = \delta_X \mu F_Y, \qquad F_Z = \delta_Z \sqrt{\mu^2 F_Y^2 - F_X^2} \tag{27}$$

$$z_y = -|\delta_h l_h|, \qquad z_z = \delta_s (l_w/2) \tag{28}$$

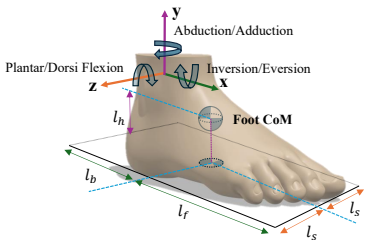


Figure 6: Foot local coordinate system and dimensions, with length $l_l = l_f + l_b$, width $l_w = 2l_s$, and CoM height l_h.

A.3 Muscle torque generators

We compute MTG torque $\boldsymbol{\tau}_{MTG}$ using the equations below that are parameterized by the γ coefficients that can be found in the tables of [54, 53]. For each joint rotational DoF $q \in \boldsymbol{q}_{[6:]}$, with angular velocity $\dot{q}$, let muscle signal $\alpha \in [0, 1]$ represent the joint's corresponding activation level for this DoF, we separate its τ_{MTG} into active torque generation τ_{active} and passive impedance $\tau_{passive}$

$$\tau_{MTG} = \tau_{active} + \tau_{passive} \tag{29}$$

We compute the **active torque** as

$$\tau_{active} = \alpha \tau_\omega \tau_\theta \tau_0 \tag{30}$$

where τ_ω models the **active-torque–angular-speed relationship** [69, 67] and is paramterized as a piecewise function with coefficients $\gamma_{1:3}$.

$$\tau_\omega(\dot{q}) = \mathbb{1}_{\dot{q}<0}\left(\frac{(1-\gamma_1)|\omega_{max}| - (\gamma_2+1)\gamma_1\gamma_3\dot{q}}{(1-\gamma_1)|\omega_{max}| + (\gamma_2+1)\gamma_1\dot{q}}\right) + \mathbb{1}_{\dot{q}\geq 0}\left(\frac{|\omega_{max}| - \dot{q}}{|\omega_{max}| + \gamma_2\dot{q}}\right) \tag{31}$$

The peak velocity ω_{max} for each joint we use the values from [54, 53]. The coefficient γ_1 is the ratio of the maximum eccentric isokinetic torque over the maximum isometric torque [69, 15], γ_2 is the slope of the eccentric and concentric functions when the angular velocity is zero [69], and γ_3 is a shape factor that influences the curvature of the hyperbola in the torque-velocity concentric relationship [4].

τ_θ models the **active-torque-angle relationship** [21, 33] and is represented by the non-negative portion of a polynomial (32) with coefficients $\gamma_{4:6}$

$$\tau_\theta(q) = (\gamma_4 + \gamma_5 q + \gamma_6 q^2)_+ \tag{32}$$

τ_0 is the **peak isokinetic torque** that controls peak MTG output at zero joint velocity, which can be measured via dynamometry.

$\tau_{passive}$ is the **passive torque** [1] of a joint that arises when the surrounding muscles, tendons, and ligaments are strained and intensifies near anatomical joint limits [1, 24, 81]. A joint's viscous damping and nonlinear stiffness are commonly described by a double exponential function [79]

$$\tau_{passive} = \gamma_{10e}^{-\gamma_{11}(\boldsymbol{q}-\boldsymbol{q}_{min})} - \gamma_{12} e^{\gamma_{13}(\boldsymbol{q}-\boldsymbol{q}_{max})} - \gamma_{14}\omega \tag{33}$$

where γ_{10-14} are passive coefficients from [48] and γ_{11} is the rotational damping linear coefficient [78] to reflect viscoelasticity. This encourages the joint to move within its **range of motion** (RoM), as a large restoring torque is produced otherwise.

A.4 Inverse dynamics [1]

We compute $\boldsymbol{\tau}_q$ using Lagrange's equations derived from d'Alembert's Principle of virtual work

$$\boldsymbol{\tau}_q = \mathfrak{M}\ddot{\boldsymbol{q}} + \mathfrak{C} - \mathfrak{F} \tag{34}$$

We can write the terms on the right hand side as:

$$\mathfrak{M} = \sum_k J_k^\intercal \mathcal{M}_k J_k, \tag{35}$$

$$\mathfrak{C} = \sum_k (J_k^\intercal \mathcal{M}_k \dot{J}_k + J_k^\intercal \begin{bmatrix} 0 & 0 \\ 0 & [J_{\Omega,k}\dot{q}]_s \end{bmatrix} \mathcal{M}_k J_k)\dot{\boldsymbol{q}} \tag{36}$$

$$\mathfrak{F} = J_{LFoot}^\intercal \mathcal{F}_{LFoot} + J_{RFoot}^\intercal \mathcal{F}_{RFoot} \tag{37}$$

where $\mathbf{I}_3$ is the identity matrix, $([\cdot]_s)$ denotes the skew-symmetric form, and

$$\mathcal{M}_k = \begin{bmatrix} m_k \mathbf{I}_3 & 0 \\ 0 & R_k^0 I_{0,k} (R_k^0)^\intercal \end{bmatrix} \tag{38}$$

To deal with potential energy, we offset the root acceleration in the direction of gravity by -9.8 m/s^2. Jacobian matrix J is the mapping from the generalized space to the global Cartesian coordinates, such that for linear and angular velocities $\boldsymbol{V}_k, \boldsymbol{\Omega}_k$ in global Cartesian coordinates, we have:

$$J_k \dot{\boldsymbol{q}} = \begin{bmatrix} J_{V,k} \\ J_{\Omega,k} \end{bmatrix} \dot{\boldsymbol{q}} = \begin{bmatrix} \boldsymbol{V}_k \\ \boldsymbol{\Omega}_k \end{bmatrix} \tag{39}$$

J can be computed analytically using a recursive algorithm such as in [17]. For segment k, we define its parent segment $p(k)$ as its neighboring segment that is closer to the root. Other than the root, each segment has one and only one parent. We define k's children $ch(k)$ as its neighboring segments further away from the root. Let $\boldsymbol{r}_{a\to b}$ denote the 3D displacement from point a to b. For segment k, with linear velocity $\boldsymbol{V}_k$ at its CoM and linear velocity $\boldsymbol{V}_k^{joint}$ at its corresponding joint, we have

$$\boldsymbol{V}_k^{joint} = \boldsymbol{V}_{p(k)} + \boldsymbol{\Omega}_{p(k)} \times \boldsymbol{r}_{p(k)\to k^{joint}} \Rightarrow J_{V,k}^{joint} = J_{V,p(k)} - [\boldsymbol{r}_{p(k)\to k^{joint}}] J_{\Omega,p(k)} \tag{40}$$

and the velocity at the CoM of segment k becomes:

$$\boldsymbol{V}_k = \boldsymbol{V}_k^{joint} + \boldsymbol{\Omega}_k \times \boldsymbol{r}_{k^{joint}\to k} \Rightarrow J_{V,k} = J_{V,p(k)} - [\boldsymbol{r}_{p(k)\to k^{joint}}] J_{\Omega,p(k)} - [\boldsymbol{r}_{k^{joint}\to k}] J_{\Omega,k} \tag{41}$$

From (41), we can compute the time derivative recursively as:

$$\dot{J}_{V,k} = \dot{J}_{V,p(k)} - [\boldsymbol{r}_{p(k)\to k^{joint}}] \dot{J}_{\Omega,p(k)} - [\boldsymbol{r}_{k^{joint}\to k}] \dot{J}_{\Omega,k} \tag{42}$$

The global angular velocity of k in skew symmetric form is:

$$[\boldsymbol{\Omega}_k] = \dot{R}_k^0 (R_k^0)^\intercal = (\dot{R_{p(k)}^0 R_k})(R_{p(k)}^0 R_k)^\intercal \tag{43}$$

$$= ... = \dot{R}_{p(k)}^0 (R_{p(k)}^0)^\intercal + R_{p(k)}^0 (\dot{R}_k R_k^\intercal)(R_{p(k)}^0)^\intercal \tag{44}$$

$$= [\boldsymbol{\Omega}_{p(k)}] + R_{p(k)}^0 [\boldsymbol{\omega}_k] (R_{p(k)}^0)^\intercal \qquad (\because [\boldsymbol{\omega}_k] = \dot{R}_k R_k^\intercal) \tag{45}$$

$$\Rightarrow \boldsymbol{\Omega}_k = \boldsymbol{\Omega}_{p(k)} + R_{p(k)}^0 \boldsymbol{\omega}_k \qquad (\because [A\boldsymbol{b}] = A[\boldsymbol{b}]A^\intercal) \tag{46}$$

To avoid confusion of notation, we also write joint k's rotation $\boldsymbol{\theta}_k \triangleq \boldsymbol{q}_k$, *i.e.* we have generalized coordinates $\boldsymbol{q} = \begin{bmatrix} \boldsymbol{X}_0 \\ \boldsymbol{\theta} \end{bmatrix} \in \mathbb{R}^{N_{DoF}\times 1}$ where $\boldsymbol{X}_0$ is the global root translation, $\boldsymbol{\theta}_0$ is the global root rotation, $\boldsymbol{\theta}_k$ describes the local rotation of segment k relative to its parent $p(k)$, and $\boldsymbol{\theta}^\intercal = [\boldsymbol{\theta}_0^\intercal \;\; \boldsymbol{\theta}_1^\intercal \;\; ... \;\; \boldsymbol{\theta}_{N_k}^\intercal]$. Let $J_{\omega,k}$ be the local Jacobian such that $\boldsymbol{\omega}_k = J_{\omega,k}\dot{\boldsymbol{\theta}}_k$. We can compute $\boldsymbol{\Omega}_k$ recursively:

$$\boldsymbol{\Omega}_k = \boldsymbol{\Omega}_{p(k)} + R_{p(k)}^0 J_{\omega,k} \dot{\boldsymbol{\theta}}_k \tag{47}$$

$$= 0 + J_{\omega,0}\dot{\boldsymbol{\theta}}_0 + ... + R_{p(p(k))}^0 J_{\omega,p(k)} \dot{\boldsymbol{\theta}}_{p(k)} + R_{p(k)}^0 J_{\omega,k} \dot{\boldsymbol{\theta}}_k \tag{48}$$

$$\triangleq J_{\Omega,k} \dot{\boldsymbol{q}} \tag{49}$$

[1] Derivations in this section are based on C. Karen Liu and Sumit Jain's multibody dynamics notes: `https://fab.cba.mit.edu/classes/865.18/design/optimization/dynamics_1.pdf`.

Let $\mathcal{P}_k$ denote the set of all ancestors of k and itself ($k \in \mathcal{P}_k$), we split $J_{\Omega,k}$ into $N_k + 1$ blocks of size 3×3:

$$J_{\Omega,k} = \begin{bmatrix} 0_{3\times3} & J_{\omega,0} & \mathbb{1}_{1\in\mathcal{P}_k} R^0_{p(1)} J_{\omega,1} & \dots & \mathbb{1}_{N_k\in\mathcal{P}_k} R^0_{p(K)} J_{\omega,N_k} \end{bmatrix} \in \mathbb{R}^{3\times N_{DoF}} \tag{50}$$

For segment k, if we represent rotation $\boldsymbol{\theta}_k = [\theta_{k,1} \quad \theta_{k,2} \quad \theta_{k,3}] \triangleq [\alpha \quad \beta \quad \gamma] \in \mathbb{R}^3$ using 3 Euler angles, removing subscript k for notation simplicity, we have

$$[\boldsymbol{\omega}] = \dot{R}R^\intercal = \sum_i \frac{\partial R}{\partial \theta_i} R^\intercal \dot{\theta}_i = \frac{\partial R}{\partial \alpha} R^\intercal \dot{\alpha} + \frac{\partial R}{\partial \beta} R^\intercal \dot{\beta} + \frac{\partial R}{\partial \gamma} R^\intercal \dot{\gamma} \tag{51}$$

and it remains to compute $\mathbf{J}$'s s.t.

$$J_\omega \triangleq [\mathbf{J}_1 \quad \mathbf{J}_2 \quad \mathbf{J}_3], \quad \text{where } [\mathbf{J}_1] = \frac{\partial R}{\partial \alpha} R^\intercal, \quad [\mathbf{J}_2] = \frac{\partial R}{\partial \beta} R^\intercal, \quad [\mathbf{J}_3] = \frac{\partial R}{\partial \gamma} R^\intercal \tag{52}$$

$$\text{which satisfies } [\boldsymbol{\omega}] = \sum_i [\mathbf{J}_i] \dot{\theta}_i \quad \text{and } \boldsymbol{\omega} = J_\omega \dot{\boldsymbol{\theta}} \tag{53}$$

Finally, let segment $l \in \mathcal{P}_k$ be an ancestor of k and denote $\{\boldsymbol{J}_{\Omega,k}\}_l \triangleq R^0_{p(l)} J_{\omega,l}$ as the $(l+2)$-th 3×3 block in $J_{\Omega,k}$ from (50), we can compute its time derivative as:

$$\{\dot{\boldsymbol{J}}_{\Omega,k}\}_l = \dot{R}^0_{p(l)} J_{\omega,l} + R^0_{p(l)} \dot{J}_{\omega,l}, \quad \text{with } \dot{J}_{\omega,l} = \sum_{l,i} \frac{\partial J_{\omega,l}}{\partial \theta_{l,i}} \dot{\theta}_{l,i} \tag{54}$$

following (52), it remains to compute $\dot{\mathbf{J}}$'s s.t.

$$\dot{J}_{\omega,l} \triangleq [\dot{\mathbf{J}}_{l,1} \quad \dot{\mathbf{J}}_{l,2} \quad \dot{\mathbf{J}}_{l,3}], \quad \text{where } \dot{\mathbf{J}}_{l,j} = \sum_{l,i} \frac{\partial \mathbf{J}_{l,j}}{\partial \theta_{l,i}} \dot{\theta}_{l,i} \tag{55}$$

A.5 Neural network

We trained a transformer encoder consisting of 8 layers with a latent dimension of 256, using total loss $\mathcal{L}_{total}$ with weights $\boldsymbol{\lambda}_{kin} = [0.5, 10, 1000, 1, 20, 1000]$ and $\boldsymbol{\lambda}_{dyn} = [100, 100, 20]$. We trained on AMASS with a sequence input length of 16 frames, after removing sequences containing non feet-ground contact, with contact labels from [83], for 25 epochs. We used the AdamW optimizer [44] with a weight decay of 10^{-4} and an initial learning rate of 10^{-4} that decreases by 20% every 5 epochs. The entire process can be trained in about 12 hours on a single Titan Xp GPU.