PhysPT: Physics-aware Pretrained Transformer for Estimating Human Dynamics from Monocular Videos

Yufei Zhang[1], Jeffrey O. Kephart[2], Zijun Cui[1,3], Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]IBM Research, [3]University of Southern California
{zhangy76, jiq}@rpi.edu, kephart@us.ibm.com, ceejkl@gmail.com

Abstract

*While current methods have shown promising progress on estimating 3D human motion from monocular videos, their motion estimates are often physically unrealistic because they mainly consider kinematics. In this paper, we introduce **Phys**ics-aware **P**retrained **T**ransformer (**PhysPT**), which improves kinematics-based motion estimates and infers motion forces. PhysPT exploits a Transformer encoder-decoder backbone to effectively learn human dynamics in a self-supervised manner. Moreover, it incorporates physics principles governing human motion. Specifically, we build a physics-based body representation and contact force model. We leverage them to impose novel physics-inspired training losses (i.e., force loss, contact loss, and Euler-Lagrange loss), enabling PhysPT to capture physical properties of the human body and the forces it experiences. Experiments demonstrate that, once trained, PhysPT can be directly applied to kinematics-based estimates to significantly enhance their physical plausibility and generate favourable motion forces. Furthermore, we show that these physically meaningful quantities translate into improved accuracy of an important downstream task: human action recognition.*

1. Introduction

Monocular 3D human motion estimation is essential for applications like human-computer interaction [8, 74], motion analysis [14], and robotics [17]. This task is inherently challenging due to the absence of depth information and the intricate interplay of forces and human body movements.

Recent advances in deep learning, along with progress in 3D human modeling [47, 85], have substantially improved the reconstruction of 3D humans from a single image [28, 33, 38, 76, 105]. With video inputs, current research aims to enhance model performance by exploiting temporal information. Some authors devise temporal models that extract meaningful features from videos to improve performance [6, 10, 11, 13, 23, 36, 55, 64, 69, 73, 80, 98, 100]. Other authors learn motion priors that capture natural 3D body movement patterns. Integrating the learned priors into model training can promote smooth motion estimates [31, 48, 61, 66]. While these approaches enhance reconstruction to some extent, they often produce unrealistic estimates characterized by noticeable physical artifacts such as motion jittering and foot skating.

To address this limitation, a promising strategy is to leverage physical principles governing body movements. In this approach, the human body is treated as an articulated rigid body, and human dynamics are described through the Euler-Lagrange equations. These equations link body mass, inertia, and physical forces (including joint actuations and contact forces) to body motion through ordinary differential equations. Some researchers [19, 20, 40, 60, 67, 84, 88] formulate optimization frameworks that jointly estimate unknown physical parameters and refine kinematics-based estimates by aligning them with the physics equations. Alternatively, others [24, 39, 49, 68, 93] employ learning-based frameworks. They sidestep the cumbersome manual parameter tuning inherent in optimization-based methods by training neural networks to predict the parameters.

Yet, a key challenge remains: physics information, including physical properties of human bodies and motion forces, is absent in current 3D motion capture datasets [51]. To incorporate physics, existing methods generally rely on physics engines [15, 75]. This entails creating proxy bodies with geometric primitives to capture body properties, importing these proxies into a physics engine, and then leveraging the physics engine to compute the necessary physical parameters and simulate body motion. The problem with this approach arises from the difficulty of efficiently computing gradients from physics engine outputs [19], thereby limiting their seamless integration with deep learning models. Moreover, existing physics-based models are primarily trained with 3D annotated videos, which are challenging to acquire in practice. Consequently, the trained models may not generalize well to unseen scenarios.

In this paper, we propose a novel framework for learning human dynamics that circumvents the need for 3D an-

notated videos and effectively integrates physics with advanced deep models. Specifically, drawing inspiration from recent success of pre-trained Transformers [77] in temporal modeling, we propose leveraging a Transformer encoder-decoder architecture and training in a self-supervised manner by reconstructing input human motion. When incorporating physics, we bypass unrealistic body proxies by directly computing body physical properties from the widely adopted 3D body model, SMPL [47]. We also introduce a contact model to effectively model the contact forces. We utilize these models to derive motion forces from training sequences and impose novel physics-inspired training losses, including force loss, contact loss, and Euler-Lagrange loss. We train the Transformer model only using existing motion capture data. Once trained, our physics-aware pretrained Transformer (PhysPT) can be applied on top of any kinematics-based reconstruction model to produce enhanced motion and force estimates from monocular videos. In summary, our main contributions include:

- We introduce PhysPT, a Transformer encoder-decoder model trained through self-supervised learning with incorporation of physics. Once trained, PhysPT can be combined with any kinematic-based model to estimate human dynamics without additional model fine-tuning.
- We present a novel framework for incorporating physics. This includes a physics-based body representation and a contact force model, and, subsequently, the imposition of a set of novel physics-inspired losses for model training.
- We demonstrate through experiments that PhysPT significantly enhances the physical plausibility of motion estimates and infers favourable motion forces. Furthermore, we demonstrate that the enhanced motion and force estimates translate into accuracy improvements in an important downstream task: human action recognition.

2. Related Work

Kinematics-based Human Motion Estimation. Methods modeling body kinematics estimate body geometry configuration solely. Among these approaches, one line of work involves optimization-based pipelines that iteratively fit a prior body model to 2D observations to reconstruct a 3D human [1, 4, 22, 78, 83, 89, 111]. Others embrace deep learning models to directly predict 3D human bodies. Given a single input image, existing methods have proposed different model architectures with various intermediate and output representations to improve the reconstruction accuracy [18, 32, 34, 38, 41, 43, 50, 53, 63, 81, 91, 104].

Given input of a monocular video, current kinematics-based methods aim to fully harness temporal information to obtain improved results. Various temporal models are developed based on Temporal Convolutional Networks [29, 36, 55, 99, 100], Graph Convolutional Networks [6, 10, 11, 80, 98], Recurrent Neural Networks [13, 23, 31, 48, 72, 92], Transformer [21, 42, 58, 59, 64, 73, 108, 112], or those explicitly capturing and exploiting attention [44, 71, 79, 82]. Another common approach to encouraging realistic temporal predictions is to incorporate smoothness constraints or motion priors during training [31, 48, 61, 94, 101]. These kinematics-based methods, however, often produce noticeable physical artifacts due to their failure to realistically capture the complexity of human motion.

Physics-based Human Motion Estimation. Physics-based approaches explicitly leverage physics principles, particularly the Euler-Lagrange equations, to capture human dynamics. Prior works have adopted optimization frameworks to jointly estimate motion forces and refine initial kinematics-based motion estimates by minimizing the residuals introduced by the Euler-Lagrange equations [40, 60, 84]. Directly estimating the exerted forces is challenging; therefore, others employ a character control methodology. In this paradigm, kinematics-based estimates act as reference motions, and the forces needed to emulate these motions in a physics engine are predicted by estimating the parameters of a controller [19, 37, 67, 88, 90, 110]. However, these optimization-based methods often require careful tuning of the control parameters. Some approaches instead leverage neural networks to estimate the parameters, where the models are trained through fully supervised learning [68] or reinforcement learning [24, 26, 49, 93]. In these approaches, incorporating the physics engine alongside learning models falls short of achieving an effective end-to-end integration of physics. Li *et al.* [39] enhance the learning process by analytically computing some of the physical parameters coupled with the usage of 3D supervisions. While the produced results are promising, existing learning-based methods rely on 3D annotated videos for training and often exhibit poor generalization. In contrast, our model adopts self-supervised learning, trained solely using existing 3D motion data without images. Additionally, we introduce a novel framework that seamlessly bridges the gap between body kinematics and physics without relying on physics engines, facilitating the effective integration of physics with advanced deep learning models.

Full Human Dynamics Estimation. Fully capturing human dynamics requires determining both body movements and the forces exerted by individuals [3, 12, 62, 65]. Prevailing methods for estimating human dynamics primarily focus on inferring forces from 3D motion capture data [5, 95–97]. These estimated forces are used to facilitate tasks such as human action recognition [52] or human motion prediction [45, 106, 107]. Our approach can address a more challenging task: the estimation of full human dynamics from a monocular video. We do so without utilizing any ground truth force labels. To our knowledge, we are the first to demonstrate that forces inferred from monocular videos can improve human action recognition.

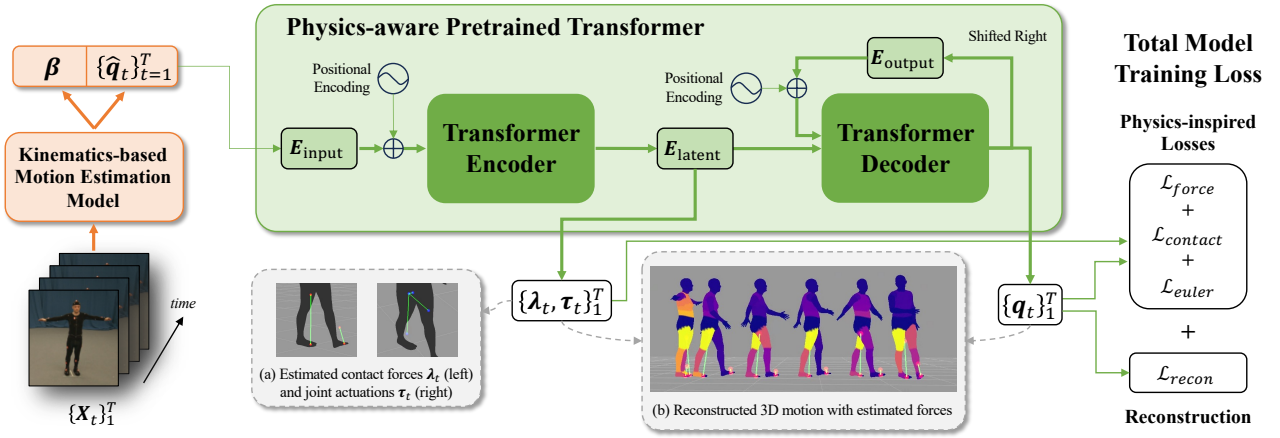

Figure 1. **Method Overview.** The proposed framework consists of a kinematics-based motion estimation model (orange) and a physics-aware pre-trained Transformer (green) for estimating human dynamics from a monocular video. Inset (a) illustrates joint actuation of right pelvis and contact forces at each foot. (b) illustrates reconstructed body motion and inferred forces with lighter colors representing greater joint actuation magnitudes (e.g. upper body joints when the figure is standing, and leg joints when it is walking).

3. Proposed Method

Fig. 1 shows an overview of our method. Given T video frames $\{\mathbf{X}_t\}_{t=1}^T$, a kinematics-based model is employed to generate initial motion estimates $\{\hat{\mathbf{q}}_t\}_{t=1}^T$, followed by the proposed PhysPT to estimate refined motion $\{\mathbf{q}_t\}_{t=1}^T$ and infer forces $\{\boldsymbol{\lambda}_t, \boldsymbol{\tau}_t\}_{t=1}^T$. In Sec. 3.1, we introduce the physics equations for modeling human dynamics and describe how we generate the kinematics-based estimates. In Sec. 3.2, we delve into the key components of PhysPT.

3.1. Preliminary

Euler-Lagrange Equations. As a complex physical system composed of multiple interacting body parts, the human body is often modeled using rigid body dynamics. In this context, the Euler-Lagrange equations provide a concise mathematical description of human dynamics within a generalized coordinate system.

The generalized coordinates are defined by variables that fully specify the system's state. Based on the successful geometry body model SMPL [47], we represent a 3D human body in terms of a mesh model using body pose $\boldsymbol{\theta} \in \mathbb{R}^{23\times3}$ and body shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$. The pose parameters characterize the rotations of 23 body joints, while the shape parameters control the variations in body attributes, e.g. girth. Given that the body shape remains constant within a video, the 3D body trajectory in the world frame can be fully specified by the pose $\boldsymbol{\theta}$ plus body translation $\mathbf{T} \in \mathbb{R}^3$ and rotation $\mathbf{R} \in \mathbb{R}^3$ via a generalized coordinate $\mathbf{q}$:

$$\mathbf{q} = \{\mathbf{T}, \mathbf{R}, \boldsymbol{\theta}\}. \tag{1}$$

Denoting the velocity and the acceleration in the generalized coordinates as $\dot{\mathbf{q}}$ and $\ddot{\mathbf{q}}$, respectively, the body dynamics governed by the Euler-Lagrange equations are:

$$\mathbf{M}(\mathbf{q};\ \mathbf{m}, \mathbf{I})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}};\ \mathbf{m}, \mathbf{I}) + \mathbf{g}(\mathbf{q};\ \mathbf{m}) = \mathbf{J}_C^T\boldsymbol{\lambda} + \boldsymbol{\tau}, \tag{2}$$

where $\mathbf{M}(\mathbf{q};\ \mathbf{m}, \mathbf{I})$ is the generalized inertia matrix determined by the position $\mathbf{q}$, the body mass $\mathbf{m}$, and the inertia $\mathbf{I}$. $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}};\ \mathbf{m}, \mathbf{I})$ represents the Coriolis and centrifugal forces. $\mathbf{g}(\mathbf{q};\ \mathbf{m})$ indicates the gravitational forces. $\boldsymbol{\lambda} \in \mathbb{R}^{3n_c}$ denotes the contact forces, where n_c is the number of points of contact. $\mathbf{J}_C \in \mathbb{R}^{3n_c\times75}$ is the contact Jacobian matrix that describes the mapping between the contact points' Cartesian velocity, $\mathbf{v}_C \in \mathbb{R}^{3n_c}$, and the generalized velocity, $\dot{\mathbf{q}}$, according to the equation $\mathbf{v}_C = \mathbf{J}_C\dot{\mathbf{q}}$. Additionally, $\boldsymbol{\tau} \in \mathbb{R}^{75}$ represents joint actuations, as exemplified in Fig. 1-a for right pelvis joint.

Kinematics-based Motion Estimation Model. We first employ an established method to obtain per-frame 3D body pose and shape $\{\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\beta}}_t\}_{t=1}^T$ from the video input. It places no restrictions on which method is used; for our experiments we use recent publicly-available models. The pose and shape estimated by those traditional 3D human reconstruction models only capture body movements in the body frame, lacking a global motion trajectory to fully specify the generalized positions defined in Eq. 1. As in [94], we train a global trajectory predictor to provide per-frame global translation and rotation $\{\hat{\mathbf{T}}_t, \hat{\mathbf{R}}_t\}_{t=1}^T$ based on the local body movements. The global trajectory predictor is trained independently and produces the global estimates without additional model fine-tuning. Further details of the model architecture and training are in Supp. A. Finally, by combining the estimated global trajectory with the local body pose, we obtain the initial generalized position estimates $\{\hat{\mathbf{q}}_t\}_{t=1}^T$, which are input to PhysPT for further refinement. Meanwhile, we consider the final shape estimate $\boldsymbol{\beta} = \frac{1}{T}\sum_{t=1}^T \hat{\boldsymbol{\beta}}_t$ since the subject's shape remains unchanged over time.

3.2. Physics-aware Pretrained Transformer

The initial kinematics-based motion estimates maintain reasonable per-frame reconstruction accuracy. The Physics-aware Pretrained Transformer introduced in this section further enhances the motion estimates and infers motion forces. In the following, we first introduce the Transformer encoder-decoder backbone of PhysPT in Sec. 3.2.1. To incorporate physics into the model, we build a physics-based body representation (Sec. 3.2.2) and a contact force model (Sec. 3.2.3), which enable the formulation of physics-inspired training losses (Sec. 3.2.4).

3.2.1 Transformer Encoder-Decoder Backbone

Differing from existing works that primarily utilize a Transformer encoder to learn representations, we exploit a Transformer encoder-decoder architecture [77]. As illustrated in Fig. 1 (green region), the model first extracts embedding $\mathbf{E}_{input} \in \mathbf{R}^{T \times n_f}$ from the kinematics-based estimates $\{\hat{\mathbf{q}}_t\}_{t=1}^{T}$ using a linear layer. This $\mathbf{E}_{input}$, combined with a time positional encoding, is then fed into the Transformer encoder to generate a latent embedding $\mathbf{E}_{latent} \in \mathbf{R}^{T \times n_l}$. Here, n_f and n_l are embedding dimensions. Subsequently, the decoder generates refined estimates $\{\mathbf{q}_t\}_{t=1}^{T}$ via autoregressive prediction. Specifically, at time frame $m+1$, the previous m estimates are projected into embeddings $\mathbf{E}_{output} \in \mathbf{R}^{m \times n_f}$, to which a positional encoding is added. Together with $\mathbf{E}_{latent}$, this is input to the Transformer decoder to produce the motion prediction.

Leveraging the Transformer encoder-decoder backbone can effectively capture temporal information in a self-supervised manner by reconstructing the input. Specifically, denoting an input sequence from existing 3D motion capture data as $\{\bar{\mathbf{q}}_t\}_{t=1}^{T}$, we compute a mean squared error on the generalized positions and 3D joint positions, leading to the added reconstruction loss $\mathcal{L}_{recon}$:

$$\begin{aligned} \mathcal{L}_{recon} &= \sum_{t=1}^{T} \gamma_q \mathcal{L}_{q,t} + \gamma_J \mathcal{L}_{J,t}, \\ \mathcal{L}_{q,t} &= \|\mathbf{q}_t - \bar{\mathbf{q}}_t\|_2^2, \\ \mathcal{L}_{J,t} &= \|\mathbf{J}_t - \bar{\mathbf{J}}_t\|_2^2, \end{aligned} \tag{3}$$

where the 3D joint positions $\mathbf{J}_t \in \mathbb{R}^{n_J \times 3}$ are computed from the generalized positions and the body shape parameters using forward kinematics and n_J is the number of body joints. γ_q and γ_J are training loss weights. To enhance model robustness, we introduce random Gaussian noise into the input during training while the model is still tasked with reconstructing the clean input. Up to this point, the Transformer model is trained to effectively learn the geometry information from motion data, but it is agnostic to physics and insufficient to faithfully capture human dynamics.

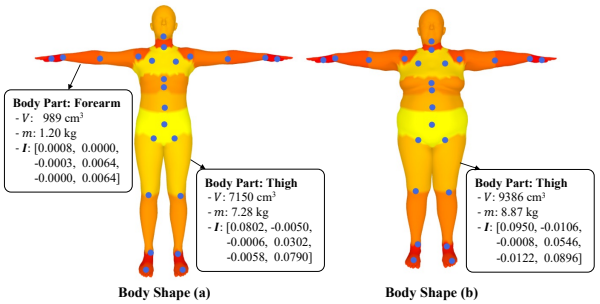

Figure 2. **Phys-SMPL.** Besides 3D positions, Phys-SMPL models the volume (V), mass (m), and inertia ($\mathbf{I}$) of every body parts. Lighter colors represent larger body weight distributions.

3.2.2 Physics-based Body Representation

To empower the model to capture physics, we need to first model the physical properties of human bodies. For this purpose, we introduce a physics-based body representation, Phys-SMPL. As shown in Fig. 2, Phys-SMPL incorporates the mass $\mathbf{m}(\boldsymbol{\beta}) \in \mathbb{R}^{24}$ and inertia $\mathbf{I}(\boldsymbol{\beta}) \in \mathbb{R}^{24 \times 3 \times 3}$ of 24 body parts in additional to SMPL's geometry information. Specifically, we first close the meshes of each body part. This allows us to compute the volume of a body part as the sum of the tetrahedrons formed by its centroid and mesh faces. Based on the average mass density of the human body [56], we calculate the mass and, subsequently, the inertia of different body parts. Note that these physical body properties are computed directly from SMPL's geometry information specified by the shape parameters $\boldsymbol{\beta}$, without the need for creating unrealistic body proxies. Expanding on Phys-SMPL, we analytically calculate the physical terms in the Euler-Lagrange equations (Eq. 2). The analytical computation of physical parameters is fully differentiable, enabling the seamless integration of physics with learning models during the training process. Further details of Phys-SMPL and the analytical calculation are in Supp. B.

3.2.3 Continuous Contact Force Model

To capture human dynamics, the motion forces must be modeled as well. For the joint actuations and contact forces, modeling the contact forces can be particularly challenging. The contact status often needs to be determined beforehand — and this is in itself difficult to do accurately. The discrete contact status also introduces a non-differentiable process in estimating the forces. To address this issue, we draw inspiration from the continuous contact model proposed by [5] for estimating ground reaction forces from 3D motion, which entails introducing a spring-mass system as illustrated in Fig. 3. Specifically, the ground contact force experienced by a point p at time t is modeled as:

$$\boldsymbol{\lambda}_{p,t} = s_{p,t}(-k_{h,t}\mathbf{b}_{h,t} - k_{n,t}\mathbf{b}_{n,t} - c_t\mathbf{v}_{C,t}). \tag{4}$$

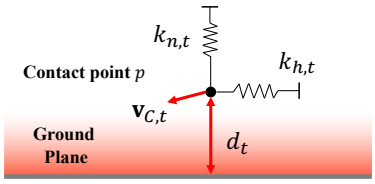

Figure 3. **Continuous Contact Force Model.** The contact forces received by a point p at time frame t are determined by its velocity and distance to the ground through a spring-mass system built along the horizontal ($k_{h,t}$) and normal ($k_{n,t}$) directions.

where $k_{h,t}$ and $k_{n,t}$ denote the stiffness of the spring-mass in the horizontal and normal directions, respectively, while c_t represents the damping factor. The scalar $s_{p,t} = 2\sigma(-60d_t)\sigma(-60\|\mathbf{v}_{C,t}\|)$ regulates the force magnitude, where $\sigma(\cdot)$ represents the Sigmoid function, d_t denotes the point's distance to the ground, and $\|\mathbf{v}_{C,t}\|$ is its velocity. Additionally, $\mathbf{b}_{h,t} = [d_{t,x} - 0.5; d_{t,y} - 0.5; 0]$ and $\mathbf{b}_{n,t} = [0; 0; d_t - 2]$ is the distance to the reference point in the horizontal and normal direction, respectively. Here, $d_{t,x}$ and $d_{t,y}$ are projections of d_t onto the x and y axes using the normal of the contact point. The units are in meters. For the sake of computational efficiency (and as further discussed in Supp. C), we apply the contact model to a subset of vertices within each body part. The contact model captures the essentials of natural contact behavior, where points closest to the ground and most stable experience larger forces. Utilizing the contact model also avoids the problems presented in estimating the discrete contact status.

3.2.4 Physics-inspired Training Losses

Building upon the physics-based body representation and force model, we can effectively integrate physics with the model by utilizing several physics-inspired training losses.

To formulate these losses, the first step involves deriving valuable motion force information from training sequences. Given a 3D trajectory $\{\bar{\mathbf{q}}_t\}_{t=1}^T$ from training data, we utilize the finite difference to obtain the velocity and acceleration $\{\dot{\bar{\mathbf{q}}}_t, \ddot{\bar{\mathbf{q}}}_t\}_{t=1}^T$. We then formulate the following optimization problem to recover the motion forces at time frame t as:

$$\begin{aligned} \arg\min_{\mathbf{x}_t, \boldsymbol{\tau}_t} \quad & \|\bar{\mathbf{M}}_t\ddot{\bar{\mathbf{q}}}_t + \bar{\mathbf{C}}_t + \bar{\mathbf{g}}_t - \bar{\mathbf{J}}_{C,t}^T\bar{\mathbf{A}}_t\mathbf{x}_t - \boldsymbol{\tau}_t\|_2^2 \\ s.t. \quad & \mathbf{0} < \mathbf{x}_t < \bar{\mathbf{x}}_{max}. \quad \text{(stiffness and damping constraints)} \end{aligned} \tag{5}$$

The objective is a least squared error introduced by the Euler-Lagrange equations in Eq. 2. (Variables for each physical term are omitted for simplicity.) The optimization variables consist of joint actuations $\boldsymbol{\tau}_t$ and the spring-mass model parameters $\mathbf{x}_t$. Specifically, the contact force model in Eq. 4 is written in a vector representation as $\boldsymbol{\lambda}_{p,t} = \mathbf{A}_{p,t}\mathbf{x}_{p,t}$, where $\mathbf{A}_{p,t} = s_{p,t}[-\mathbf{b}_{h,t}, -\mathbf{b}_{n,t}, -\mathbf{v}_{C,t}]$ include the position-related parameters, and $\mathbf{x}_{p,t} = [k_{h,t}; k_{n,t}; c_t]$ involve the unknown stiffness and damping parameters. Concatenating the contact forces of all modeled points, we have $\boldsymbol{\lambda}_t = \mathbf{A}_t\mathbf{x}_t$. Additionally, the linear inequality constraints on $\mathbf{x}_t$ are specified considering the maximal contact forces that can be experienced when a human is standing normally. The formulated optimization problem is a standard Quadratic Programming problem for which the global minimum is found by utilizing CVXOPT [16]. The final solution yields forces that comply with the constraints set by the spring-mass model and optimally satisfy the Euler-Lagrange equations. Utilizing the inferred forces enables effective incorporation of physics into the model by imposing the following physics-inspired losses.

Force Loss. We first employ the derived motion forces to guide the model to produce realistic motion forces and extract meaningful latent representations for predicting physically plausible motion. Specifically, we introduce a linear layer to project the latent representation $\mathbf{E}_{latent}$ to motion forces $\{\boldsymbol{\lambda}_t, \boldsymbol{\tau}_t\}_{t=1}^T$ based on the contact force model. Given the derived forces $\bar{\boldsymbol{\tau}}_t$ and $\bar{\boldsymbol{\lambda}}_t$, we train the model by minimizing the mean absolute error as:

$$\mathcal{L}_{force} = \sum_{t=1}^{T} \gamma_\tau\|\boldsymbol{\tau}_t - \bar{\boldsymbol{\tau}}_t\|_1 + \gamma_\lambda\|\boldsymbol{\lambda}_t - \bar{\boldsymbol{\lambda}}_t\|_1. \tag{6}$$

Contact Loss. Moreover, we apply constraints to the vertices experiencing contact forces, obtaining realistic contact behavior through the following contact loss:

$$\mathcal{L}_{contact} = \sum_{t=1}^{T} \frac{1}{n_{C_t}} \sum_{n_i \in \mathcal{C}_t} \gamma_v\|\mathbf{v}_{n_i,t}\|_1 + \gamma_z|z_{n_i,t}|, \tag{7}$$

where $\mathcal{C}_t$ denotes the set of vertices that exhibit contact forces (calculated via Eq. 5), and n_{C_t} is the set size. $\mathbf{v}_{n_i,t}$ and $z_{n_i,t}$ represent the velocity and vertical distance to the ground of the n_i^{th} vertex, respectively. Minimizing Eq. 7 encourages those vertices that experience large contact forces to have smaller velocity and be closer to the ground.

Euler-Lagrange Loss. Furthermore, we incorporate a loss derived from the Euler-Lagrange equations to ensure the reconstructed motion adheres to the physics equations:

$$\mathcal{L}_{euler} = \textstyle\sum_{t=1}^{T} \|\mathbf{M}_t\ddot{\mathbf{q}}_t + \mathbf{C}_t + \mathbf{g}_t - \mathbf{J}_{C,t}^T\bar{\boldsymbol{\lambda}}_t - \bar{\boldsymbol{\tau}}_t\|_1. \tag{8}$$

It is worth noting that all terms in the loss function are analytically computed and fully differentiable with respect to the model outputs thanks to the physics-based body model.

Total Model Training Loss. Combining all the physics-inspired losses with the reconstruction loss, we obtain the total training loss function:

$$\mathcal{L} = \mathcal{L}_{recon} + \mathcal{L}_{force} + \mathcal{L}_{contact} + \mathcal{L}_{euler}. \tag{9}$$

We utilize Eq. 9 to train the Transformer encoder-decoder backbone solely using motion capture data. Once the model

is trained, it is directly added on top of the kinematics-based model to obtain improved motion estimates and infer motion forces, without the need of model fine-tuning.

4. Experiment

Datasets. During training, we use AMASS [51], a collection of motion capture datasets featuring a diverse range of subjects and actions. For evaluation, we utilize the test set of Human3.6M [25] and 3DOH [102]. Human3.6M encompasses common activities such as walking and sitting down. In contrast to certain physics-based methods that focus solely on sequences involving interactions with the ground, we adhere to the standard protocol and evaluate our method on all actions. 3DOH includes sequences of human-object interactions, such as opening a box — representing a challenging testing setting with significant occlusions. Furthermore, we utilize PennAction [103] to demonstrate that our approach helps improve human action recognition. PennAction comprises over 2K online videos of 15 sports actions, such as baseball pitching and bowling.

Evaluation Metrics. We evaluate 3D reconstruction error (*Rec. Error*) and physical plausibility (*Phys. Plausibility*). *Rec. Error* includes the Mean Per-Joint Position Error (MJE in mm) and MJE after the Procrustes Alignment (P-MJE in mm). *Phys. Plausibility* involves metrics introduced by prior methods [31, 67, 93], including: (1) the average difference between the predicted and the ground truth acceleration (acceleration error ACCL in mm/frame^2); (2) the difference between the predicted and the ground truth joint velocity magnitude (velocity error VEL in mm/frame); (3) the average displacement between two adjacent frames of those in-contact vertices (foot sliding FS in mm); (4) the average distance to the ground of those mesh vertices below the ground (ground penetration GP in mm).

Implementation. The Transformer backbone consists of standard encoder and decoder layers, with 6 layers, 8 attention heads, and 1024 embedding dimensions. The model's input sequence length is 16, aligning with most existing methods. For efficient Transformer training [77], we initially use the ground truth to extract the output embeddings for 20 epochs, followed by an additional 5 epochs using the prediction. We employ the Adam optimizer [30] with a weight decay of 10^{-4}. The initial learning rate is 10^{-5} and decreases to 0.8 after every 5 epochs. The hyperparameters are empirically set as: $\gamma_q = 2e^3$, $\gamma_J = 1e^5$, $\gamma_\tau = 5$, $\gamma_\lambda = 1$, $\gamma_v = 100$, and $\gamma_z = 200$.

4.1. Comparison with State-of-the-Arts (SOTAs)

Improvements to Kinematics-based Methods. As seen in Tab. 1, PhysPT, significantly improves the physical plausibility of kinematics-based motion estimates. Whether they take images or video frames as input, the kinematics-based methods often struggle with physical plausibility. For example, CLIFF retains competitive per-frame reconstruction accuracy but provides poor performance on all the physical plausibility evaluation metrics. Applying PhysPT to CLIFF significantly enhances its physics plausibility. Notably, the acceleration error (ACCL) and foot skating (FS) are reduced by 83.8% and 68.7% respectively. Like PhysPT, PoseBert leverages a Transformer-encoder pre-trained on 3D motion capture data, but it does not consider physics. PoseBert reduces ACCL and VEL somewhat, but unlike PhysPT it fails to decrease the foot skating and ground penetration error. In Supp. D, we demonstrate that PhysPT also improves other kinematics-based models besides CLIFF (SPIN [33] and IPMAN [76]) and the improvements are consistently more significant than PoseBert.

Method	Physics Engine	Video Label	*Rec. Error*		*Phys. Plausibility*			
			MJE	P-MJE	ACCL	VEL	FS	GP
HybriK† [38]	-	-	55.4	**33.6**	-	-	-	-
*CLIFF [41]	-	-	**52.2**	36.8	15.4	6.8	8.3	9.3
VIBE [31]	-	-	61.3	43.1	15.2	25.5	15.1	12.6
*PoseBert [2]	-	-	54.9	37.5	5.0	4.0	10.0	12.8
GLoT [64]	-	-	67.0	46.3	3.6	-	-	-
PMCE [92]	-	-	53.5	37.7	3.1	-	-	-
PhysCap [67]	Yes	-	97.4	65.1	-	7.2	-	-
NeurPhys [68]	Yes	Yes	76.5	58.2	-	4.5	-	-
Xie *et al.* [84]	Yes	-	68.1	-	-	4.0	-	-
SimPoE [93]	Yes	Yes	56.7	41.6	6.7	-	3.4	1.6
NeurMoCon [24]	Yes	Yes	72.5	54.6	-	3.8	-	-
TrajOpt [20]	Yes	-	84	56	-	-	-	-
DiffPhy [19]	Yes	-	81.7	55.6	-	-	-	-
D&D† [39]	No	Yes	52.5	35.5	6.1	-	5.8	**1.5**
Huang *et al.* [26]	Yes	Yes	55.4	41.3	-	3.5	-	-
PhysPT (Ours)	**No**	**No**	52.7	36.7	**2.5** ↓83.8	**3.4** ↓50.0	**2.6** ↓68.7	**1.5** ↓83.9

Table 1. **Evaluation on Human3.6M.** Methods in the top block use image inputs, those in the middle use video inputs, and those in the bottom are physics-based. Current physics-based methods require 3D annotated videos ("Video Label") for training or adopt a optimization-based framework. Methods marked by † are evaluated on 3D joints computed from fitted body models [46] instead of the one provided in the original datasets. For those marked "*", the results are from their officially released models. All other results are taken from the respective papers. Evaluation of PoseBert and PhysPT is based on CLIFF. The green numbers represent percentages of the relative improvement of our approach over CLIFF. For all metrics, smaller values are preferred.

Advantages over Physics-based Methods. PhysPT surpasses existing physics-based methods without relying on physics engines or 3D annotated videos for training. As illustrated in Tab. 1, existing physics-based methods generally exhibit improved physical plausibility compared to kinematics-based methods. They typically employ a physics engine separate from their learning models to compute physical parameters and simulate body motion. They require 3D annotations paired with input videos for train-

Method	*Rec. Error*		*Phys. Plausibility*			
	MJE	P-MJE	ACCL	VEL	FS	GP
*CLIFF [41]	**53.0**	34.4	26.0	12.0	10.8	12.6
VIBE [31]	98.1	61.8	-	26.5	-	-
*PoseBert [2]	54.8	34.1	6.6	6.9	14.0	10.4
NeurPhys [68]	107.8	93.3	-	12.2	-	-
NeurMoCon [24]	93.4	86.7	-	9.2	-	-
Huang *et al.* [26]	79.3	72.8	-	8.9	-	-
PhysPT (Ours)	**53.0**	**33.3**	**4.6** ↓82.3	**6.5** ↓45.8	**4.7** ↓56.5	**0.1** ↓99.2

Table 2. **Evaluation on 3DOH.** Evaluation of PoseBert and PhysPT is based on CLIFF. The green numbers represent percentages of the relative improvement of our approach over CLIFF.

ing or are confined to optimization-based approaches. In contrast, our approach avoids the need for 3D annotated videos by exploiting an innovative self-supervised learning framework. We bridge the gap between body kinematics and physics through a physics-based body representation and contact force model, allowing the seamless integration of physics with deep models. Comparing the performance, our model achieves competitive reconstruction accuracy with more significant advancements in physical plausibility (Tab. 1). For instance, although all other methods utilize training data from Human3.6M, our approach yields an acceleration error that is 2.4 times less than that of its nearest competitor D&D (2.5 vs. 6.1 mm/frame^2), and foot skating that is 76% of second-best SimPoE's (2.6mm vs. 3.4mm). In Supp. E, we demonstrate that our approach produces improved global motion recovery as well.

Robustness under Occlusion. Our approach is robust to occlusion, as demonstrated through the evaluation on 3DOH (Tab. 2). Despite the significant occlusions and complex human-object interaction motions included in 3DOH, applying PhysPT to CLIFF produces consistent improvement on motion estimates. The final model performance outperforms existing physics-based methods. For example, our approach achieves a velocity error of 6.5 mm/frame (Tab. 2, VEL), 45.8% less than CLIFF's 12.0, and 27.0% less than the 8.9 attained by SOTA Huang *et al.*. Furthermore, on 3DOH, our approach surpasses existing physics-based methods in reconstruction accuracy by a large margin. Specifically, our approach achieves P-MJE of 33.5, 54.0% less than Huang *et al.*'s 72.8. Existing physics-based methods do not utilize 3DOH for training, and their performance degrades on new testing sequences. In contrast, our approach maintains better generalization ability and effectively leverages the favourable per-frame 3D body reconstruction of kinematics-based estimates to generate accurate and physical plausible motion.

Training Losses				*Rec. Error*		*Phys. Plausibility*			
$\mathcal{L}_{recon}$	$\mathcal{L}_{force}$	$\mathcal{L}_{contact}$	$\mathcal{L}_{euler}$	MJE	P-MJE	ACCL	VEL	FS	GP
-	-	-	-	52.2	36.8	15.4	6.8	8.3	9.3
✓				52.7	36.7	2.5	3.5	7.1	6.9
✓	✓			52.7	36.7	2.5	3.4	6.5	5.6
✓	✓	✓		53.0	36.8	2.5	3.4	4.1	1.7
✓	✓	✓	✓	52.7	36.7	2.5	3.4	2.6	1.5

Table 3. **Ablation on the Training Losses.** The evaluation is on Human3.6M. The first row denotes the kinematics-based model

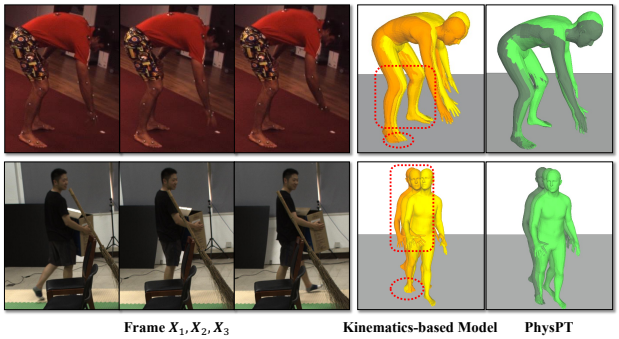

Figure 4. **Qualitative Evaluation on Utilizing PhysPT.** The body color of each figure represents the reconstruction at different time frames (lighter colors indicate later time frames). Ground penetration and motion jittering exhibited in the reconstructed motion are marked by red circle and rectangle, respectively.

4.2. Ablation Study

Effectiveness of PhysPT. We first study the effectiveness of the Transformer encoder-decoder backbone and the physics-inspired training losses. According to the evaluation results detailed in Tab. 3, when trained exclusively with the reconstruction loss (Eq. 3), the model maintains the reconstruction accuracy while reducing the acceleration and velocity errors of the kinematics-based estimates (Tab. 3-row2 over row1). The reduction is more significant than that observed in PoseBert (Tab. 1), demonstrating the advantages of leveraging the Transformer encoder-decoder rather than using the encoder solely. However, the foot skating and ground penetration errors are reduced to a much lesser extent. Reducing them significantly requires further imposing the physics-inspired losses. Specifically, when force labels are used for training, the foot skating error drops from 7.1mm to 6.5mm and the ground penetration error drops from 6.9mm to 5.6mm. Imposing the contact loss (Eq. 7) further reduces the errors but sacrifices the reconstruction accuracy (Tab. 3-row3). To obtain the best model, PhysPT, we leverage all the physics-inspired losses for training. In Fig. 4, we showcase that utilizing PhysPT effectively reduces the motion jitter and foot penetration exhibited by kinematics-based estimates. For example, the

Figure 5. **Qualitative Evaluation with Force Estimation Visualization.** The testing image frames are from Human3.6M (left), 3DOH (middle), and PennAction (right).

kinematics-based model can produce excessive motion jitter of the lower body even when only the upper body moves (Fig. 4-row1) or be affected by occlusion in the input video frame (Fig. 4-row2). By contrast, PhysPT resolves these issues by integrating physics with the Transformer.

Motion Reconstruction with Force Estimation. Our approach can generate accurate 3D motion and infer forces, as illustrated qualitatively in Fig. 5. The inferred forces offer valuable insights into body dynamic behavior. For instance, in the left column of Fig. 5, significant contact forces and joint actuations are evident on the left foot when the subject walks forward with the left foot (top) or on the left leg and when the body leans forward to the left (middle). The estimated forces also capture the contact behavior of various body parts, such as the knee (bottom). Moreover, our approach effectively handles occlusion (Fig. 5, 3DOH) and is applicable to in-the-wild videos (Fig. 5, PennAction).

4.3. Improvements to Human Action Recognition

In the preceding sections, we illustrate that PhysPT generates more physically-realistic motion and produces reasonable force estimates. The improved motion and additional force estimates can successfully improve downstream tasks such as human action recognition. We demonstrate this through the evaluation of human action recognition on PennAction. Specifically, we employ a recent skeleton-based recognition model proposed by [9]. We utilize the motion and forces generated by our approach as model input and evaluate the corresponding recognition accuracy. We summarize the evaluation results with comparison to recent skeleton-based recognition models in Tab. 4. In this comparison, we exclude the physics-based recognition model discussed in the related work ([52]) as it relies on 3D motion

Method	HDM [109]	UNIK [87]	Ours			
			$\mathbf{J}_{kin}$	$\mathbf{J}_{phys}$	**F**	$\mathbf{J}_{phys}$+**F**
Top-1 Acc. (↑)	93.4	94.0	96.0	96.8	94.4	**98.0**

Table 4. **Human Action Recognition.** The evaluation is on PennAction. We report the Top-1 accuracy in percentage. $\mathbf{J}_{kin}$ and $\mathbf{J}_{phys}$ stands for the 3D body joint positions estimated by the kinematics-based model and PhysPT, respectively. **F** indicates the motion forces output by PhysPT.

capture data and is not adept at handling in-the-wild videos. Conventional methods primarily rely on 2D body pose as model input. In contrast, our approach excels by leveraging 3D body pose information. Particularly, as shown in Tab. 4, utilizing the motion generated by PhysPT yields better performance compared to using kinematics-based estimates (96.8% over 96.0% in accuracy). Note that, using estimated forces alone, the accuracy is superior to that of existing methods (94.4% over UNIK's 94.0%), further demonstrating the effectiveness of our approach in estimating human dynamics. Finally, the combination of motion and force estimates leads to a significant performance boost, achieving the best recognition accuracy of 98.0%. In Supp. G, we provide action-wise evaluation, illustrating that utilizing forces enhances performance, particularly in cases where relying solely on 3D joint positions falls short, such as when different actions have similar body movement patterns.

5. Conclusion

In summary, we describe PhysPT (Physics-aware Pretrained Transformer), which generates more physically plausible motion estimates than previous methods and infers motion forces. PhysPT exploits a Transformer encoder-decoder backbone trained through self-supervised learning and it integrates physics principles. Specifically, we craft a physics-based body representation and a continuous contact force model. We introduce novel physics-inspired training losses. Leveraging them for model training enables PhysPT to effectively capture physical properties of the human body and the forces it experiences. Through extensive experiments, we demonstrate the direct applicability of PhysPT to kinematics-based estimates results in the reconstruction of more physically-realistic motion and the inference of motion forces from monocular videos. Notably, for the first time, we demonstrate that these more accurate estimates of motion and force translate to improvements in an important downstream task: human action recognition.

Acknowledgement This work is supported in part by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2

[2] Fabien Baradel, Romain Brégier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez. Posebert: A generic transformer module for temporal 3d human modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6, 7, 4

[3] Dan Biderman, Christian A Naesseth, Luhuan Wu, Taiga Abe, Alice C Mosberger, Leslie J Sibener, Rui Costa, James Murray, and John P Cunningham. Inverse articulated-body dynamics from video via variational sequential monte carlo. *NeurIPS 2020 Workshop*, 2021. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[5] Marcus A Brubaker, Leonid Sigal, and David J Fleet. Estimating contact dynamics. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2389–2396. IEEE, 2009. 2, 4

[6] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281, 2019. 1, 2

[7] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 1

[8] Biplab Ketan Chakraborty, Debajit Sarma, Manas Kamal Bhuyan, and Karl F MacDorman. Review of constraints on vision-based gesture recognition for human–computer interaction. *IET Computer Vision*, 12(1):3–15, 2018. 1

[9] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 8

[10] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. *arXiv preprint arXiv:2012.11806*, 2020. 1, 2

[11] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7649–7659, 2021. 1, 2

[12] Mia Chiquier and Carl Vondrick. Muscles in action. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22091–22101, 2023. 2

[13] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021. 1, 2

[14] Jessica Colombel, Vincent Bonnet, David Daney, Raphaël Dumas, Antoine Seilles, and François Charpillet. Physically consistent whole-body kinematics assessment based on an rgb-d sensor. application to simple rehabilitation exercises. *Sensors*, 20(10):2848, 2020. 1

[15] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016. 1

[16] Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016. 5

[17] Zackory Erickson, Vamsee Gangaram, Ariel Kapusta, C Karen Liu, and Charles C Kemp. Assistive gym: A physics simulation framework for assistive robotics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10169–10176. IEEE, 2020. 1

[18] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8781–8791, 2023. 2

[19] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13190–13200, 2022. 1, 2, 6

[20] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2022. 1, 6

[21] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2

[22] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10472–10481, 2021. 2

[23] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018. 1, 2

[24] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6417–6426, 2022. 1, 2, 6, 7

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian 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):1325–1339, 2013. 6

[26] Jingyi Ju, Buzhen Huang, Chen Zhu, Zhihao Li, and Yangang Wang. Physics-guided human motion capture with pose probability modeling. *arXiv preprint arXiv:2308.09910*, 2023. 2, 6, 7

[27] Michael Kallay. Computing the moment of inertia of a solid defined by a triangle mesh. *Journal of Graphics Tools*, 11 (2):51–57, 2006. 2

[28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1

[29] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 2

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6, 1

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 1, 2, 6, 7

[32] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. *arXiv preprint arXiv:2104.08527*, 2021. 2

[33] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 6, 4

[34] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021. 2

[35] Robotic Systems Lab. Robot dynamics lecture notes. *ETH Zurich*, 2017. 3

[36] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12375–12384, 2021. 1, 2

[37] Sergey Levine and Jovan Popović. Physically plausible simulation for character animation. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 221–230, 2012. 2

[38] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 1, 2, 6

[39] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, pages 479–496. Springer, 2022. 1, 2, 6, 4

[40] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8649, 2019. 1, 2

[41] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 2, 6, 7

[42] Ziwen Li, Bo Xu, Han Huang, Cheng Lu, and Yandong Guo. Deep two-stream video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 430–439, 2022. 2

[43] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 2

[44] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 2

[45] Wansong Liu, Xiao Liang, and Minghui Zheng. Dynamic model informed human motion prediction based on unscented kalman filter. *IEEE/ASME Transactions on Mechatronics*, 27(6):5287–5295, 2022. 2

[46] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6):220–1, 2014. 6

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 3

[48] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2

[49] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *arXiv preprint arXiv:2206.09106*, 2022. 1, 2

[50] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 534–543, 2023. 2

[51] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive

of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 1, 6, 5

[52] Al Mansur, Yasushi Makihara, and Yasushi Yagi. Inverse dynamics for action recognition. *IEEE transactions on cybernetics*, 43(4):1226–1236, 2012. 2, 8

[53] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 2

[54] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 1

[55] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 1, 2

[56] Stanley Plagenhoef, F Gaynor Evans, and Thomas Abdelnour. Anatomical data for analyzing human motion. *Research quarterly for exercise and sport*, 54(2):169–178, 1983. 4, 2

[57] Barry M Prior, Christopher M Modlesky, Ellen M Evans, Mark A Sloniger, Michael J Saunders, Richard D Lewis, and Kirk J Cureton. Muscularity and the density of the fat-free mass in athletes. *Journal of Applied Physiology*, 90(4): 1523–1531, 2001. 2

[58] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21254–21263, 2023. 2

[59] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. *arXiv preprint arXiv:2111.07868*, 2021. 2

[60] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020. 1, 2

[61] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. *arXiv preprint arXiv:2105.04668*, 2021. 1, 2

[62] Jesse Scott, Bharadwaj Ravichandran, Christopher Funk, Robert T Collins, and Yanxi Liu. From image to stability: learning dynamics from human pose. In *European Conference on Computer Vision*, pages 536–554. Springer, 2020. 2

[63] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so (3) manifolds for human pose and shape distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4779–4789, 2023. 2

[64] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8887–8896, 2023. 1, 2, 6

[65] Michael A Sherman, Ajay Seth, and Scott L Delp. Simbody: multibody dynamics for biomedical research. *Procedia Iutam*, 2:241–261, 2011. 2

[66] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14725–14737, 2023. 1

[67] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 1, 2, 6

[68] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 1, 2, 6, 7

[69] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. *arXiv preprint arXiv:2312.07531*, 2023. 1

[70] Mark W Spong and Mathukumalli Vidyasagar. *Robot dynamics and control*. John Wiley & Sons, 2008. 2

[71] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019. 2

[72] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8866, 2023. 2

[73] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 1, 2

[74] Theophilus Teo, Louise Lawrence, Gun A Lee, Mark Billinghurst, and Matt Adcock. Mixed reality remote collaboration combining 360 video and 3d reconstruction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14, 2019. 1

[75] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012. 1

[76] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4713–4725, 2023. 1, 6, 4

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4, 6

[78] Minh Vo, Yaser Sheikh, and Srinivasa G Narasimhan. Spatiotemporal bundle adjustment for dynamic 3d human reconstruction in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1066–1080, 2020. 2

[79] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021. 2

[80] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 1, 2

[81] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. *arXiv preprint arXiv:2303.13796*, 2023. 2

[82] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13211–13220, 2022. 2

[83] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 2

[84] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11532–11541, 2021. 1, 2, 6

[85] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 1

[86] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 1

[87] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580*, 2021. 8

[88] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. Ppr: Physically plausible reconstruction from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3914–3924, 2023. 1, 2

[89] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023. 2

[90] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. *arXiv preprint arXiv:2203.08528*, 2022. 2, 5

[91] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2023. 2

[92] Yingxuan You, Hong Liu, Ti Wang, Wenhao Li, Runwei Ding, and Xia Li. Co-evolution of pose and mesh for 3d human body estimation from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14963–14973, 2023. 2, 6

[93] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7159–7169, 2021. 1, 2, 6

[94] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 2, 3, 1, 4

[95] Petrissa Zell and Bodo Rosenhahn. Learning-based inverse dynamics of human motion. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 842–850, 2017. 2

[96] Petrissa Zell and Bodo Rosenhahn. Learning inverse dynamics for human locomotion analysis. *Neural Computing and Applications*, 32(15):11729–11743, 2020.

[97] Petrissa Zell, Bodo Rosenhahn, and Bastian Wandt. Weakly-supervised learning of human dynamics. In *European Conference on Computer Vision*, pages 68–84. Springer, 2020. 2

[98] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11436–11445, 2021. 1, 2

[99] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, pages 625–642. Springer, 2022. 2

[100] Boyang Zhang, Kehua Ma, Suping Wu, and Zhixiang Yuan. Two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5662–5670, 2023. 1, 2

[101] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body

capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11343–11353, 2021. 2

[102] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 6

[103] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 6

[104] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan Yuille. 3d-aware neural body fitting for occlusion robust 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9399–9410, 2023. 2

[105] Yufei Zhang, Hanjing Wang, Jeffrey O Kephart, and Qiang Ji. Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9020–9032, 2023. 1

[106] Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6164–6174, 2024. 2

[107] Zhibo Zhang, Yanjun Zhu, Rahul Rai, and David Doermann. Pimnet: Physics-infused neural network for human motion prediction. *IEEE Robotics and Automation Letters*, 7(4):8949–8955, 2022. 2

[108] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023. 2

[109] Rui Zhao, Wanru Xu, Hui Su, and Qiang Ji. Bayesian hierarchical dynamic model for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7733–7742, 2019. 8

[110] Yu Zheng and Katsu Yumane. Human motion tracking control with strict contact force constraints for floating-base humanoid robots, 2015. US Patent 9,120,227. 2

[111] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 2

[112] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: Unified pretraining for human motion analysis. *arXiv preprint arXiv:2210.06551*, 2022. 2