D&D: Learning Human Dynamics from Dynamic Camera

Jiefeng Li[1], Siyuan Bian[1], Chao Xu[2], Gang Liu[2], Gang Yu[2], and Cewu Lu[1]*

[1] Shanghai Jiao Tong University [2] Tencent
{ljf_likit,biansiyuan,lucewu}@sjtu.edu.cn
{dasxu,sylvainliu,skicyyu}@tencent.com

Abstract. 3D human pose estimation from a monocular video has recently seen significant improvements. However, most state-of-the-art methods are kinematics-based, which are prone to physically implausible motions with pronounced artifacts. Current dynamics-based methods can predict physically plausible motion but are restricted to simple scenarios with static camera view. In this work, we present D&D (Learning Human **D**ynamics from **D**ynamic Camera), which leverages the laws of physics to reconstruct 3D human motion from the in-the-wild videos with a moving camera. D&D introduces *inertial force control (IFC)* to explain the 3D human motion in the non-inertial local frame by considering the inertial forces of the dynamic camera. To learn the ground contact with limited annotations, we develop *probabilistic contact torque (PCT)*, which is computed by differentiable sampling from contact probabilities and used to generate motions. The contact state can be weakly supervised by encouraging the model to generate correct motions. Furthermore, we propose an attentive PD controller that adjusts target pose states using temporal information to obtain smooth and accurate pose control. Our approach is entirely neural-based and runs without offline optimization or simulation in physics engines. Experiments on large-scale 3D human motion benchmarks demonstrate the effectiveness of D&D, where we exhibit superior performance against both state-of-the-art kinematics-based and dynamics-based methods. Code is available at https://github.com/Jeff-sjtu/DnD.

Keywords: 3D Human Pose Estimation, Physical Awareness, Human Motion Dynamics

1 Introduction

Recovering 3D human pose and shape from a monocular image is a challenging problem. It has a wide range of applications in activity recognition [20, 21], character animation, and human-robot interaction. Despite the recent progress,

* Cewu Lu is the corresponding author, the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute

estimating 3D structure from the 2D observation is still an ill-posed and challenging task due to the inherent ambiguity.

A number of works [13, 14, 23, 6, 50, 62] turn to temporal input to incorporate body motion priors. Most state-of-the-art methods [12, 15, 14, 33, 19, 50, 56] are only based on kinematics modeling, i.e., body motion modeling with body part rotations and joint positions. Kinematics modeling directly captures the geometric information of the 3D human body, which is easy to learn by neural networks. However, methods that entirely rely on kinematics information are prone to physical artifacts, such as motion jitter, abnormal root actuation, and implausible body leaning.

Recent works [40, 44, 43, 59, 7] have started modeling human motion dynamics to improve the physical plausibility of the estimated motion. Dynamics modeling considers physical forces such as contact force and joint torque to control human motion. These physical properties can help analyze the body motion and understand human-scene interaction. Compared to widely adopted kinematics, dynamics gains less attention in 3D human pose estimation. The reason is that there are lots of limitations in current dynamics methods. For example, existing methods fail in daily scenes with dynamic camera movements (e.g., the 3DPW dataset [26]) since they require a static camera view, known ground plane and gravity vector for dynamics modeling. Besides, they are hard to deploy for real-time applications due to the need for highly-complex offline optimization or simulation with physics engines.

In this work, we propose a novel framework, D&D, a 3D human pose estimation approach with learned *Human **D**ynamics from **D**ynamic Camera*. Unlike previous methods that build the dynamics equation in the world frame, we redevise the dynamics equations in the non-inertial camera frame. Specifically, when the camera is moving, we introduce inertial forces in the dynamics equation to relate physical forces to local pose accelerations. We develop dynamics networks that directly estimate physical properties (forces and contact states). Then we can use the physical properties to compute the pose accelerations and obtain final human motion based on the accelerations. To train the dynamics network with only a limited amount of contact annotations, we propose *probabilistic contact torque (PCT)* for differentiable contact torque estimation. Concretely, we use a neural network to predict contact probabilities and conduct differentiable sampling to draw contact states from the predicted probabilities. Then we use the sampled contact states to compute the torque of the ground reaction forces and control the human motion. In this way, the contact classifier can be weakly supervised by minimizing the difference between the generated human motion and the ground-truth motion. To further improve the smoothness of the estimated motion, we propose a novel control mechanism called *attentive PD controller*. The output of the conventional PD controller [44, 44, 59] is proportional to the distance of the current pose state from the target state, which is sensitive to the unstable and jittery target. Instead, our attentive PD controller allows accurate control by globally adjusting the target state and is robust to the jittery target.

We benchmark D&D on the 3DPW [26] dataset captured with moving cameras and the Human3.6M [9] dataset captured with static cameras. D&D is compared against both state-of-the-art kinematics-based and dynamics-based methods and obtains state-of-the-art performance.

The contributions of this paper can be summarized as follows:

- We present the idea of inertial force control (IFC) to perform dynamics modeling for 3D human pose estimation from a dynamic camera view.
- We propose probabilistic contact torque (PCT) that leverages large-scale motion datasets without contact annotations for weakly-supervised training.
- Our proposed attentive PD controller enables smooth and accurate character control against jittery target motion.
- Our approach outperforms previous state-of-the-art kinematics-based and dynamics-based methods. It is fully differentiable and runs without offline optimization or simulation in physics engines.

2 Related Work

Kinematics-based 3D Human Pose Estimation. Numerous prior works estimate 3D human poses by locating the 3D joint positions [2, 35, 54, 34, 8, 27, 46, 37, 41, 28, 63, 30, 47, 32, 51, 61, 18]. Although these methods obtain impressive performance, they cannot provide physiological and physical constraints of the human pose. Many works [5, 16, 12, 15, 33, 19] adopt parametric statistical human body models [22, 36, 52] to improve physiological plausibility since they provide a well-defined human body structure. Optimization-based approaches [5, 16, 49, 36] automatically fit the SMPL body model to 2D observations, e.g., 2D keypoints and silhouettes. Alternatively, learning-based approaches use a deep neural network to regress the pose and shape parameters directly [12, 15, 14, 33, 19]. Several works [16, 15, 11] combine the optimization-based and learning-based methods to produce pseudo supervision or conduct test-time optimization.

For better temporal consistency, recent works have started to exploit temporal context [31, 13, 4, 48, 29, 14, 6, 50, 39]. Kocabas et al. [14] propose an adversarial framework to leverage motion prior from large-scale motion datasets [25]. Sun et al. [48] model temporal information with a bilinear Transformer. Rempe et al. [39] propose a VAE model to learn motion prior and optimize ground contacts. All the aforementioned methods disregard human dynamics. Although they achieve high accuracy on pose metrics, e.g., Procrustes Aligned MPJPE, the resulting motions are often physically implausible with pronounced physical artifacts such as improper balance and inaccurate body leaning.

Dynamics-based 3D Human Pose Estimation. To reduce physical artifacts, a number of works leverage the laws of physics to estimate human motion [57, 40, 44, 43, 59, 7]. Some of them are optimization-based approaches [40, 44, 7]. They use trajectory optimization to obtain the physical forces that induce

human motion. Shimada et al. [44] consider a complete human motion dynamics equation for optimization and obtain motion with fewer artifacts. Dabral et al. [7] propose a joint 3D human-object optimization framework for human motion capture and object trajectory estimation. Recent works [60, 43] have started to use regression-based methods to estimate human motion dynamics. Shimada et al. [43] propose a fully-differentiable framework for 3D human motion capture with physics constraints. All previous approaches require a static camera, restricting their applications in real-world scenarios.

On the other hand, deep reinforcement learning and motion imitation are widely used for 3D human motion estimation [57, 38, 59, 24, 55]. These works rely on physics engines to learn the control policies. Peng et al. [38] propose a control policy that allows a simulated character to mimic realistic motion capture data. Yuan et al. [59] present a joint kinematics-dynamics reinforcement learning framework that learns motion policy to reconstruct 3D human motion. Luo et al. [24] propose a dynamics-regulated training procedure for egocentric pose estimation. The work of Yu et al. [55] is most related to us. They propose a policy learning algorithm with a scene fitting process to reconstruct 3D human motion from a dynamic camera. Training their RL model and the fitting process is time-consuming. It takes 24 $\sim$ 96 hours to obtain the human motion of one video clip. Unlike previous methods, our regression-based approach is fully differentiable and does not rely on physics engines and offline fitting. It predicts accurate and physically plausible 3D human motion for in-the-wild scenes with dynamic camera movements.

3 Method

The overall framework of the proposed D&D (*Learning Human **D**ynamics from **D**ynamic Camera*) is summarized in Fig. 1. The input to D&D is a video $\{\mathbf{I}^t\}_{t=1}^T$ with T frames. Each frame $\mathbf{I}^t$ is fed into the kinematics backbone network to estimate the initial human motion $\widehat{q}^t$ in the local camera frame. The dynamics networks take as input the initial local motion $\{\widehat{q}^t\}_{t=1}^T$ and estimate physical properties (forces and contact states). Then we apply the forward dynamics modules to compute the pose and trajectory accelerations from the estimated physical properties. Finally, we use accelerations to obtain 3D human motion with physical constraints iteratively.

In this section, before introducing our solution, we first review the formulation of current dynamics-based methods in §3.1. In §3.2, we present the formulation of *Inertial Force Control (IFC)* that introduces inertial forces to explain the human motion in the dynamic camera view. Then we elaborate on the pipeline of D&D: **i**) learning physical properties with neural networks in §3.3, **ii**) analytically computing accelerations with forward dynamics in §3.4, **iii**) obtaining final pose with the constrained update in §3.5. The objective function of training the entire framework is further detailed in §3.6.

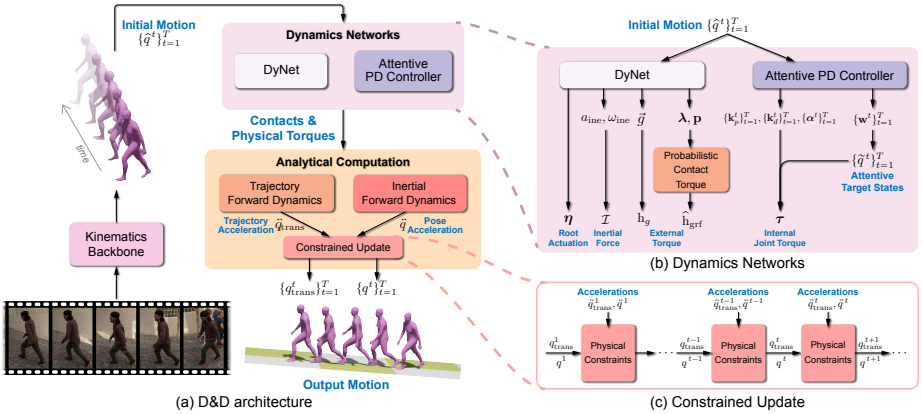

Fig. 1. Overview of the proposed framework. The video clip is fed into the kinematics backbone network to estimate the initial motion $\{\widehat{q}^t\}_{t=1}^T$. The dynamics networks take as input the initial motion and estimate physical properties. Then we compute pose and trajectory accelerations analytically via forward dynamics. Finally, we utilize accelerations to obtain human motion with physical constraints.

3.1 Preliminaries

The kinematic state of the human body can be represented by a pose vector q. The pose vector is parameterized as Euler angles with root translation and orientation to build the dynamics formula, i.e., $q \in \mathbb{R}^{3N_j+6}$, where N_j denotes the total number of human body joints. In previous methods [44, 58, 43], the first six entries of q are set as the root translation and orientation in the *world frame*. All remaining $3N_j$ entries encode joint angles of the human body. The laws of physics are imposed by considering Newtonian rigid body dynamics as:

$$\mathrm{M}(q)\ddot{q} - \boldsymbol{\tau} = \mathrm{h}_{\mathrm{grf}}(q, \mathbf{b}, \boldsymbol{\lambda}) - \mathrm{h}_g(q, \dot{q}) - \mathrm{h}_c(q, \dot{q}), \tag{1}$$

where $\mathrm{M} \in \mathbb{R}^{(3N_j+6)\times(3N_j+6)}$ denotes the inertia matrix of the human body; $\dot{q} \in \mathbb{R}^{3N_j+6}$ and $\ddot{q} \in \mathbb{R}^{3N_j+6}$ denote the velocity and the acceleration of q, respectively; $\mathrm{h}_{\mathrm{grf}} \in \mathbb{R}^{3N_j+6}$ denotes the resultant torque of the ground reaction forces; $\mathbf{b} \in \mathbb{R}^{N_c}$ is the discrete contact states vector; $\boldsymbol{\lambda} \in \mathbb{R}^{3N_c}$ is the linear contact forces; N_c denotes the number of joints to which the contact forces are applied; $\mathrm{h}_g \in \mathbb{R}^{3N_j+6}$ is the gravity torque; $\mathrm{h}_c \in \mathbb{R}^{3N_j+6}$ encompasses Coriolis and centripetal forces; $\boldsymbol{\tau} \in \mathbb{R}^{3N_j+6}$ represents the internal joint torque of the human body, with the first six entries being the direct root actuation. In this formulation, the translation and orientation must be in the static world frame, which restricts the application of the model in real-world scenarios. Therefore, previous dynamics-based methods are not applicable in current in-the-wild datasets with moving cameras, e.g., the 3DPW dataset [26].

3.2 Inertial Force Control

In this work, to facilitate in-the-wild 3D human pose estimation with physics constraints, we reformulate the dynamics equation to impose the laws of physics in the dynamic-view video. When the camera is moving, the local frame is an inertial frame of reference. In order to satisfy the force equilibrium, we introduce the inertial force $\mathcal{I}$ in the dynamics system:

$$\mathrm{M}(q)\ddot{q} - \boldsymbol{\tau} = \mathrm{h}_{\mathrm{grf}}(q, \mathbf{b}, \boldsymbol{\lambda}) - \mathrm{h}_g(q, \dot{q}) - \mathrm{h}_c(q, \dot{q}) + \mathcal{I}(q, \dot{q}, a_{\mathrm{ine}}, \omega_{\mathrm{ine}}), \tag{2}$$

where the first six entries of q are set as the root translation and orientation in the *local camera frame*, and the inertial force $\mathcal{I}$ is determined by the current motion state (q and $\dot{q}$) and camera movement state (linear acceleration $a_{\mathrm{ine}} \in \mathbb{R}^3$ and angular velocity $\omega_{\mathrm{ine}} \in \mathbb{R}^3$). Specifically, the inertial force encompasses linear, centripetal, and Coriolis forces. It is calculated as follows:

$$\mathcal{I} = \sum_{i}^{N_j} \underbrace{m_i J_{v_i}^{\mathsf{T}} a_{\mathrm{ine}}}_{\text{linear force}} + \underbrace{m_i J_{v_i}^{\mathsf{T}} \omega_{\mathrm{ine}} \times (\omega_{\mathrm{ine}} \times r_i)}_{\text{centripetal force}} + \underbrace{2 m_i J_{v_i}^{\mathsf{T}} (\omega_{\mathrm{ine}} \times v_i)}_{\text{Coriolis force}}, \tag{3}$$

where $J_{v_i} \in \mathbb{R}^{3\times(3N_j+6)}$ denotes the linear Jacobian matrix that describes how the linear velocity of the i-th joint changes with pose velocity $\dot{q}$, m_i denotes the mass of the i-th joint, r_i denotes the position of the i-th joint in the local frame, and v_i is the velocity of the i-th joint. J_{v_i}, r_i, and v_i can be analytically computed using the pose q and the velocity $\dot{q}$.

The inertial force control (IFC) establishes the relation between the physical properties and the pose acceleration in the local frame. The pose acceleration can be subsequently used to calculate the final motion. In this way, we can estimate physically plausible human motion from *forces* to *accelerations* to *poses*. The generated motion is smooth and natural. Besides, it provides extra physical information to understand human-scene interaction for high-level activity understanding tasks.

Discussion. The concept of *residual force* [58] is widely adopted in previous works [17, 3, 44, 58, 43] to explain the direct root actuation in the global static frame. Theoretically, we can adopt a residual term to explain the inertia in the local camera frame implicitly. However, we found explicit inertia modeling obtains better estimation results than implicit modeling with a residual term. Detailed comparisons are provided in §4.4.

3.3 Learning Physical Properties

In this subsection, we elaborate on the neural networks for physical properties estimation. We first use a kinematics backbone to extract the initial motion $\{\widehat{q}^t\}_{t=1}^T$. The initial motion is then fed to a dynamics network (DyNet) with *probabilistic contact torque* for contact, external force, and inertial force estimation and the *attentive PD controller* for internal joint torque estimation.

Contact, External Force, and Inertial Force Estimation. The root motion of the human character is dependent on external forces and inertial forces. To explain root motion, we propose DyNet that directly regresses the related physical properties, including the ground reaction forces $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_{N_c})$, the gravity $\boldsymbol{g}$, the direct root actuation $\boldsymbol{\eta}$, the contact probabilities $\mathbf{p} = (p_1, p_2, \cdots, p_{N_c})$, the linear camera acceleration a_{ine}, and the angular camera velocity ω_{ine}. The detailed network structure of DyNet is provided in the supplementary material.

The inertial force $\mathcal{I}$ can be calculated following Eqn. 3 with the estimated a_{ine} and ω_{ine}. The gravity torque h_g can be calculated as:

$$\mathrm{h}_g = -\sum_{i}^{N_j} m_i J_{v_i}^{\mathsf{T}} \boldsymbol{g}. \tag{4}$$

When considering gravity, bodyweight will affect human motion. In this paper, we let the shape parameters $\boldsymbol{\beta}$ control the body weight. We assume the standard weight is 75kg when $\boldsymbol{\beta}_0 = \mathbf{0}$, and there is a linear correlation between the body weight and the bone length. We obtain the corresponding bodyweight based on the bone-length ratio of $\boldsymbol{\beta}$ to $\boldsymbol{\beta}_0$.

Probabilistic Contact Torque: For the resultant torque of the ground reaction forces, previous methods [44, 58, 43] compute it with the discrete contact states $\mathbf{b} = (b_1, b_2, \cdots, b_{N_c})$ of N_c joints:

$$\mathrm{h}_{\text{grf}}(q, \mathbf{b}, \boldsymbol{\lambda}) = \sum_{j}^{N_c} b_j J_{v_j}^{\mathsf{T}} \lambda_j, \tag{5}$$

where $b_j = 1$ for contact and $b_j = 0$ for non-contact. Note that the output probabilities $\mathbf{p}$ are continuous. We need to discretize p_j with a threshold of 0.5 to obtain b_j. However, the discretization process is not differentiable. Thus the supervision signals for the contact classifier only come from a limited amount of data with contact annotations.

To leverage the large-scale motion dataset without contact annotations, we propose *probabilistic contact torque (PCT)* for weakly-supervised learning. During training, PCT conducts differentiable sampling [10] to draw a sample $\widehat{\mathbf{b}}$ that follows the predicted contact probabilities $\mathbf{p}$ and computes the corresponding ground reaction torques:

$$\widehat{\mathrm{h}}_{\text{grf}}(q, \widehat{\mathbf{b}}, \boldsymbol{\lambda}) = \sum_{j}^{N_c} \widehat{b}_j J_{v_j}^{\mathsf{T}} \lambda_j = \sum_{j}^{N_c} \frac{p_j e^{g_{j1}}}{p_j e^{g_{j1}} + (1 - p_j) e^{g_{j2}}} J_{v_j}^{\mathsf{T}} \lambda_j, \tag{6}$$

where $g_{j1}, g_{j2} \sim \text{Gumbel}(0, 1)$ are i.i.d samples drawn from the Gumbel distribution. When conducting forward dynamics, we use the sampled torque $\widehat{\mathrm{h}}_{\text{grf}}(q, \widehat{\mathbf{b}}, \boldsymbol{\lambda})$ instead of the torque $\mathrm{h}_{\text{grf}}(q, \mathbf{b}, \boldsymbol{\lambda})$ from the discrete contact states $\mathbf{b}$. To generate accurate motion, DyNet is encouraged to predict higher probabilities for the correct contact states so that PCT can sample the correct states as much as possible. Since PCT is differentiable, the supervision signals for the physical force and contact can be provided by minimizing the motion error. More details of differentiable sampling are provided in the supplementary material.

Internal Joint Torque Estimation. Another key process to generate human motions is internal joint torque estimation. PD controller is widely adopted for physics-based human motion control [44, 43, 59]. It controls the motion by outputting the joint torque $\boldsymbol{\tau}$ in proportion to the difference between the current state and the target state. However, the target pose states estimated by the kinematics backbone are noisy and contain physical artifacts. Previous works [43, 59] adjust the gain parameters dynamically for smooth motion control. However, we find that this local adjustment is still challenging for the model and the output motion is still vulnerable to the jittery and incorrect input motion.

Attentive PD Controller: To address this problem, we propose the attentive PD controller, a method that allows global adjustment of the target pose states. The attentive PD controller is fed with initial motion $\{\widehat{q}^t\}_{t=1}^T$ and dynamically predicts the proportional parameters $\{\mathbf{k}_p^t\}_{t=1}^T$, derivative parameters $\{\mathbf{k}_d^t\}_{t=1}^T$, offset torques $\{\boldsymbol{\alpha}^t\}_{t=1}^T$, and attention weights $\{\mathbf{w}^t\}_{t=1}^T$. The attention weights $\mathbf{w}^t = (w^{t1}, w^{t2}, \cdots, w^{tT})$ denotes how the initial motion contributes to the target pose state at the time step t and $\sum_{j=1}^T w^{tj} = 1$. We first compute the attentive target pose state $\widetilde{q}^t$ as:

$$\widetilde{q}^t = \sum_{j=1}^{T} w^{tj} \widehat{q}^j, \tag{7}$$

where $\widehat{q}^j$ is the initial kinematic pose at the time step j. Then the internal joint torque $\boldsymbol{\tau}^t$ at the time step t can be computed following the PD controller rule with the compensation term h_c^t [53]:

$$\boldsymbol{\tau}^t = \mathbf{k}_p^t \circ (\widetilde{q}^{t+1} - q^t) - \mathbf{k}_d^t \circ \dot{q}^t + \boldsymbol{\alpha}^t + \mathrm{h}_c^t, \tag{8}$$

where $\circ$ denotes Hadamard matrix product and h_c^t represents the sum of centripetal and Coriolis forces at the time step t. This attention mechanism allows the PD controller to leverage the temporal information to refine the target state and obtain a smooth motion. Details of the network structure are provided in the supplementary material.

3.4 Forward Dynamics

To compute the accelerations analytically from physical properties, we build two forward dynamics modules: *inertial forward dynamics* for the local pose acceleration and *trajectory forward dynamics* for the global trajectory acceleration.

Inertial Forward Dynamics. Prior works [44, 43, 59] adopt a proxy model to simulate human motion in physics engines or simplify the optimization process. In this work, to seamlessly cooperate with the kinematics-based backbone, we directly build the dynamics equation for the SMPL model [22]. The pose acceleration $\ddot{q}$ can be derived by rewriting Eqn. 2 with PCT:

$$\ddot{q} = \mathrm{M}^{-1}(q)(\boldsymbol{\tau} + \widehat{\mathrm{h}}_{\mathrm{grf}} - \mathrm{h}_g - \mathrm{h}_c + \mathcal{I}). \tag{9}$$

To obtain $\ddot{q}$, we need to compute the inertia matrix M and other physical torques in each time step using the current pose q. The time superscript t is omitted for simplicity. M can be computed recursively along the SMPL kinematics tree. The derivation is provided in the supplementary material.

Trajectory Forward Dynamics. To train DyNet without ground-truth force annotations, we leverage a key observation: the gravity and ground reaction forces should explain the global root trajectory. We devise a trajectory forward dynamics module that controls the global root motion with external forces. It plays a central role in the success of weakly supervised learning.

Let q_{trans} denote the root translation in the *world frame*. The dynamics equation can be written as:

$$\ddot{q}_{\text{trans}} = \frac{1}{m_0} R_{\text{cam}}^{\mathsf{T}} (\boldsymbol{\eta} + \widehat{\mathrm{h}}_{\text{grf}}^{\{0:3\}} - \mathrm{h}_{g}^{\{0:3\}}), \tag{10}$$

where m_0 is the mass of the root joint, R_{cam} denotes the camera orientation computed from the estimated angular velocity $\omega_{\text{ine}} = (\omega_x, \omega_y, \omega_z)$, $\boldsymbol{\eta}$ denotes the direct root actuation, and $\widehat{\mathrm{h}}_{\text{grf}}^{\{0:3\}}$ and $\mathrm{h}_{\text{g}}^{\{0:3\}}$ denote the first three entries of $\widehat{\mathrm{h}}_{\text{grf}}$ and h_{g}, respectively.

3.5 Constrained Update

After obtaining the pose and trajectory accelerations via forward dynamics modules, we can control the human motion and global trajectory by discrete simulation. Given the frame rate $1/\Delta t$ of the input video, we can obtain the kinematic 3D pose using the finite differences:

$$\dot{q}^{t+1} = \dot{q}^{t} + \Delta t\, \ddot{q}^{t}, \tag{11}$$

$$q^{t+1} = q^{t} + \Delta t\, \dot{q}^{t}. \tag{12}$$

Similarly, we can obtain the global root trajectory:

$$\dot{q}_{\text{trans}}^{t+1} = \dot{q}_{\text{trans}}^{t} + \Delta t\, \ddot{q}_{\text{trans}}^{t}, \tag{13}$$

$$q_{\text{trans}}^{t+1} = q_{\text{trans}}^{t} + \Delta t\, \dot{q}_{\text{trans}}^{t}. \tag{14}$$

In practice, since we predict the local and global motions simultaneously, we can impose contact constraints to prevent foot sliding. Therefore, instead of using Eqn. 12 and 14 to update q^{t+1} and q_{trans}^{t+1} directly, we first refine the velocities $\dot{q}^{t+1}$ and $\dot{q}_{\text{trans}}^{t+1}$ with contact constraints. For joints in contact with the ground at the time step t, we expect they have zero velocity in the world frame. The velocities of non-contact joints should stay close to the original velocities computed from the accelerations. We adopt the differentiable optimization layer following the formulation of Agrawal et al. [1]. This custom layer can obtain

the solution to the optimization problem and supports backward propagation. However, the optimization problem with zero velocity constraints does not satisfy the DPP rules (Disciplined Parametrized Programming), which means that the custom layer cannot be directly applied. Here, we use soft velocity constraints to follow the DPP rules:

$$\begin{aligned} \dot{q}^*, \dot{q}^*_{\text{trans}} = & \underset{\dot{q}^*, \dot{q}^*_{\text{trans}}}{\operatorname{argmin}} \|\dot{q}^* - \dot{q}\| + \|\dot{q}^*_{\text{trans}} - \dot{q}_{\text{trans}}\|, \\ & s.t. \quad \forall i \in \{i | p_i > 0.5\}, \ \|R^{\mathsf{T}}_{\text{cam}}(J_{v_i}\dot{q}^* - \dot{q}^{*\{0:3\}}) + \dot{q}^*_{\text{trans}}\| \leq \epsilon, \end{aligned} \tag{15}$$

where $\epsilon = 0.01$ and $\dot{q}^{*\{0:3\}}$ is the first three entries of $\dot{q}^*$. We omit the superscript t for simplicity. After solving Eqn. 15, the estimated $\dot{q}^*$ and $\dot{q}^*_{\text{trans}}$ are used to compute the final physically-plausible 3D pose q and the global trajectory q_{trans}.

3.6 Network Training

The overall loss of D&D is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{reg}}. \tag{16}$$

The 3D loss $\mathcal{L}_{\text{3D}}$ includes the joint error and the pose error:

$$\mathcal{L}_{\text{3D}} = \|X - \breve{X}\|_1 + \|q \ominus \breve{q}\|_2^2, \tag{17}$$

where X denotes the 3D joints regressed from the SMPL model, "$\ominus$" denotes a difference computation after converting the Euler angle into a rotation matrix, and the symbol "$\breve{}$" denotes the ground truth. The time superscript t is omitted for simplicity. The 2D loss $\mathcal{L}_{\text{2D}}$ calculates the 2D reprojection error:

$$\mathcal{L}_{\text{2D}} = \|\Pi(X) - \Pi(\breve{X})\|_1, \tag{18}$$

where Π denotes the projection function. The loss $\mathcal{L}_{\text{trans}}$ is added for the supervision of the root translation and provides weak supervision signals for external force and contact estimation:

$$\mathcal{L}_{\text{trans}} = \|q_{\text{trans}} - \breve{q}_{\text{trans}}\|_1. \tag{19}$$

The contact loss is added for the data with contact annotations:

$$\mathcal{L}_{\text{con}} = \frac{1}{N_c}\sum_{i}^{N_c} \big[-\breve{b}_i \log p_i - (1 - \breve{b}_i) \log (1 - p_i) \big]. \tag{20}$$

The regularization loss $\mathcal{L}_{\text{reg}}$ is defined as:

$$\mathcal{L}_{\text{reg}} = \|\boldsymbol{\eta}\|_2^2 + \frac{1}{N_c}\sum_{i}^{N_c} \big[-p_i \log p_i - (1 - p_i) \log (1 - p_i) \big], \tag{21}$$

where the first term minimizes the direct root actuation, and the second term minimizes the entropy of the contact probability to encourage confident contact predictions.

Table 1. Quantitative comparisons with state-of-the-art methods on the 3DPW dataset. Symbol "-" means results are not available, and "∗" means self-implementation.

Method	Dynamics	MPJPE ↓	PA-MPJPE ↓	PVE ↓	ACCEL ↓
HMR [12]	✗	130.0	81.3	-	37.4
SPIN [15]	✗	96.9	59.2	116.4	29.8
VIBE [14]	✗	82.9	51.9	99.1	23.4
TCMR [6]	✗	86.5	52.7	102.9	7.1
HybrIK* [19]	✗	76.2	45.1	89.1	22.8
MAED [50]	✗	79.1	45.7	92.6	17.6
Ours	✓	**73.7**	**42.7**	**88.6**	**7.0**

4 Experiment

4.1 Datasets

We perform experiments on two large-scale human motion datasets. The first dataset is 3DPW [26]. 3DPW is a challenging outdoor benchmark for 3D human motion estimation. It contains 60 video sequences obtained from a hand-held moving camera. The second dataset we use is Human3.6M [9]. Human3.6M is an indoor benchmark for 3D human motion estimation. It includes 7 subjects, and the videos are captured at 50Hz. Following previous works [15, 14, 19, 59], we use 5 subjects (S1, S5, S6, S7, S8) for training and 2 subjects (S9, S11) for evaluation. The videos are subsampled to 25Hz for both training and testing. We further use the AMASS dataset [25] to obtain annotations of foot contact and root translation for training.

4.2 Implementation Details

We adopt HybrIK [19] as the kinematics backbone to provide the initial motion. The original HybrIK network only predicts 2.5D keypoints and requires a separate RootNet [32] to obtain the final 3D pose in the camera frame. Here, for integrity and simplicity, we implement an extended version of HybrIK as our kinematics backbone that can directly predict the 3D pose in the camera frame by estimating the camera parameters. The network structures are detailed in the supplementary material. The learning rate is set to 5×10^{-5} at first and reduced by a factor of 10 at the 15th and 25th epochs. We use the Adam solver and train for 30 epochs with a mini-batch size of 32. Implementation is in PyTorch. During training on the Human3.6M dataset, we simulate a moving camera by cropping the input video with bounding boxes.

4.3 Comparison to state-of-the-art methods

Results on Moving Camera We first compare D&D against state-of-the-art methods on 3DPW, an in-the-wild dataset captured with the hand-held

Table 2. Quantitative comparisons with state-of-the-art methods on the Human3.6M dataset. Symbol "-" means results are not available, "∗" means self-implementation, and "†" means the method reports results on 17 joints.

Method	Dynamics	MPJPE ↓	PA-MPJPE ↓	PVE ↓	ACCEL ↓	FS ↓	GP ↓
VIBE [14]	✗	61.3	43.1	-	15.2	15.1	12.6
NeurGD [45]	✗	57.3	42.2	-	14.2	16.7	24.4
MAED† [50]	✗	56.3	38.7	-	-	-	-
HybrIK* [19]	✗	56.4	36.7	-	10.9	18.3	10.6
PhysCap [44]	✓	113.0	68.9	-	-	-	-
EgoPose [57]	✓	130.3	79.2	-	31.3	5.9	3.5
NeurPhys [43]	✓	76.5	-	-	-	-	-
SimPoE [59]	✓	56.7	41.6	-	6.7	**3.4**	1.6
Ours	✓	**52.5**	**35.5**	**72.9**	**6.1**	5.8	**1.5**

moving camera. Since previous dynamics-based methods are not applicable in the moving camera, prior arts on the 3DPW dataset are all kinematics-based. Mean per joint position error (MPJPE) and Procrustes-aligned mean per joint position error (PA-MPJPE) are reported to assess the 3D pose accuracy. The acceleration error (ACCEL) is reported to assess the motion smoothness. We also report Per Vertex Error (PVE) to evaluate the entire estimated body mesh.

Tab. 1 summarizes the quantitative results. We can observe that D&D outperforms the most accurate kinematics-based methods, HybrIK and MAED, by 2.5 and 5.4 mm on MPJPE, respectively. Besides, D&D improves the motion smoothness significantly by 69.3% and 60.2% relative improvement on ACCEL, respectively. It shows that D&D retains the benefits of the accurate pose in kinematics modeling and physically plausible motion in dynamics modeling.

Results on Static Camera To compare D&D with previous dynamics-based methods, we evaluate D&D on the Human3.6M dataset. Following the previous method [59], we further report two physics-based metrics, foot sliding (FS) and ground penetration (GP), to measure the physical plausibility. To assess the effectiveness of IFC, we simulate a moving camera by cropping the input video with bounding boxes, i.e., the input to D&D is the video from a moving camera. Tab. 2 shows the quantitative comparison against kinematics-based and dynamics-based methods. D&D outperforms previous kinematics-based and dynamics-based methods in pose accuracy. For physics-based metrics (ACCEL, FS, and GP), D&D shows comparable performance to previous methods that require physics simulation engines.

We further follow GLAMR [56] to evaluate the global MPJPE (G-MPJPE) and global PVE (G-PVE) on the Human3.6M dataset with the simulated moving camera. The root translation is aligned with the GT at the first frame of the

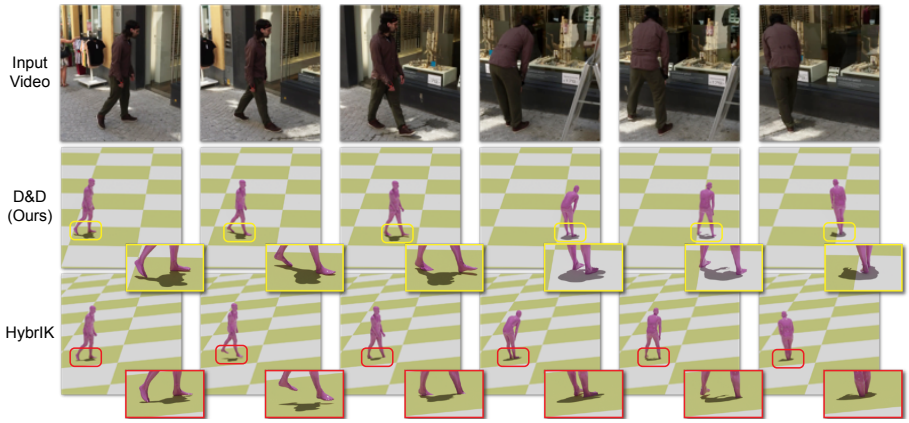

Fig. 2. Qualitative comparisons on the 3DPW dataset. D&D estimates accurate poses with physically plausible foot contact and global movement.

Table 3. Ablation experiments on 3DPW and Human3.6M dataset.

	3DPW			Human3.6M		
	MPJPE ↓	PA-MPJPE ↓	ACCEL ↓	MPJPE ↓	PA-MPJPE ↓	ACCEL ↓
w/o IFC	76.0	45.2	10.0	53.8	36.4	6.7
w/o PCT	74.6	43.4	9.8	53.4	36.1	6.7
w/o Att PD Controller	73.8	42.8	8.0	52.5	35.7	6.3
D&D (Ours)	**73.7**	**42.7**	**7.0**	**52.5**	**35.5**	**6.1**

video sequence. D&D obtains 785.1mm G-MPJPE and 793.3mm G-PVE. More comparisons are reported in the supplementary material.

4.4 Ablation Study

Inertial Force *vs.* Residual Force. In this experiment, we compare the proposed inertial force control (IFC) with residual force control (RFC). To control the human motion with RFC in the local camera frame, we directly estimate the residual force instead of the linear acceleration and angular velocity. Quantitative results are reported in Tab. 3. It shows that explicit modeling of the inertial components can better explain the body movement than implicit modeling with residual force. IFC performs more accurate pose control and significantly reduces the motion jitters, showing a 30% relative improvement of ACCEL on 3DPW.

Effectiveness of PCT. To study the effectiveness of the probabilistic contact torque, we remove PCT in the baseline model. When training the baseline, the output contact probabilities are discretized to 0 or 1 with the threshold of 0.5

and we compute the discrete contact torque instead of the probabilistic contact torque. Quantitative results in Tab. 3 show that PCT is indispensable to have smooth and accurate 3D human motion.

Effectiveness of Attentive PD Controller. To further validate the effectiveness of the attentive mechanism, we report the results of the baseline model without the attentive PD controller. In this baseline, we adopt the meta-PD controller [59, 43] that dynamically predicts the gain parameters based on the state of the character, which only allows local adjustment. Tab. 3 summarizes the quantitative results. The attentive PD controller contributes to a more smooth motion control as indicated by a smaller acceleration error.

4.5 Qualitative Results

In Fig. 3, we plot the contact forces estimated by D&D of the walking motion from the Human3.6M test set. Note that our approach does not require any ground-truth force annotations for training. The estimated forces fall into a reasonable force range for walking motions [42]. We also provide qualitative comparisons in Fig. 2. It shows that D&D can estimate physically plausible motions with accurate foot-ground contacts and no ground penetration.

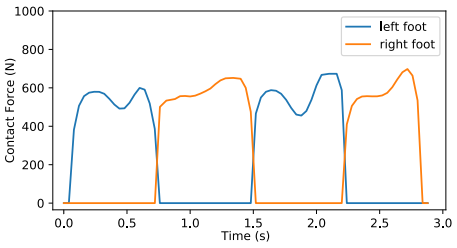

Fig. 3. Estimated contact forces of the walking sequences. The forces remain in a reasonable range for walking.

5 Conclusion

In this paper, we propose D&D, a physics-aware framework for 3D human motion capture with dynamic camera movements. To impose the laws of physics in the moving camera, we introduce inertial force control that explains the 3D human motion by taking the inertial forces into consideration. We further develop the probabilistic contact torque for weakly-supervised training and the attentive PD controller for smooth and accurate motion control. We demonstrate the effectiveness of our approach on standard 3D human pose datasets. D&D outperforms state-of-the-art kinematics-based and dynamics-based methods. Besides, it is entirely neural-based and runs without offline optimization or physics simulators. We hope D&D can serve as a solid baseline and provide a new perspective for dynamics modeling in 3D human motion capture.

Acknowledgments. This work was supported by the National Key R&D Program of China (No. 2021ZD0110700), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, SHEITC (2018-RGZN-02046) and Tencent GY-Lab.

References

1. Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., Kolter, J.Z.: Differentiable convex optimization layers. NeurIPS (2019) 9
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: CVPR (2015) 3
3. Andrews, S., Huerta, I., Komura, T., Sigal, L., Mitchell, K.: Real-time physics-based motion capture with sparse sensors. In: CVMP (2016) 6
4. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3d human pose estimation in the wild. In: CVPR (2019) 3
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV (2016) 3
6. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: CVPR (2021) 2, 3, 11
7. Dabral, R., Shimada, S., Jain, A., Theobalt, C., Golyanik, V.: Gravity-aware monocular 3d human-object reconstruction. In: ICCV (2021) 2, 3, 4
8. Fang, H., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. AAAI (2017) 3
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI (2013) 3, 11
10. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017) 7
11. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: 3DV (2021) 3
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018) 2, 3, 11
13. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019) 2, 3
14. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (2020) 2, 3, 11, 12
15. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019) 2, 3, 11
16. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR (2017) 3
17. Levine, S., Popović, J.: Physically plausible simulation for character animation. In: SIGGRAPH (2012) 6
18. Li, J., Chen, T., Shi, R., Lou, Y., Li, Y.L., Lu, C.: Localization with sampling-argmax. Advances in Neural Information Processing Systems **34**, 27236–27248 (2021) 3
19. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: CVPR (2021) 2, 3, 11, 12
20. Li, Y.L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., Lu, C.: Detailed 2d-3d joint representation for human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10166–10175 (2020) 1

21. Li, Y.L., Xu, L., Liu, X., Huang, X., Xu, Y., Wang, S., Fang, H.S., Ma, Z., Chen, M., Lu, C.: Pastanet: Toward human activity knowledge engine. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 382–391 (2020) 1
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. TOG (2015) 3, 8
23. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: ACCV (2020) 2
24. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. In: NeurIPS (2021) 4
25. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019) 3, 11
26. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018) 2, 3, 5, 11
27. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017) 3
28. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV (2017) 3
29. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. TOG (2020) 3
30. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 3DV (2018) 3
31. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. TOG (2017) 3
32. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: ICCV (2019) 3, 11
33. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: ECCV (2020) 2, 3
34. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: CVPR (2017) 3
35. Park, S., Hwang, J., Kwak, N.: 3d human pose estimation using convolutional neural networks with 2d pose information. In: ECCV (2016) 3
36. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019) 3
37. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR (2017) 3
38. Peng, X.B., Chang, M., Zhang, G., Abbeel, P., Levine, S.: Mcp: Learning composable hierarchical control with multiplicative compositional policies. arXiv preprint arXiv:1905.09808 (2019) 4
39. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: ICCV (2021) 3
40. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video. In: ECCV (2020) 2, 3

41. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR (2017) 3
42. Shahabpoor, E., Pavic, A.: Measurement of walking ground reactions in real-life environments: a systematic review of techniques and technologies. Sensors (2017) 14
43. Shimada, S., Golyanik, V., Xu, W., Pérez, P., Theobalt, C.: Neural monocular 3d human motion capture with physical awareness. TOG (2021) 2, 3, 4, 5, 6, 7, 8, 12, 14
44. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: Physcap: Physically plausible monocular 3d motion capture in real time. TOG (2020) 2, 3, 5, 6, 7, 8, 12
45. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: ECCV (2020) 12
46. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017) 3
47. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018) 3
48. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: ICCV (2019) 3
49. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: ECCV (2018) 3
50. Wan, Z., Li, Z., Tian, M., Liu, J., Yi, S., Li, H.: Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In: ICCV (2021) 2, 3, 11, 12
51. Wang, C., Li, J., Liu, W., Qian, C., Lu, C.: Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In: European Conference on Computer Vision. pp. 242–259. Springer (2020) 3
52. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: CVPR (2020) 3
53. Yang, C., Huang, Q., Jiang, H., Peter, O.O., Han, J.: Pd control with gravity compensation for hydraulic 6-dof parallel manipulator. Mechanism and Machine theory (2010) 8
54. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3d pose estimation from a single image. In: CVPR (2016) 3
55. Yu, R., Park, H., Lee, J.: Human dynamics from monocular video with dynamic camera movements. TOG (2021) 4
56. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: CVPR (2022) 2, 12
57. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: ICCV (2019) 3, 4, 12
58. Yuan, Y., Kitani, K.: Residual force control for agile human behavior imitation and extended motion synthesis. NeurIPS (2020) 5, 6, 7
59. Yuan, Y., Wei, S.E., Simon, T., Kitani, K., Saragih, J.: Simpoe: Simulated character control for 3d human pose estimation. In: CVPR (2021) 2, 3, 4, 8, 11, 12, 14
60. Zell, P., Rosenhahn, B., Wandt, B.: Weakly-supervised learning of human dynamics. In: ECCV (2020) 4
61. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: ECCV (2020) 3

62. Zeng, A., Yang, L., Ju, X., Li, J., Wang, J., Xu, Q.: Smoothnet: A plug-and-play network for refining human poses in videos. arXiv preprint arXiv:2112.13715 (2021) 2
63. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017) 3