Trajectory Optimization for Physics-Based Reconstruction of 3d Human Pose from Monocular Video

Erik Gärtner[1,2] Mykhaylo Andriluka[1] Hongyi Xu[1] Cristian Sminchisescu[1]

[1]Google Research, [2]Lund University

erik.gartner@math.lth.se

{mykhayloa,hongyixu,sminchisescu}@google.com

Abstract

We focus on the task of estimating a physically plausible articulated human motion from monocular video. Existing approaches that do not consider physics often produce temporally inconsistent output with motion artifacts, while state-of-the-art physics-based approaches have either been shown to work only in controlled laboratory conditions or consider simplified body-ground contact limited to feet. This paper explores how these shortcomings can be addressed by directly incorporating a fully-featured physics engine into the pose estimation process. Given an uncontrolled, real-world scene as input, our approach estimates the ground-plane location and the dimensions of the physical body model. It then recovers the physical motion by performing trajectory optimization. The advantage of our formulation is that it readily generalizes to a variety of scenes that might have diverse ground properties and supports any form of self-contact and contact between the articulated body and scene geometry. We show that our approach achieves competitive results with respect to existing physics-based methods on the Human3.6M benchmark [13], while being directly applicable without re-training to more complex dynamic motions from the AIST benchmark [36] and to uncontrolled internet videos.

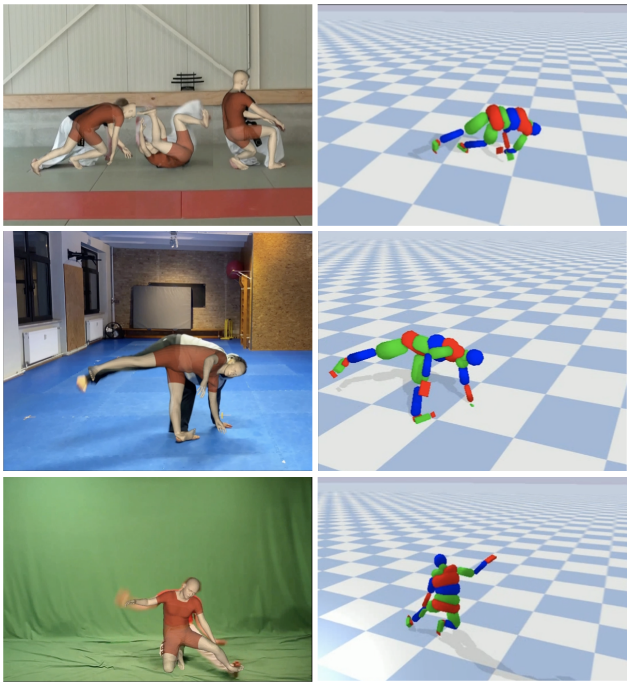

Figure 1. Example results of our approach on internet videos of dynamic motions. Note that our model can reconstruct physically plausible articulated 3d motion even in the presence of complex contact with the ground: full body contact (top row), feet and hands (middle), and feet and knee contacts (bottom).

1. Introduction

In this paper, we address the challenge of reconstructing physically plausible articulated 3d human motion from monocular video aiming to complement the recent methods [15, 16, 23, 42, 42, 48] that achieve increasingly more accurate 3d pose estimation results in terms of standard joint accuracy metrics, but still often produce reconstructions that are visually unnatural.

Our primary mechanism to achieve physical plausibility is to incorporate laws of physics into the pose estimation process. This naturally allows us to impose a variety of desirable properties on the estimated articulated motion, such as temporal consistency and balance in the presence of gravity. Perhaps one of the key challenges in using physics for pose estimation is the inherent complexity of adequately modeling the diverse physical phenomena that arise due to interactions of people with the scene. In the recent literature [29–31, 43] it is common to keep the physics model simple to enable efficient inference. For example, most of

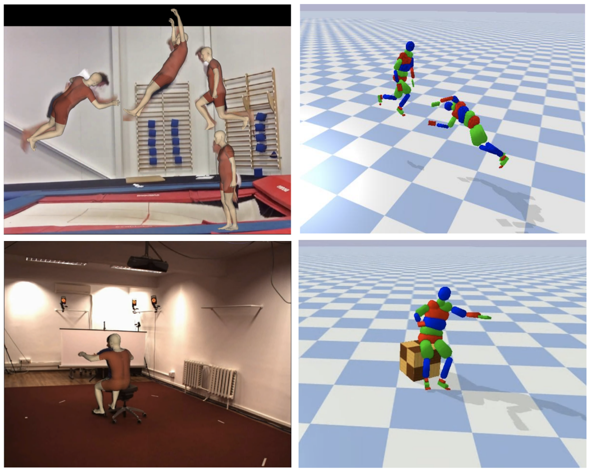

Figure 2. Examples results of our approach for scene with soft ground (top) and interaction with a chair (bottom).

the recent approaches opt for using simplified contact models (considering foot contact only), ignore potential effects due to interaction with objects other than the ground-plane, and do not model more subtle physical effects such as sliding and rolling friction, or surfaces with varying degrees of softness. Clearly there are many real-world scenarios where leveraging a more feature-complete physical model is necessary. We explore physics-based articulated pose estimation using feature-complete physical simulation as a building block to address this shortcoming. The advantage of such an approach is that it allows our method to be readily applicable to a variety of motions and scenarios that have not previously been tackled in the literature (see fig. 1 and 2). Specifically, in contrast to [29–31, 43] our approach can reconstruct motions with any type of contact between the body and the ground plane (see fig. 1). Our approach can also model interaction with obstacles and supporting surfaces such as furniture and allows for varying the stiffness and damping of the ground-plane to represent special cases such as trampoline floor (see fig. 2). We rely on the Bullet [7] engine, which was previously used for simulating human motion in [24]. However, none of our implementation details are engine-specific, so we envision that the quality of our results might continue to improve with further development in physical simulation.

The main contribution of this paper is to experimentally evaluate the use of trajectory optimization for physics-based articulated motion estimation on laboratory and real-world data using a generic physics engine as a building block. We demonstrate that combining a feature-complete physics engine and trajectory optimization can reach competitive or better accuracy than state-of-the-art methods while being applicable to a large variety of scenes and motion types. Furthermore, to the best of our knowledge, we are the first to apply physics-based reconstruction to complex real-world motions such as the ones shown in fig. 1 and 2. As a second contribution, we generate technical insights such as demonstrating that we can reach excellent alignment of estimated physical motion with 2d input images by automatically adapting the 3d model to the person in the image, and employing appropriate 2d alignment losses. This is in contrast to related work [29–31, 43] that typically does not report 2d alignment error and qualitatively may not achieve good 2d alignment of the physical model with the image. We also contribute to the understanding of the use of the residual root force control [45]. Such residual root force has been hypothesized as essential to bridge the simulation-to-reality gap and compensate for inaccuracies in the physical model. We experimentally demonstrate that the use of physically unrealistic residual force control might not be necessary, even in cases of complex and dynamic motions.

2. Related work

In the following, we first discuss recent literature on 3d human pose estimation that does not incorporate physical reasoning. We then review the related work on physics-based human modeling and compare our approach to other physics-based 3d pose estimation approaches.

3d pose estimation without physics. State-of-the-art methods are highly effective in estimating 2d and 3d people poses in images [5, 15, 49], and recent work has been able to extend this progress to 3d pose estimation in video [16, 23, 42]. The key elements driving the performance of these methods is the ability to estimate data-driven priors on articulated 3d poses [16, 47] and learn sophisticated CNN-based representations from large corpora of annotated training images [13, 14, 21, 37]. As such, these methods perform very well on common poses but are still challenged by rare poses. Occlusions, difficult imaging conditions, and dynamic motions (e.g. athletics) remain a challenge as these are highly diverse and hard to represent in the training set. As pointed out in [29], even for common poses state-of-the-art methods still often generate reconstructions prone to artifacts such as floating, footskating, and non-physical leaning. We aim to complement the statistical models used in the state-of-the-art approaches by incorporating laws of physics into the inference process and thus adding a component that is universally applicable to any human motion regardless of the statistics of the training or test set.

In parallel with recent progress in pose estimation, we now have accurate statistical shape and pose models [3, 20, 44]. These body models are typically estimated from thousands of scans of people and can generate shape deformations for a given pose. In this paper, we take advantage of these improvements and use a statistical body shape model [44] to define the dimensions of our physical model and derive the mass from the volume of the body parts.

Physics-based human motion modeling. Human motion modeling has been a subject of active research in com-

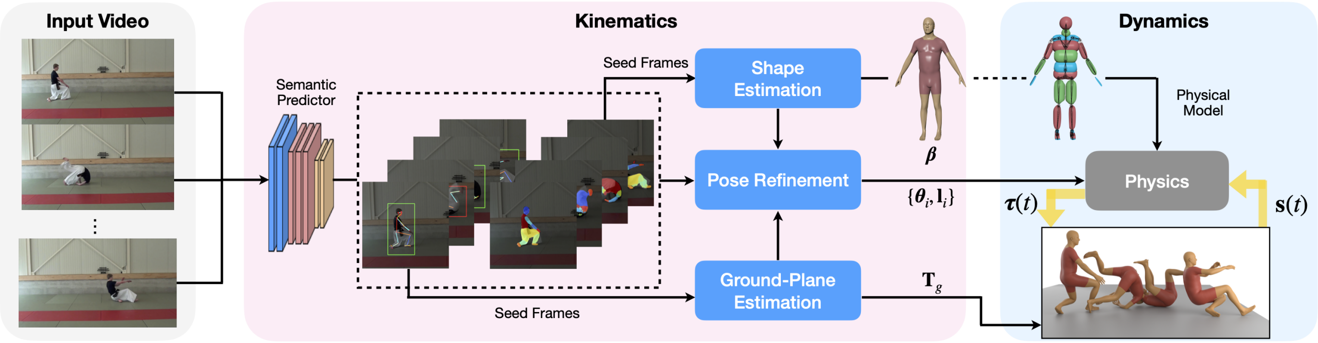

Figure 3. **Overview.** Given a monocular video of a human motion, we estimate the parameters of a physical human model and motor control trajectories $\boldsymbol{\tau}(t)$ such that the physically simulated human motion aligns with the video. We first use an inference network that predicts 2d landmarks $\mathbf{l}_i$ and body semantic segmentation masks from the video frames. From n seed frames we estimate a time-consistent human shape $\boldsymbol{\beta}$ and the ground-plane location $\mathbf{T}_g$. These are then kept fixed during a per-frame pose refinement step which provides the 3d kinematic initialization $\{\boldsymbol{\theta}_i\}$ to the physics optimization. The dynamics stage creates a physical model that mirrors the statistical shape model with appropriate shape and mass. Our dynamics optimization improves 3d motion estimation taking into account 3d kinematics, 2d landmarks and physical constraints. We refer to §3 for details.

	Contact model	Real-time	Physics implementation	Residual force	Body model	Real-world videos
Li *et al.* [19]	body joints	no	custom	no	fixed	yes
Rempe *et al.* [29]	feet	no	custom	no	fixed	yes
PhysCap [31]	feet	yes	custom	yes	fixed	yes
Shimada *et al.* [30]	feet	yes	custom	yes	fixed	yes
SimPoE [46]	full body	yes	MuJoCo [35]	yes	adapt.	no
Xie *et al.* [43]	feet	no	custom	no	adapt.	no
DiffPhy [9]	full body	no	TDS [12]	no	adapt.	yes
Ours	full body	no	Bullet [7]	no	adapt.	yes

Table 1. Comparison of recent physics-based articulated pose estimation approaches. "Contact model" indicates what contact points between body and ground are considered, "Residual force" indicates if the physical model allows application of additional external force to move the person (see [45]), "Body model" specifies if approach adapts the physical model to person in the video, and "Real-world videos" specifies if approach has also been evaluated on real-world videos or only on videos captured in laboratory conditions.

puter graphics [2, 17], robotics [8] and reinforcement learning [11,24,40] literature. With a few exceptions, most of the models in these domains have been constructed and evaluated using the motion capture data [2]. Some work such as [26] use images as input, aiming to train motion controllers for a simulated character capable of performing the observed motion under various perturbations. That work focuses on training motion controllers for a fixed character, whereas our focus is on estimating the motion of the subject observed in the image. Furthermore, the character's size, shape, and mass are independent of the observed subject. [17] propose a realistic human model that directly represents muscle activations and a method to learn control policies for it. [41] generate motions for a variety of character sizes and learn control policies that adapt to each size. [17, 41] and similar results in the graphics literature do not demonstrate this for characters observed in real images and do not deal with challenges of jointly estimating physical motion and coping with ambiguity in image measurements or the 2d to 3d lifting process [33].

Physics-based 3d pose estimation. Physics-based human pose estimation has a long tradition in computer vision [4, 22, 38]. Early works such as [38] already incorporated physical simulation as prior for 3d pose tracking but only considered simple motions such as walking and mostly evaluated in the multi-view setting in the controlled laboratory conditions. We list some of the properties of the recent works in tab. 1. [19] demonstrate joint physics-based estimation of human motion and interaction with various tool-like objects. [29] proposes a formulation that simplifies physics-based reasoning to feet and torso only, and infers positions of other body parts through inverse kinematics, whereas [19] jointly model all body parts and also include forces due to interaction with an object. [30, 31] use a specialized physics-based formulation that solves for ground-reaction forces given pre-detected foot contacts and kinematic estimates. In contrast, we do not assume that contacts can be detected a-priori, and in our approach, we estimate these as part of the physical inference. Hence we are not limited to predefined types of contact as [19, 29–31] or their accurate a-priori estimates. We show that we quantitatively improve over [29, 31], and qualitatively show how we can address more difficult in-the-wild internet videos of activities such as somersaults and sports, which would be difficult to reconstruct using previous methods. Our work is conceptually similar to SimPoE [46] in that both works use physics simulation. In contrast to SimPoE, we introduce a complete pipeline that is applicable to real-world videos, whereas SimPoE has been tested only in laboratory conditions and requires a calibrated camera. Furthermore, since SimPoE relies on reinforcement learning to train dataset-specific neural network models to control the simulated body, it is not clear how well SimPoE would generalize to variable motions present in real-world videos. One clear advantage of the SimPoE approach is its fast execution at test time, which comes at the cost of lengthy pre-training. Our approach is related to the approach of [43] which also estimates 3d human motion by minimizing an objective function that incorporates physics constraints. Perhaps the most significant differences to [43] are that (1) we use the full-featured physics model whereas they consider simplified physical model, (2) their model considers physics-based loss, but the output is not required to correspond to actual physical motion, and (3) they do not discuss performance of the approach on real-world data. The advantage of [43] is that they define a differentiable model that can be readily optimized with gradient descent. Finally, the concurrent work [9] tackles physics-based human pose reconstruction by minimizing a loss using a differentiable physics simulator given estimated kinematics.

3. Our approach

We present an overview of our approach in fig. 3. Given monocular video as input, we first reconstruct the initial kinematic 3d pose trajectory using a kinematic approach of [48] and use it to estimate body shape and the position of the ground plane relative to the camera. Subsequently, we instantiate a physical person model with body dimensions and weight that match the estimated body shape. Next, we formulate an objective function that measures the similarity between the motion of the physical model and image measurements and includes regularization terms that encourage plausible human poses and penalize jittery motions. Finally, we reconstruct the physical motion by minimizing this objective function with respect to the joint torque trajectories. To realize the physical motion, we rely on the implementation of rigid body dynamics available in Bullet [7].

3.1. Body model and control

We model the human body as rigid geometric primitives connected by joints. Our model consists of 26 capsules and has 16 3d body joints for a total of 48 degrees of freedom. We rely on a statistical model of human shape [44] to instantiate our model for a variety of human body types. To that end, given the 3d mesh representing the body shape, we estimate dimensions of the geometric primitives to approximate the mesh following the approach of [2]. We then compute the mass and inertia of each primitive based on its volume and estimate the mass based on an anatomical weight distribution [28] from the statistical human shape dataset CAESAR [27].

We do not model body muscle explicitly and instead actuate the model by directly applying the torque at the body joints. We denote the vector of torques applied at time t as $\boldsymbol{\tau}_t$, the angular position, and velocity of each joint at time t as $\mathbf{q}_t$ and $\dot{\mathbf{q}}_t$, and the set of 3d Cartesian coordinates of each joint at time t as $\mathbf{x}_t$. Similarly to [25], we control the motion of the physical model by introducing a sequence of control targets $\hat{\mathbf{q}}_{1:T} = \{\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \dots, \hat{\mathbf{q}}_t\}$ which are used to derive the torques via a control loop. The body motion in our model is then specified by the initial body state $\mathbf{s}_0 = (\mathbf{q}_0, \dot{\mathbf{q}}_0)$, the world geometry $\mathbf{G}$ specifying the position and orientation of the ground plane, the control trajectory for each joint $\hat{\mathbf{q}}_{1:T}$ and the corresponding control rule. We assume the initial acceleration to be $\mathbf{0}$. To implement the control loop we rely on the articulated islands algorithm[1] (AIA) [34] that incorporates motor control targets as constraints in the linear complementarity problem (LCP) (*cf*. (6.3) a, b in [34]) alongside contact constraints. AIA enables stable simulation already at 100 Hz compared to 1000-2000 Hz for PD control used in [2, 9, 25].

3.2. Physics-based articulated motion estimation

Our approach to the task of physical motion estimation is generally similar to other trajectory and spacetime optimization approaches in the literature [1, 2, 39]. We perform optimization over a sequence of overlapping temporal windows, initializing the start of each subsequent window with the preceding state in the previous window. To reduce the dimensionality of the search space, we use cubic B-spline interpolation to represent the control target $\hat{\mathbf{q}}_{1:T}$ and perform optimization over the spline coefficients [6]. Given the objective function L introduced in §3.3 we aim to find the optimal motion by minimizing L with respect to the spline coefficients of the control trajectory $\hat{\mathbf{q}}_{1:T}$. We initialize the control trajectory with the kinematic estimates of the body joints (see §3.4). The initial state is initialized from the corresponding kinematic estimate. We use the finite difference computed on the kinematic motion to estimate the initial velocity. As in [1, 2] we minimize the objective function with the evolutionary optimization approach CMA-ES [10] since our simulation environment does not support differentiation with respect to the dynamics variables. We generally observe convergence with CMA-ES after 2000 iterations per window with 100 samples per iteration. The inference takes $20 - 30$ minutes when evaluating 100 samples in parallel.

[1] "POSITION_CONTROL" mode in Bullet.

3.3. Objective functions

We use a composite objective function given by a weighted combination of several components.

3d pose. To encourage reconstructed physical motion to be close to the estimated kinematic 3d poses $\mathbf{q}^{\mathrm{k}}_{1:T}$ we use the following objective functions

$$
\begin{aligned}
L_{COM}(\hat{\mathbf{q}}_{1:T}) &= \sum_t (\|\mathbf{c}_t - \mathbf{c}^{\mathrm{k}}_t\|_2^2 + \|\dot{\mathbf{c}}_t - \dot{\mathbf{c}}^{\mathrm{k}}_t\|_2^2) \quad (1)\\
L_{pose} &= \sum_t \sum_{j \in \mathbf{J}} \arccos(|\langle \mathbf{q}_{tj}, \mathbf{q}^{\mathrm{k}}_{tj} \rangle|) \quad (2)
\end{aligned}
$$

where $\mathbf{c}_t$ and $\mathbf{c}^{\mathrm{k}}_t$ denote the position of the center of mass at time t in the reconstructed motion and kinematic estimate. L_{pose} measures the angle between observed joint angles and their kinematic estimates and the summation (2) is over the set J of all body joints including the base joint which defines the global orientation of the body.

2d re-projection. To encourage alignment of 3d motion with image observations, we use a set of $N = 28$ landmark points that include the main body joints, eyes, ears, nose, fingers, and endpoints of the feet. Let $\mathbf{l}_t$ denote the positions of 3d landmarks on the human body at time t, $\mathbf{C}$ be the camera projection matrix that maps world points into the image via perspective projection, $\mathbf{l}^d_t$ be the vector of landmark detections by the CNN-detector, and $\mathbf{s}_t$ the corresponding detection score vector. The 2d landmark re-projection loss is then defined as

$$L_{2d} = \sum_t \sum_n \mathbf{s}_{tn} \|\mathbf{C}\mathbf{l}_{tn} - \mathbf{l}^d_{tn}\|_2. \quad (3)$$

See §3.4 for details on estimating the 2d landmarks.

Regularization. We include several regularizers into our objective function. Firstly, we use the normalizing flow prior on human poses introduced in [47] which penalize unnatural poses. The loss is given by

$$L_{nf} = \sum_t \|\mathbf{z}(\mathbf{q}_t)\|_2, \quad (4)$$

where $\mathbf{z}(\mathbf{q}_t)$ is the latent code corresponding to the body pose $\mathbf{q}_t$. To discourage jittery motions we a add total variation loss on the acceleration of joints

$$L_{TV} = \frac{1}{J} \sum_t \sum_j \|\ddot{\mathbf{x}}_{tj} - \ddot{\mathbf{x}}_{t-1,j}\|_1 \quad (5)$$

Finally, we include a L_{lim} term that adds exponential penalty on deviations from anthropomorphic joint limits. The overall objective L used in physics-based motion estimation is given by the weighted sum of (1- 5) and of the term L_{lim}. See the supplemental material for details.

Model	MPJPE-G	MPJPE	MPJPE-PA
HUND [48]	239	116	72
+ S	233	110	71
+ SO	178	85	62
+ SO + G	148	84	63
+ SO + T	186	85	61
+ SO + GT	135	80	58

Table 2. Ablation of kinematics improvements on HUND on a validation subset of Human3.6M. *+S* indicates time-consistent body shape, *+O* indicates additional non-linear optimization, *+G* using ground-plane constraints, and *+T* temporal smoothness constraints.

3.4. Kinematic 3d pose and shape estimation

In this section, we describe our approach to extracting 2d and 3d evidence from the input video sequence.

Body shape. Given the input sequence, we proceed first to extract initial per-frame kinematic estimates of the 3d pose and shape using HUND [48]. As part of its optimization pipeline HUND also recovers the camera intrinsics $\mathbf{c}$ and estimates the positions of 2d landmarks, which we use in the 2d re-projection objective in (3). HUND is designed to work on single images, so our initial shape and pose estimates are not temporally consistent. Therefore, to improve the quality of kinematic 3d pose initialization, we extend HUND to pose estimation in video. We evaluate the additional steps introduced in this section in the experiments shown in tab. 2 using a validation set of 20 sequences from Human3.6M dataset. In our adaptation, we do not re-train the HUND neural network predictor and instead, directly minimize the HUND loss functions with BFGS. As a first step, we re-estimate the shape jointly over multiple video frames. To keep optimization tractable, we first jointly estimate shape and pose over a subset of $n = 5$ seed frames and then re-estimate the pose in all video frames keeping the updated shape fixed. The seed frames are selected by the highest average 2d keypoint confidence score. We refer to the HUND approach with re-estimated shape as *HUND+S* and to our approach where we subsequently also re-estimate the pose as *HUND+SO*. In tab. 2 we show results for both variants. Note that *HUND+SO* improves considerably compared to the original HUND results.

Ground plane. We define the location of the ground plane by the homogeneous transformation $\mathbf{T}_g$ that maps from the HUND coordinates to the canonical coordinate system in which the ground plane is passing through the origin, and its normal is given by the "y" axis. Let $\mathbf{M}^t$ be a subset of points on the body mesh at frame t. The signed distance from the mesh points to the ground plane is given by $D(\mathbf{M}^t) = \mathbf{T}_g \mathbf{M}^t \mathbf{e}_y$, where $\mathbf{e}_y = [0, 1, 0, 0]^T$ is the unit vector of the "y" axis in homogeneous coordinates. To estimate the transformation $\mathbf{T}_g$ we introduce an objective function

$$L_{gp}(\mathbf{T}_g, \mathbf{M}) = \sum_t \| \min(\delta, L_k(D(\mathbf{M}^t)))\|_2, \qquad (6)$$

where $L_k(D^t)$ corresponds to the smallest $k = 20$ signed distances in D^t. This objective favors $\mathbf{T}_g$ that places body mesh in contact with the ground without making preference for a specific contact points. This objective is also robust to cases when person is in the air by clipping the distance at δ, which we set to 0.2m in the experiments in this paper. We recover $\mathbf{T}_g$ by minimizing

$$\begin{aligned} L_{gp}(\mathbf{T}_g) =& L_{gp}(\mathbf{T}_g, \mathbf{M}_l) + L_{gp}(\mathbf{T}_g, \mathbf{M}_r) \\ &+ 2L_{gp}(\mathbf{T}_g, \mathbf{M}_b), \end{aligned} \qquad (7)$$

where $\mathbf{M}_l$, $\mathbf{M}_r$ and $\mathbf{M}_b$ are the meshes of the left foot, right foot and whole body respectively. This biases the ground plane to have contact with the feet, but is still robust to cases when person is jumping or touching the ground with other body parts (e.g. as in the case of a somersault).

3d pose. In the final step, we re-estimate the poses in all frames using the estimated shape and ground plane while adding the temporal consistency objective

$$L_{temp} = \sum_t \|\mathbf{M}^t - \mathbf{M}^{t-1}\|_2 + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_2, \qquad (8)$$

where $\mathbf{M}^t$ is a body mesh and $\boldsymbol{\theta}_t$ is a HUND body pose vector in frame t. To enforce ground plane constraints we use (6), but now keep $\mathbf{T}_g$ fixed and optimize with respect to body pose. In the experiments in tab. 2 we refer to the variant of our approach that uses temporal constraints in (8) as *HUND+SO+T* and to the full kinematic optimization that uses both temporal and ground plane constraints as *HUND+SO+GT*. Tab. 2 demonstrates that both temporal and ground-truth constraints considerably improve the accuracy of kinematic 3d pose estimation. Even so, the results of our best variant *HUND+SO+GT* still contain artifacts such as motion jitter and footskating, which are substantially reduced by the dynamical model (see tab. 3).

4. Experimental results

Datasets. We evaluate our method on three human motion datasets: Human3.6M [13], HumanEva-I [32] and AIST [36]. In addition, we qualitatively evaluate on our own "in-the-wild" internet videos. To compare different variants of our approach in tab. 2 and tab. 3 we use a validation set composed of 20 short 100-frame sequences from the Human3.6M dataset. We use the same subset of full-length sequences as proposed in [43] for the main evaluation in tab. 4. We use a preprocessed version of the AIST dataset [36] from [18] which contains pseudo 3d body pose ground-truth obtained through multi-view reconstruction. For our experiments, we select a subset of fifteen videos featuring diverse dances of single subjects. For the evaluation on HumanEva-I, we follow the protocol defined in [29] and evaluate on the walking motions from the validation split of the dataset using images from the first camera. We assume known camera extrinsic parameters in the Human3.6M experiments and estimate them for other datasets. In order to speed up the computation of the long sequences of Human3.6M in tab. 4 we compute all temporal windows in parallel and join them together in post-processing.

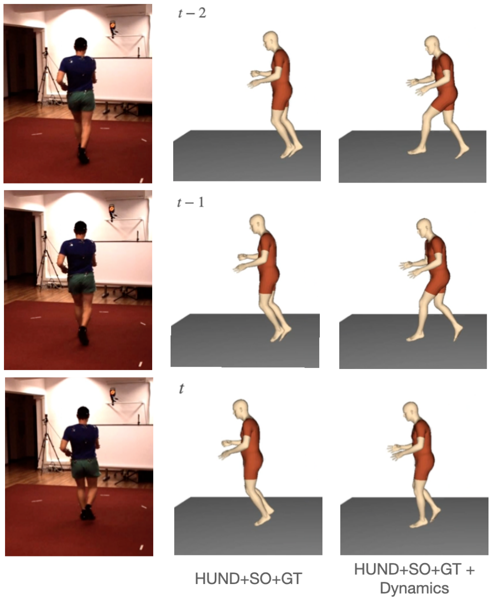

Figure 4. Qualitative results on the Human3.6M dataset. Note how the dynamical model (right) recovers plausible locomotion.

We report results using mean global per-joint position error (mm) overall joints (MPJPE-G), as well as translation aligned (MPJPE) and Procrustes aligned (MPJPE-PA) error metrics. Note that to score on the MPJPE-G metric an approach should be able to both estimate the articulated pose and correctly track the global position of the person in world coordinates. In addition to standard evaluation metrics, we implement the foot skate and floating metrics similar to those introduced in [29] but detect contacts using a threshold rather than through contact annotation. Finally, we report image alignment (MPJPE-2d) and 3d joint velocity error in m/s. See supplementary for further details.

Analysis of model components. In tab. 3 we present ablation results of our approach. Our full dynamical model uses kinematic inputs obtained with *HUND+SO+GT* introduced in §3.4 and is denoted as *HUND+SO+GT + Dynamics*. Our dynamical model performs comparably or slightly better compared to *HUND+SO+GT* on joint localization metrics (e.g. MPJPE-G improves slightly from 135 to 132 mm) but greatly reduces motion artifacts. The percentage of frames with footskate is reduced from 64 to

Model	MPJPE-G	MPJPE	MPJPE-PA	MPJPE-2d	Velocity	Footskate (%)	Float (%)
HUND+SO	178	85	62	12	1.3	25	40
HUND+SO + Dynamics	167	87	62	12	0.45	**7**	1
HUND+SO+GT	135	80	58	12	0.58	64	0
HUND+SO+GT + Dynamics	**132**	80	**57**	**11**	**0.27**	8	0
HUND+SO+GT + Dynamics							
w/o 2d re-projection, (3)	154	104	68	17	0.32	-	-
w/o 3d joints, (2)	134	84	60	11	0.27	-	-
w/o COM, (1)	149	81	57	11	0.31	-	-
w/o COM and 3d joints, (1, 2)	151	85	59	11	0.33	-	-
w/o pose prior, (4)	138	80	57	11	0.24	-	-

Table 3. Ablation experiments of the dynamics model on a validation set of 20 sequences from the Human3.6M dataset.

Dataset	Model	MPJPE-G	MPJPE	MPJPE-PA	MPJPE-2d	Velocity	Footskate (%)
Human3.6M	VIBE [16]	208	69	44	16	0.32	27
	PhysCap [31]	-	97	65	-	-	-
	SimPoE [46]	-	**57**	**42**	-	-	
	Shimada *et al.* [30]	-	77	58	-	-	-
	Xie *et al.* [43] (Kinematics)	-	74	-	-	-	-
	Xie *et al.* [43] (Dynamics)	-	68	-	-	-	-
	Ours: HUND+SO+GT	145	83	56	14	0.46	48
	Ours: HUND+SO+GT + Dynamics	**143**	84	56	**13**	**0.24**	**4**
HumanEva-I	Rempe et al. [29] (Kinematics)	408	-	-	-	-	-
	Rempe et al. [29] (Dynamics)	422	-	-	-	-	-
	Ours: HUND+SO+GT	208	**90**	76	14	0.51	40
	Ours: HUND+SO+GT + Dynamics	**196**	91	**74**	14	**0.27**	**4**
AIST	Ours: HUND+SO+GT	156	**107**	**67**	**10**	0.59	51
	Ours: HUND+SO+GT + Dynamics	**154**	113	69	13	**0.41**	**4**

Table 4. Quantitative results of our models compared to prior work on Human3.6M [13], HumanEva-I [32] and a subset of AIST [18, 36].

8 and error in velocity from 0.58 to 0.27 m/s. We also evaluate a dynamic model based on a simpler kinematic variant *HUND+SO* that does not incorporate ground-plane and temporal constraints when re-estimating poses from video. For *HUND+SO*, the inference with dynamics similarly improves perceptual metrics considerably. Note that *HUND+SO* produces output that suffers from both footskating (25% of frames) and floating (40% of frames). Adding ground-plane constraints in (*cf*. (6)) removes floating artifacts in *HUND+SO+GT*, but the output still suffers from footskating (64% of the frames). Dynamical inference helps to substantially reduce both types of artifacts both for *HUND+SO* and *HUND+SO+GT*. In fig. 4 we show example output of *HUND+SO+GT + Dynamics* and compare it to *HUND+SO+GT* which it uses for initialization. Note that for *HUND+SO+GT* the person in the output appears to move forward by floating in the air, whereas our dynamics approach infers plausible 3d poses consistent with the subject's global motion. In the bottom part of tab. 3 we report results for our full model *HUND+SO+GT + Dynamics* while ablating components of the objective function (*cf*. §3.3). We observe that all components of the objective function contribute to the overall accuracy. The most important components are the 2d re-projection (*cf*. (3)) and difference in COM position (*cf*. (1)). Without these, the MPJPE-G increases from 132 to 154 and 151 mm, respectively. Excluding the 3d joints component leads to only a small loss of accuracy from 132 to 134 mm.

Comparison to state-of-the-art. In tab. 4 we present the results of our full model on the Human3.6M, HumanEva-I, and AIST datasets. We compare to VIBE [16] using the publicly available implementation by the authors and use the evaluation results of other approaches as reported in the original publications. Since VIBE generates only root-relative pose estimates, we use a similar technique as proposed in PhysCap [31] and estimate the global position and orientation by minimizing the 2d joint reprojection error. On the Human3.6M benchmark, our approach improves over VIBE and our own *HUND+SO+GT* in terms of joint accuracy and perceptual metrics. Compared to VIBE, the MPJPE-G improves from 208 to 143 mm, MPJPE-2d improves from 16 to 13 px, and the percentage of footskating frames are reduced from 27% to 4%. Interestingly our approach achieves the best MPJPE-PA overall physics-based approaches except the pretrained SimPoE, but reaches somewhat higher MPJPE compared to [30] and fairly recent work of [43] (82 mm vs 68 mm for [43] and 77 mm for [30]). Note that [43] start with a stronger kinematic baseline (74 mm MPJPE) and that the performance of other approaches might improve as well given such better kinematic initialization. Furthermore, our dynamics approach improves over the results of [29] on HumanEva-I and achieves significantly better MPJPE-G compared to *HUND+SO+GT*. On the AIST dataset, dynamics similarly

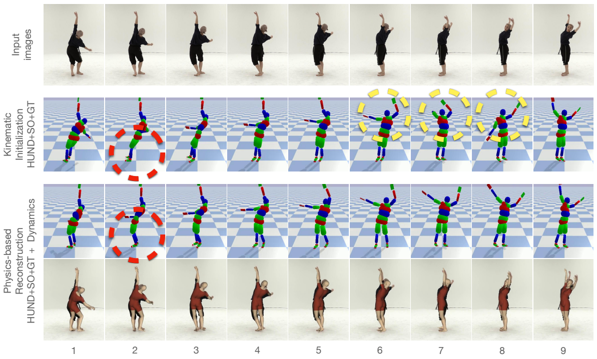
Input images
Kinematic Initialization HUND+SO+GT
Physics-based Reconstruction HUND+SO+GT + Dynamics
1
2
3
4
5
6
7
8
9

Figure 5. Example result on AIST [36]. The kinematic initialization produces poses that are unstable in the presence of gravity (red circle) or poses that are temporally inconsistent (yellow circles). Our physics-based approach corrects both errors.

improves in terms of MPJPE-G, footskating, and velocity compared to our kinematic initialization.

Results on real-world internet video. We show example results of our approach on the AIST dataset [36] in fig. 5 and on the real-world internet videos in fig. 1, 2 and 6. To obtain the results with a soft floor shown in fig. 2 we manually modify the stiffness and damping floor parameters to mimic the trampoline behavior. The sequence with the chair from the Human3.6M dataset shown in fig. 2 (bottom) is generated by manually adding a chair to the scene since our approach does not perform reasoning about scene objects.

In fig. 5 we qualitatively compare the output of our full system with physics to our best kinematic approach *HUND+SO+GT*. We strongly encourage the reader to watch the video in supplemental material[2] to appreciate the differences between the two approaches and to see the qualitative comparison to VIBE [16]. We observe that our physics approach is often able to correct out-of-balance poses produced by *HUND+SO+GT* (*e.g.* second frame in fig. 5) and substantially improves temporal coherence of the reconstruction. Note that typically both *HUND+SO+GT* and our physics-based approach produce outputs that match 2d observations, but the physics-based approach estimates 3d pose more accurately. For example, in the first sequence in fig. 6 the physics-based model infers the pose that enables the person to jump in subsequent frames, whereas *HUND+SO+GT* places the left leg at an angle that would make the jump impossible. Note that the output of the physics-based approach can deviate significantly from the kinematic initialization (second example in fig. 6).

5. Conclusion

In this paper, we have proposed a physics-based approach to 3d articulated video reconstruction of humans. By closely combining kinematic and dynamic constraints within an optimization process that is contact, mass, and inertia aware, with values informed by body shape estimates, we are able to improve the physical plausibility and reduce reconstruction artifacts compared to purely kinematic approaches. One of the primary goals of our work has been to demonstrate the advantages of incorporating an expressive physics model into the 3d pose estimation pipeline. Clearly, such a model makes inference more involved compared to specialized physics-based approaches such as [31, 43], but with the added benefit of being more capable and general.

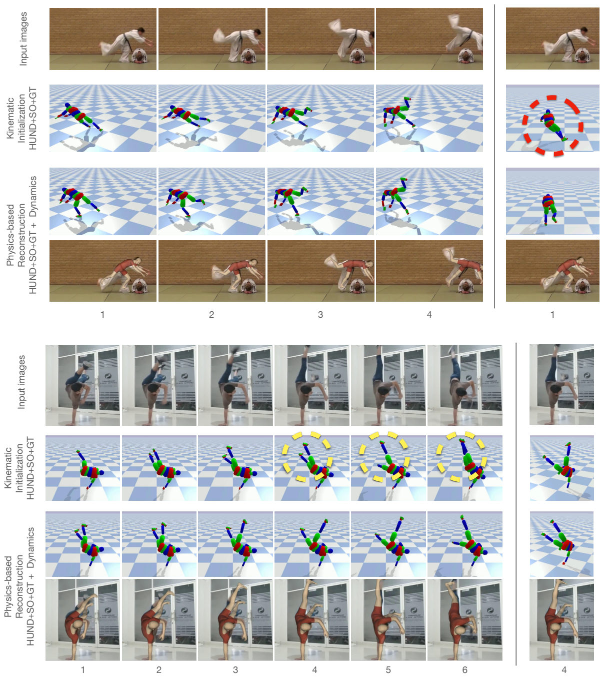
Input images
Kinematic Initialization HUND+SO+GT
Physics-based Reconstruction HUND+SO+GT + Dynamics
1
2
3
4
1
Input images
Kinematic Initialization HUND+SO+GT
Physics-based Reconstruction HUND+SO+GT + Dynamics
1
2
3
4
5
6
4

Figure 6. Example results on real-world videos. In the top row sequence, the kinematic initialization incorrectly places the left foot before the jump. We highlight the mistake by showing the scene from another viewpoint (red circle). The kinematic initialization also fails to produce temporally consistent poses in the example in the bottom row (yellow circles). Our physics-based inference corrects both errors and generates a more plausible motion. See tiny.cc/traj-opt for more results.

Ethical considerations. This work aims to improve the quality of human pose reconstruction through the inclusion of physical constraints. We believe that the level of detail in our physical model limits its applications in tasks such as person identification or surveillance. The same limitation also prevents its use in the generation of e.g. deepfakes, particularly as the model lacks a photorealistic appearance. We believe our model is inclusive towards and supports a variety of different body shapes and sizes. While we do not study this in the paper, we consider it important future work.

Acknowledgements. We would like to thank Erwin Coumans for his help with the project, as well as the supportive anonymous reviewers for their insightful comments.

[2] See tiny.cc/traj-opt.

References

[1] Mazen Al Borno, Martin de Lasa, and Aaron Hertzmann. Trajectory optimization for full-body movements with complex contacts. In *IEEE transactions on visualization and computer graphics*, volume 19, pages 1405–14, 08 2013. 4

[2] Mazen Al Borno, Ludovic Righetti, Michael J. Black, Scott L. Delp, Eugene Fiume, and Javier Romero. Robust Physics-based Motion Retargeting with Realistic Body Shapes. In *Computer Graphics Forum*, 2018. 3, 4

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[4] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating contact dynamics. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2389–2396, 2009. 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *CVPR*, 2017. 2

[6] Michael F. Cohen. Interactive spacetime control for animation. In *SIGGRAPH*, 1992. 4

[7] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2019. 2, 3, 4

[8] M. Da Silva, Y. Abe, and J. Popović. Simulation of human motion data using short-horizon model-predictive control. *Computer Graphics Forum*, 27(2):371–380, 2008. 3

[9] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3, 4

[10] Nikolaus Hansen. *The CMA Evolution Strategy: A Comparing Review*, pages 75–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. 4

[11] Nicolas Heess, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, S. M. Ali Eslami, Martin A. Riedmiller, and David Silver. Emergence of locomotion behaviours in rich environments. *CoRR*, abs/1707.02286, 2017. 3

[12] Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S Sukhatme. NeuralSim: Augmenting differentiable simulators with neural networks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 3

[13] 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, jul 2014. 1, 2, 6, 7

[14] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2

[16] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1, 2, 7, 8

[17] Seunghwan Lee, Moonseok Park, Kyoungmin Lee, and Jehee Lee. Scalable muscle-actuated human simulation and control. *ACM Transactions on Graphics*, 38:1–13, 07 2019. 3

[18] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 6, 7

[19] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 2

[22] Dimitris Metaxas. *Physics-Based Vision*. Kluwer Academic Publishing, 1997. 3

[23] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[24] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. In *SIGGRAPH*, 2018. 2, 3

[25] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 37(4):143:1–143:14, July 2018. 4

[26] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37(6), Nov. 2018. 3

[27] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 2017. 4

[28] 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

[29] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 6, 7

[30] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human

motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), aug 2021. 1, 2, 3, 7

[31] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), dec 2020. 1, 2, 3, 7, 8

[32] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, Mar. 2010. 6, 7

[33] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, 2003. 3

[34] Jakub Stepien. *Physics-Based Animation of Articulated Rigid Body Systems for Virtual Environments*. PhD thesis, 10 2013. 4

[35] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. 3

[36] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019. 1, 6, 7, 8

[37] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 2

[38] M. Vondrak, L. Sigal, and O. C. Jenkins. Physical simulation for probabilistic motion tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 3

[39] Andrew Witkin and Michael Kass. Spacetime constraints. In *SIGGRAPH*, 1988. 4

[40] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Trans. Graph.*, 39(4), 2020. 3

[41] Jungdam Won and Jehee Lee. Learning body shape variation in physics-based characters. *ACM Trans. Graph.*, 2019. 3

[42] 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*, 2019. 1, 2

[43] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 3, 4, 6, 7, 8

[44] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6184–6193, 2020. 2, 4

[45] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *Advances in Neural Information Processing Systems*, 2020. 2, 3

[46] 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 (CVPR)*, 2021. 3, 4, 7

[47] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *arXiv preprint arXiv:2003.10350*, 2020. 2, 5

[48] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 4, 5

[49] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NeurIPS*, pages 8410–8419, 2018. 2