Neural MoCon: Neural Motion Control for Physically Plausible Human Motion Capture

Buzhen Huang Liang Pan Yuan Yang Jingyi Ju Yangang Wang*

Southeast University, China

Abstract

Due to the visual ambiguity, purely kinematic formulations on monocular human motion capture are often physically incorrect, biomechanically implausible, and can not reconstruct accurate interactions. In this work, we focus on exploiting the high-precision and non-differentiable physics simulator to incorporate dynamical constraints in motion capture. Our key-idea is to use real physical supervisions to train a target pose distribution prior for sampling-based motion control to capture physically plausible human motion. To obtain accurate reference motion with terrain interactions for the sampling, we first introduce an interaction constraint based on SDF (Signed Distance Field) to enforce appropriate ground contact modeling. We then design a novel two-branch decoder to avoid stochastic error from pseudo ground-truth and train a distribution prior with the non-differentiable physics simulator. Finally, we regress the sampling distribution from the current state of the physical character with the trained prior and sample satisfied target poses to track the estimated reference motion. Qualitative and quantitative results show that we can obtain physically plausible human motion with complex terrain interactions, human shape variations, and diverse behaviors. More information can be found at https://www.yangangwang.com/papers/HBZ-NM-2022-03.html

1. Introduction

Recent years have witnessed significant development of marker-less motion capture, which promotes a wide variety of applications ranging from character animation to human-computer interaction, personal well-being, and human behavior understanding. Extensive existing works can kinematically capture accurate human pose from monocular videos and images via network regression [22,25,26,66,67] or optimization [6,38,41,51]. However, they are often hard to leverage in real-world systems due to a series of artifacts that are not satisfied biomechanical and physical plausibility (*e.g.*, jitter and floor penetration).

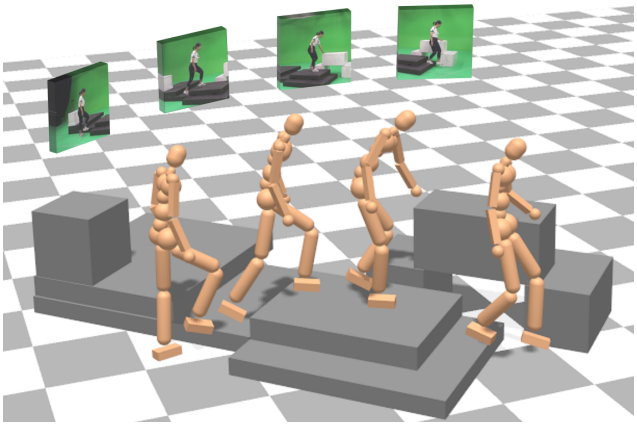

Figure 1. Our method captures physically plausible human motion from monocular RGB videos via neural motion control.

To improve motion quality and physical plausibility, a few works focus on capturing human motion using physics-based constraints. [42, 46, 47, 53, 57] incorporate physical laws as soft constraints in numerical optimization framework and reduce artifacts. To make optimization be tractable, they can only adopt simple and differentiable physical models, which may result in high approximation errors. Other methods [40,60,63] utilize non-differentiable physics simulators with deep reinforcement learning (DRL) to achieve accurate and physically plausible 3D human pose estimation. However, training a desirable policy requires complex configurations [1,5,31], and it may be sensitive to environmental changes [39, 60]. The limitations above make them be infeasible to estimate human pose with scene interactions and subject varieties for motion capture tasks. Nevertheless, motion control, typically sampling-based methods [35], have achieved an impressive performance in reproducing highly dynamic and acrobatic motions and is robust to contact-rich scenarios, which shows a way for general physics-based motion capture.

In this paper, we aim to construct a physics-based motion

*Corresponding author. E-mail: yangangwang@seu.edu.cn. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1403900, the National Natural Science Foundation of China (No. 62076061), the "Young Elite Scientists Sponsorship Program by CAST" (No. YES20200025), and the "Zhishan Young Scholar" Program of Southeast University (No. 2242021R41083).

capture framework that is more general to complex terrains, shape variations, and diverse behaviors along sampling-based motion control. However, employing sampling-based motion control in monocular motion capture tasks faces several challenges. First, conventional sampling-based methods [33, 35] often track the accurate reference motion from commercial motion capture systems, while the estimated motion from monocular RGB videos is noisy and physically implausible. An inaccurate contact results in an unnatural pose would even lead to an imbalance state for the character. Second, it is complicated to find an optimal distribution for the sampling. Although CMA (Covariance Matrix Adaptation) [11] is proved to be able to adjust distribution with black-box optimization [33], it requires evaluating plenty of samples for the distribution adaption, which is time-consuming. Furthermore, the adaption relied on random samples from an initial distribution imposes uncertainty for the motion capture.

To address the obstacles, **our key-idea is to train a motion distribution prior with physical supervisions. The prior provides feasible solutions for sampling-based motion control to capture physically plausible human motion from a monocular color video, which is named as Neural Motion Control (Neural MoCon).** We first introduce a human-scene interaction constraint to obtain a reference motion with appropriate contacts for sampling. Different from existing works [42, 47] to detect foot-ground contact status, our proposed interaction constraint adjusts the distance between two disconnected meshes via SDF, enforcing the human model to be close to the ground surface. Then, we have tried to train an encoder to regress the distribution with KL divergence (Kullback-Leibler divergence) and pseudo ground-truth from CMA. However, for the same character state and reference pose, the CMA method obtains different distributions, thus the stochastic error of CMA results in network divergence and erroneous regression. Consequently, we propose a novel two-branch decoder to address this obstacle. As shown in Fig. 3, the target pose sampled from the estimated distribution is fed into a physical branch to verify the validity. Since the simulator is non-differentiable, we use the output to supervise the pose decoder and enforce it to transfer the target pose to a dynamical pose like the simulator. Moreover, a reconstruction loss from the reference pose is applied to the decoded pose to promote correct distribution encoding. When the encoder is convergent, we use it to encode distribution and sample target poses for the physical branch to capture physically plausible motion. The main contributions of this work are summarized as follows.

- We propose an explicit physics-based motion capture framework that is more general to complex terrain, body shape variations, and diverse behaviors.
- We propose a novel two-branch decoder to avoid stochastic error from pseudo ground-truth and train the distribution prior with a non-differentiable physics simulator.
- We propose an interaction constraint based on SDF to capture accurate human-scene contact from complex terrain scenarios.

2. Related Work

Physics-based motion capture. VideoMocap [53] first employs physical constraints in motion capture by jointly optimizing the human pose and contact force, and this approach requires manual intervention to achieve satisfying results. Based on [53], [32] and [42, 47, 64] further consider the object interaction and kinematic pose estimation, respectively. Recently, Shimada *et al*. [46] proposed a neural network-based approach to estimate the ground reaction force and joint force and updated the character's pose using the derived accelerations. To make optimization tractable, their methods can only adopt simple and differentiable physics models with limited constraints, which results in high approximation errors. To address this problem, some latest works [40, 60, 62, 63] employ DRL to implement motion capture based on non-differentiable simulators. Nevertheless, training a desirable policy requires complex configurations [1, 5, 31], and it may be sensitive to motion types and body shape variations [39, 60]. Vondrak *et al*. [50] directly used the silhouette to construct a character-image consistency to train a state-machine controller. However, this approach could only be generalized to a variety of motions, and the recovered motion seems to be unnatural. In this work, we adopt neural motion control to capture motion rather than DRL. With the trained distribution prior, our method is more general to different terrain interactions, human shape variations, and diverse behaviors.

Physics-based character control. Physics-based character control is a longstanding problem [28, 29, 45, 49, 54, 55]. Early works rely on the inverted pendulum model [21], passive dynamics walking [27] and zero-moment-point-based trajectory generation [12] can handle simple motions. To solve large-DOF (degree-of-freedom) models, optimization-based methods [23, 30, 48, 56] are widely used to simulate and analyze human motions. However, it requires substantial computational effort to deal with a complex motion. Other methods [3, 59] approximate the actual human control systems and can produce both normal and pathological walking motions. These control-based methods can generalize to a variety of skills [3, 33–35, 59], but a set of hyperparameters are required to tune for the desired behaviors. Recent works adopt DRL to control physical character [28, 39, 58]. It shows that DRL can achieve high-quality motion when motion capture data are provided as a reference [39]. Curriculum learning promotes the DRL to

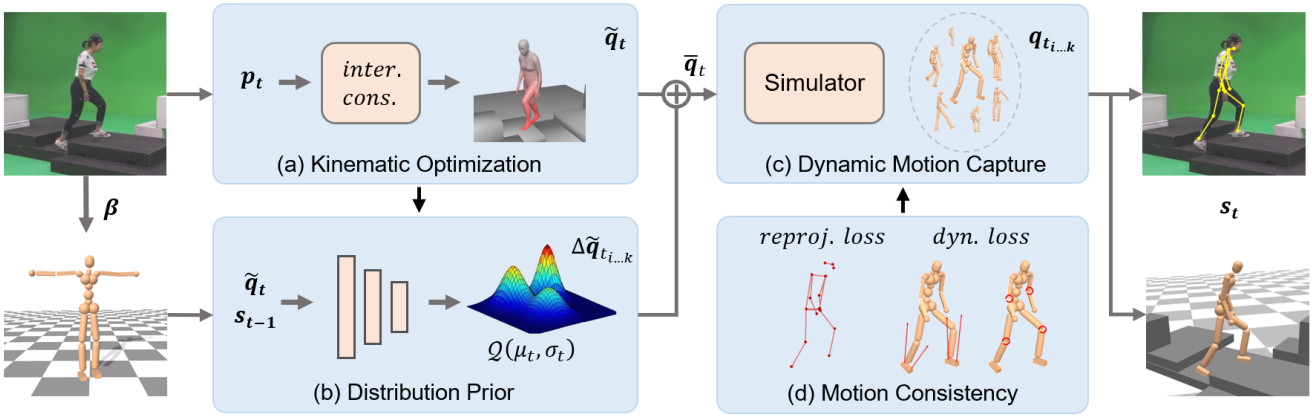

Figure 2. **Overview.** Our method first estimates reference motion with accurate human-scene interaction as well as human shape from a monocular RGB video (a). Then, a prior regresses a distribution from the state of physical character and the reference pose to sample target poses (b). The physics simulator is used to obtain a physically plausible pose for each sample (c). The sample with the lowest loss is adopted and used for the next frame after sample evaluation (d).

learn more complex tasks [58]. However, training an optimal policy takes numerous low- and high-level design decisions, which strongly affect the performance of the resulting agents. We follow sampling-based motion control [33, 35] to construct a general framework. Furthermore, we propose a network-based distribution prior to avoid the time-consuming distribution adaption and to improve the stability for their methods.

3D human with scene interaction. Modeling 3D human with scene interactions will promote the computational understanding of human behavior, which is important for metaverse and related applications. Previous works in scene labeling [19], scene synthesis [8], affordance learning [10, 24] and object arrangement [20] verified human context is helpful for scene understanding. The prior knowledge of scene geometry can also promote a more reasonable and accurate human pose estimation. [13, 15, 43, 44] generate human motion with interaction from the relationship between scene geometry and human body parts. [37] further utilizes this relationship to recover interactions from videos. To explicitly use scene information to improve pose accuracy, [14] formulates two constraints in optimization to reduce interpenetration and encourage appropriate contact. [65] also adopts the optimization-based approach and proposes a smoothness prior to improve motion quality. However, numerical optimization with soft constraints is hard to avoid artifacts like interpenetration, which is the main concern for human-scene reconstruction. In contrast, our method relies on a physics simulator [4] to provide hard physical constraints. With the network-based distribution prior, our method can obtain accurate terrain interactions via neural motion control.

3. Method

We propose a framework with a non-differentiable physics simulator [4] to capture physically plausible human motion. We first describe the representations of our kinematic and dynamical characters (Sec. 3.1). Then, an interaction constraint is designed to obtain reference motion with appropriate contact information (Sec. 3.2). In addition, we introduce a distribution prior trained with a novel two-branch structure for neural motion control (Sec. 3.3). Finally, we regress a distribution and sample satisfied target poses to track the estimated reference motion (Sec. 3.4).

3.1. Preliminaries

Representation. The kinematic motion is represented with SMPL model [36]. To represent different human shapes in the physics simulator, we design our physical character to have the same kinematic tree as SMPL. The bone length and link shape of the character can be directly obtained from the estimated SMPL parameters. We fix a few skeleton joints to have 57 DOFs. The state of character is denoted $\boldsymbol{s} = (\boldsymbol{q}, \dot{\boldsymbol{q}})$, where $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$ are the pose and velocity, respectively. The details of the model can be found in the supplementary material.

Sampling-based motion control. We briefly review the sampling-based motion control approach [35] to promote understanding of our method. A kinematic pose $\tilde{\boldsymbol{q}}_t$ is used as a reference, and we wish the physical character to dynamically track the reference pose via PD-control (Proportional Derivative). However, due to the inaccuracies of kinematic pose estimation and PD controller, the tracking always fails when directly applying the reference pose as the desired setpoint. The sampling algorithm samples a correction $\Delta\tilde{\boldsymbol{q}}_t$ for reference pose, thus employing the target pose $\bar{\boldsymbol{q}}_t = \tilde{\boldsymbol{q}}_t + \Delta\tilde{\boldsymbol{q}}_t$ can compensate the discrepancies. The quality of samples is evaluated by a loss function. By selecting the sample with the lowest loss, we can obtain the physically plausible motion. More details can be found in [35].

3.2. Reference motion estimation

The neural motion control requires reference motion with accurate ground contact to drive the physical character. To obtain the contact information, previous works [47, 60] train a network to estimate a binary foot contact status. However, no sufficient data can be utilized for training in complex terrain scenarios (*e.g.*, stairs and uneven ground). We address the problem by incorporating an SDF-based interaction constraint in an optimization-based framework.

Specifically, we optimize the latent code of pre-trained motion prior in [16] to fit SMPL models to single-view 2D poses detected by AlphaPose [7]. The overall formulation is:

$$\underset{(\mathbf{z},\mathcal{R},\mathcal{T})_{1:T},\beta}{\arg\min} \mathcal{L} = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{scene}}, \tag{1}$$

where $\mathbf{z}, \mathcal{R}, \mathcal{T}$ are the latent code, global rotation and translation for character in each frame. β is the human shape parameter, and T is the frame length. The data term is:

$$\mathcal{L}_{\text{data}} = \sum_{t=1}^{T} \sigma_t \left\| \Pi\left(\tilde{\boldsymbol{j}}_t\right) - \mathbf{p}_t \right\|^2, \tag{2}$$

where $\mathbf{p}$, σ are 2D poses and their corresponding confidence. $\tilde{\boldsymbol{j}}$ is the model joint position. We further add the regularization term:

$$\mathcal{L}_{\text{prior}} = \|\beta\|^2 + \sum_{t=1}^{T} \left(\|\mathbf{z}_t\|^2 + \|z_{t+1} - 2z_t + z_{t-1}\|^2 \right). \tag{3}$$

Due to the depth ambiguity, the recovered 3D human may float in the air or penetrate with the ground mesh with only the above constraints. With such reference motion, the simulated results are unnatural and incorrect. To reconstruct more accurate human-scene interactions from single-view videos, we generate a differentiable SDF of the scene mesh using [18]. In the optimization, we follow [14] to sample the SDF value for the pre-defined foot keypoints and use it to construct an objective function:

$$\mathcal{L}_{\text{scene}} = \rho \| \text{SDF}(\check{j}) \|^2, \tag{4}$$

where $\check{j}$ is the 3D positions of the keypoints and SDF is the sample operation. Our optimization has four stages. Since the proximate motion can be obtained in the first three stages, we only apply the interaction term to refine the ground contact in the last stage. To make our method to be compliant with airborne motions, we further apply a Geman-McClure error function ρ [9] to down-weight keypoints that are far from the scene mesh.

3.3. Distribution prior training

It is essential to find an optimal target pose distribution to achieve physically plausible motion for sampling-based motion control. Previous works [33] use (μ_W, λ)-CMA-ES method [11] to realize the distribution adaption. However, the time-consuming operation and stochastic error of the adaption make it hard to be leveraged in motion capture for real-world applications. We propose to replace this operation and improve the performance with a network-based distribution prior. To train the network, a naive idea is to directly supervise the distribution using the CMA results. Given a pair of character state and reference pose, it seems that we can provide the supervisions by running CMA online before feeding the data into the network or preparing the pseudo supervision with CMA in advance. Actually, the two strategies are both infeasible in real implementation. For the same character state and reference pose, the CMA method obtains different distributions, resulting in network divergence and erroneous regression for online and offline strategies, respectively.

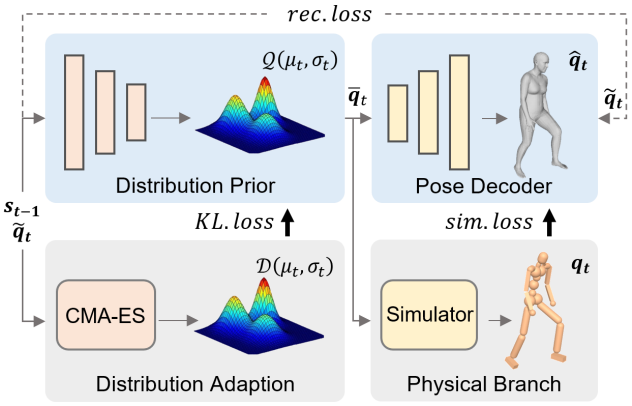

Figure 3. Different from conventional approaches (gray). We propose a two-branch decoder to avoid stochastic error from the CMA-ES method and train the distribution prior with real physical supervision. The non-differentiable physical branch simulates the sampled result, and the pose decoder intermediately employs the physical information to optimize the prior with simulation loss and reconstruction loss.

To solve this obstacle, we propose a two-branch decoder to assist training an accurate and generalized distribution encoder. As shown in the Fig. 3, we first pre-train the distribution encoder with the supervision from offline CMA. Since the network parameters trained with the inaccurate supervisions are incorrect, we then introduce a physical branch to verify the validity of the sampled target pose. Due to the non-differentiability of the simulator, we further design a pose decoder to intermediately employ physical supervision to train the distribution encoder.

Specifically, the KL divergence with pseudo ground-truth distributions is used to pre-train the encoder:

$$\mathcal{L}_{kl} = KL(\mathcal{Q}(\Delta\tilde{\boldsymbol{q}}_t \mid \boldsymbol{s}_{t-1}, \tilde{\boldsymbol{q}}_t) \| \mathcal{D}(\mu_t, \sigma_t)), \tag{5}$$

where $\mathcal{D}(\mu, \sigma)$ is the distribution prepared by (μ_W, λ)-CMA-ES method and $\mathcal{Q}(\Delta\tilde{\boldsymbol{q}} \mid \boldsymbol{s}, \tilde{\boldsymbol{q}})$ is the estimated distribution. To improve the generalization ability, we sample

correction of the reference pose from the estimated distribution, which is denoted as $\Delta\tilde{\boldsymbol{q}}_t$. Thus, the target pose is $\overline{\boldsymbol{q}}_t = \tilde{\boldsymbol{q}}_t + \Delta\tilde{\boldsymbol{q}}_t$.

To optimize the distribution encoder with real physical supervision, the sampled target pose is fed to the non-differentiable physics simulator to get the simulated pose. We design a pose decoder to imitate the physical branch by supervising it with the simulated pose.

$$\mathcal{L}_{sim} = \|\hat{\boldsymbol{q}}_t - \boldsymbol{q}_t\|^2 + \left\|\hat{\boldsymbol{j}}_t - \boldsymbol{j}_t\right\|^2, \tag{6}$$

where $\hat{\boldsymbol{q}}$, $\hat{\boldsymbol{j}}$ and $\boldsymbol{q}_t$, $\boldsymbol{j}$ are pose and joint positions of the estimated result and the simulated result, respectively. In addition, a reconstruction loss is applied to enforce optimal distribution encoding:

$$\mathcal{L}_{rec} = \|\hat{\boldsymbol{q}}_t - \tilde{\boldsymbol{q}}_t\|^2 + \left\|\hat{\boldsymbol{j}}_t - \tilde{\boldsymbol{j}}_t\right\|^2. \tag{7}$$

With the pose decoder, the encoder can gradually encode valid distribution to sample effective poses in the simulator. We further add a regularization term to ensure the network will not be easily overfitted:

$$\mathcal{L}_{reg} = \|\phi\|_2^2. \tag{8}$$

We reduce the weight of KL loss when training with the two-branch decoder. The overall loss function is:

$$\mathcal{L}_{dist} = \mathcal{L}_{sim} + \mathcal{L}_{rec} + \lambda\mathcal{L}_{kl} + \mathcal{L}_{reg}. \tag{9}$$

The λ is 0.2 in our experiments. When the training is finished, the encoder is utilized to construct a neural motion capture framework in Sec. 3.4.

3.4. Motion capture with neural motion control

With the trained distribution prior, we then capture human motion by tracking the kinematic reference motion by a sampling strategy. As shown in Fig. 2, the reference pose and the current state of character are first fed into the prior to encode target pose distribution. Then, we sample target poses and simulate them in the simulator. The quality of each sample is evaluated with character-level and image-level loss functions. The sample with the lowest loss will be adopted for the next frame. Since the reference motions from uneven terrains are noisy, we design several loss functions to evaluate sample quality.

The loss between simulated pose and reference pose is first used to measure the pose and joint position consistency.

$$\mathcal{L}_{tra} = \|\boldsymbol{q}_t - \tilde{\boldsymbol{q}}_t\|^2 + \left\|\boldsymbol{j}_t - \tilde{\boldsymbol{j}}_t\right\|^2. \tag{10}$$

We find that the dynamical state of the character is critical for physics-based motion capture. We then introduce a dynamical loss to evaluate the velocity consistency:

$$\mathcal{L}_{dyn} = \left\|\dot{\boldsymbol{q}}_t - \dot{\tilde{\boldsymbol{q}}}_t\right\|^2 + \left\|\dot{\boldsymbol{j}}_t - \dot{\tilde{\boldsymbol{j}}}_t\right\|^2, \tag{11}$$

where $\dot{\boldsymbol{q}}$ and $\dot{\boldsymbol{j}}$ are joint angular velocity and linear velocity, respectively. To let the physical character keep balance, we follow [35] to add a balance term to adjust CoM (Center of Mass):

$$\mathcal{L}_{ban} = \sum_{m=0}^{M} \left\|\boldsymbol{d}_t^m - \tilde{\boldsymbol{d}}_t^m\right\|^2 + \left\|\dot{\boldsymbol{j}}_t^{CoM} - \dot{\tilde{\boldsymbol{j}}}_t^{CoM}\right\|^2, \tag{12}$$

where, $\boldsymbol{d}^m = (\boldsymbol{j}^m - \boldsymbol{j}^{CoM})|_{z=0}$, which denotes the planar vector from end-effector m to CoM. The $\dot{\boldsymbol{j}}^{CoM}$ is the linear velocity of CoM and M is number of end-effectors.

Different from DRL, we can directly use image features to evaluate the quality of the sample. With 2D pose and corresponding confidence, the image-level loss makes our method more robust to occlusion scenarios:

$$\mathcal{L}_{reproj} = \sigma \left\|\Pi\left(\boldsymbol{j}_t\right) - \mathbf{p}_t\right\|^2. \tag{13}$$

The overall loss function for the sampling procedure is:

$$\mathcal{L}_{sam} = \mathcal{L}_{tra} + \mathcal{L}_{dyn} + \mathcal{L}_{ban} + \mathcal{L}_{reproj}. \tag{14}$$

Finally, the sample with the lowest loss in each frame consists of a complete physically plausible human motion.

4. Experiments

In this section, we conduct several qualitative and quantitative experiments to demonstrate the effectiveness of our method. We first introduce the implementation details and datasets in Sec. 4.1 and Sec. 4.2. Then, the comparisons with the state-of-the-arts are shown in Sec. 4.3. Finally, ablation studies in Sec. 4.4 are conducted to evaluate key components.

4.1. Metrics

The common metrics of the Mean Per Joint Position Error (MPJPE) and the MPJPE after rigid alignment of the prediction with ground truth using Procrustes Analysis (MPJPE-PA) are used to evaluate joint accuracy. To evaluate physical plausibility, we use the metrics proposed in [47] and [57] to measure motion jitter and foot contact. e_S is the difference in joint velocity magnitude between the ground truth motion and the predicted motion. e_S and its standard deviation σ_S are used to assess motion smoothness. $e_{f,z}$ is the foot position error on z-axis. We adopt this metric to evaluate foot floating artifacts. More details can be found in their original paper.

4.2. Datasets

Human3.6M [17] is a large-scale dataset, which consists of 3.6 million 3D human poses and corresponding images. Following previous work [63], we train our model

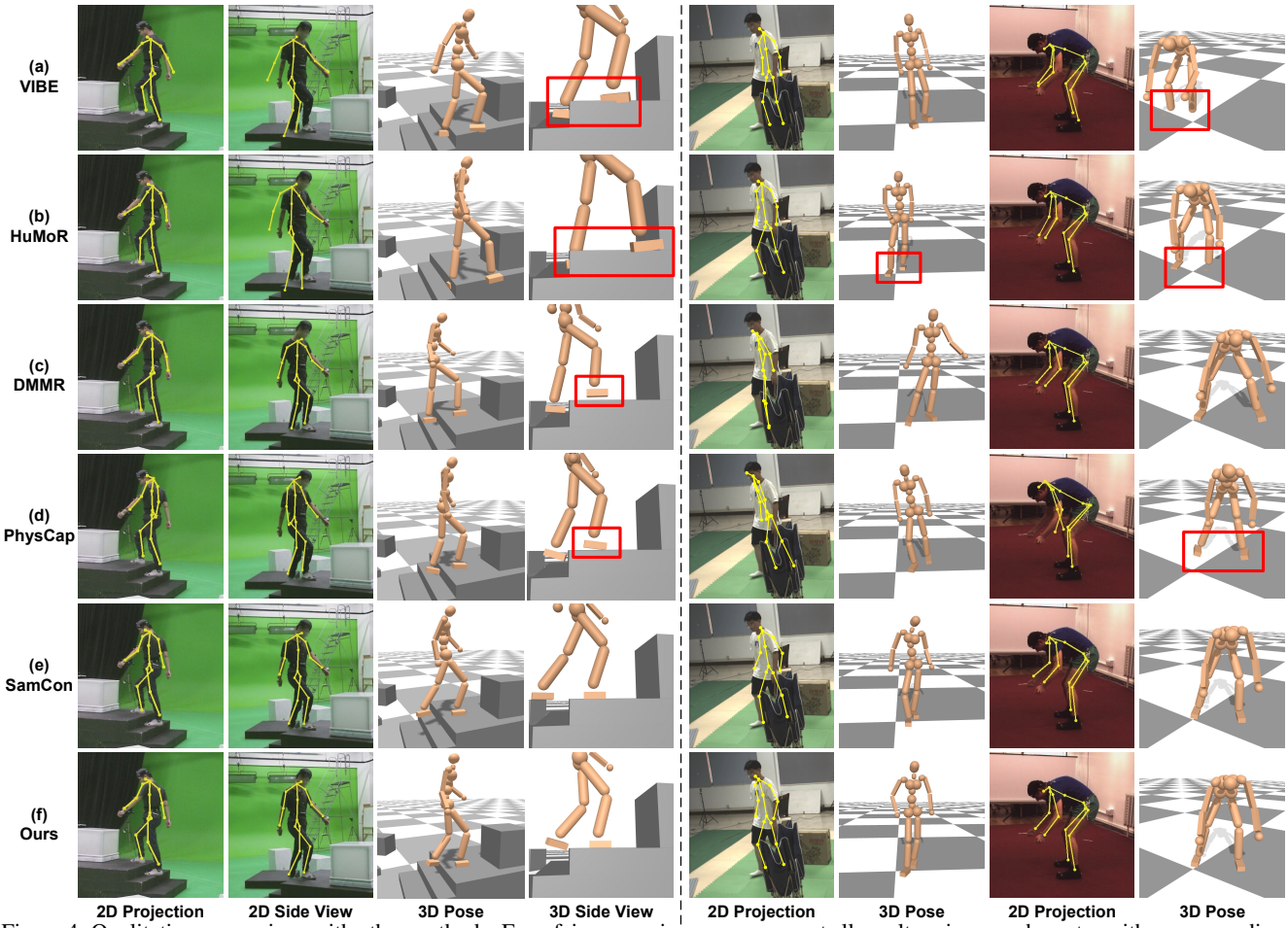

Figure 4. Qualitative comparison with other methods. For a fair comparison, we represent all results using our character with corresponding shape variations. The results show that our method can obtain physically plausible and natural human motion from monocular RGB videos.

on 5 subjects (S1,S5,S6,S7,S8), and test on the other subjects (S9,S11) with 25Hz.

GPA [52] is a 3D human dataset with both human-scene interactions and ground-truth scene geometries. It utilizes a commercial motion capture system to collect data. The sequence *0, 34, 52* are used to test, and the rest are served as training data. With the scene geometries, we verify the performance of our method on more complex terrains.

3DOH [66] is the first dataset to handle the object-occluded human body estimation problem, which contains 3D motions in occluded-scenarios. We use the sequence *13, 27, 29* in this dataset to evaluate our method on occlusion cases.

GTA-IM [2]. Since there are limited ground-truth terrain data, we use this synthetic dataset as additional human-scene interaction cases. The scene meshes are recovered from the depth map. We conduct qualitative experiments on this dataset.

4.3. Comparison to state-of-the-art methods

There are several kinematic and dynamical approaches that report results on Human3.6M datasets. As shown in Tab. 1, we first evaluated our method on this dataset to demonstrate that our neural motion control works well on flat ground. [16, 25, 41] are recent works to estimate kinematic SMPL parameters. Although the explicit dynamics of the human model are not considered, [16, 41] learn implicit dynamics via VAE and improve physical plausibility by using prior knowledge. The rest methods in Tab. 1 are dynamics-based methods. Specifically, SamCon [33] is de-

Method	MPJPE	PA-MPJPE	e_S	σ_S	$e_{f,z}$
*HuMoR [41]	97.5	68.5	24.2	25.9	43.2
*DMMR [16]	96.0	67.4	**14.4**	**12.6**	48.6
*VIBE [25]	**65.9**	**41.5**	25.5	25.7	**34.0**
EgoPose [61]	130.3	79.2	–	–	–
PhysCap [47]	97.4	65.1	7.2	6.9	–
SamCon [33]	78.4	63.2	4.0	4.3	20.4
NeuralPhysCap [46]	76.5	58.2	4.5	6.9	–
Xie *et al*. [57]	68.1	–	4.0	**1.3**	18.9
SimPoE [63]	**56.7**	**41.6**	–	–	–
Ours	72.5	54.6	**3.8**	2.4	**14.4**

Table 1. Comparisons with state-of-the-art methods on Human3.6M dataset. Our method achieves good performance in physical plausibility and motion smoothness. * denotes the kinematics-based method.

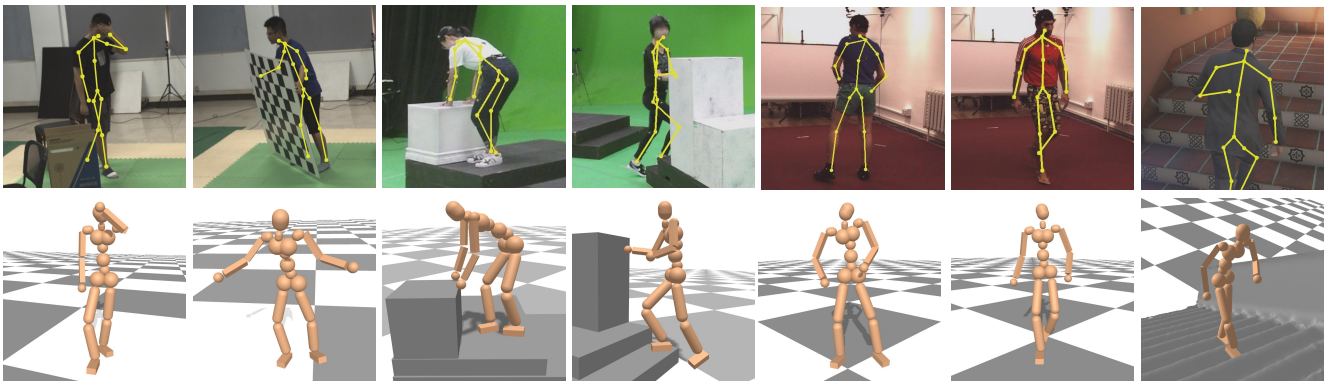

Figure 5. Our method is general to different terrain interactions, human shape variations, and diverse behaviors.

Method	*3DOH*			*GPA*		
	MPJPE	PA-MPJPE	e_S	MPJPE	PA-MPJPE	$e_{f,z}$
*DMMR [16]	102.9	65.8	**16.2**	**107.0**	87.4	**32.8**
*VIBE [25]	**98.1**	61.8	26.5	114.3	**80.6**	36.4
*HuMoR [41]	105.1	**60.6**	21.9	117.2	86.3	58.7
SamCon [33]	102.4	95.4	9.7	104.7	87.1	28.3
PhysCap [47]	107.8	93.3	12.2	103.4	91.2	36.1
Ours	**93.4**	**86.7**	**9.2**	**94.8**	**80.3**	**21.2**

Table 2. Quantitative comparison on 3DOH and GPA dataset. Our method achieves state-of-the-art in complex terrain scenarios and occlusion cases. * denotes the kinematics-based method.

signed for animation. We used this method to track our kinematic motion and adopted it as a baseline to compare among sampling-based methods.

In Tab. 1, we found that VIBE achieves the best performance in terms of PA-MPJPE. It relies on a GRU-based network to build correspondences among different frames. However, directly regressing kinematic SMPL parameters causes the largest smoothness error and results in visually noticeable motion jitter. Furthermore, VIBE shows a severe penetration with the ground in Fig. 4. Due to model discrepancies between the motion capture subject and the physical character, the joint position error for dynamics-based methods is higher than kinematics-based approaches. SimPoE [63] utilizes a model with a similar shape as Human3.6M subjects and get comparable results to VIBE. However, for different subjects with the variation of body proportion and shape, this method requires to re-train the policy. Benefited from the proposed target pose distribution prior, our method can adapt to shape variation. Thus, we can update the bone length of the physical character model with the estimated human shape and directly use it to capture human motion from images. Our method also obtained smooth motion and achieves state-of-the-art in terms of e_S.

We then compared our method to others on the 3DOH dataset. It is tricky to obtain accurate reference motion for occlusion cases. As shown in the 5th column of Fig. 4, the inaccurate reference motion will result in a large deviation between 3D pose and image observation for other physics-based methods. However, due to the image-level loss, our method got more accurate results. Moreover, SamCon also based on a sampling approach to get human motion. The results in Tab. 2 and Fig. 4 show that our network-based distribution prior can get more appropriate distribution and then produce natural and precise motion.

On the GPA dataset, we evaluated our method with complex terrains. The interactions with objects and terrains impose great difficulty for kinematics-based methods. The estimated poses float on the air or penetrate with the scene mesh for their methods (Fig. 4). Since PhysCap uses a numerical optimization framework with soft physical constraints to capture human motion, the results also show physical artifacts. The qualitative and quantitative results on GPA dataset in Fig. 4 and Tab. 2 show that neural motion control is more proper for contact-rich scenarios.

4.4. Ablation studies

Two-branch decoder. As mentioned before, directly supervising the distribution encoder without two-branch decoder will result in erroneous regression. In Fig. 6 and Tab. 3, we conducted comparisons between the distribution prior trained with and without the two-branch decoder. Without the decoder, the encoder can not regress correct distribution to sample a valid target pose, thus causing an unsatisfied simulated pose. The quantitative results in Tab. 3 show that the two-branch decoder induces major improvement and demonstrate that it is the most important component for our method.

Distribution prior. We compared different methods of distribution generation to verify the superiority of our distribution prior. We first replaced the distribution encoder with uniform distribution with a pre-defined range. The results in Fig. 6 show that it can not generalize to a large variety motion types. As shown in Tab. 3, since there is a stochastic error for the CMA method, the gaussian distribution with CMA adaption is inferior to the distribution encoder.

Interaction constraint. We further conducted several experiments to illustrate the necessity of the interaction constraint. Due to the visual ambiguity, it is difficult to reconstruct accurate human-scene interactions with complex terrains (Fig. 7). In Tab. 3, the optimization with interaction constraint gets more accurate foot position on GPA dataset.

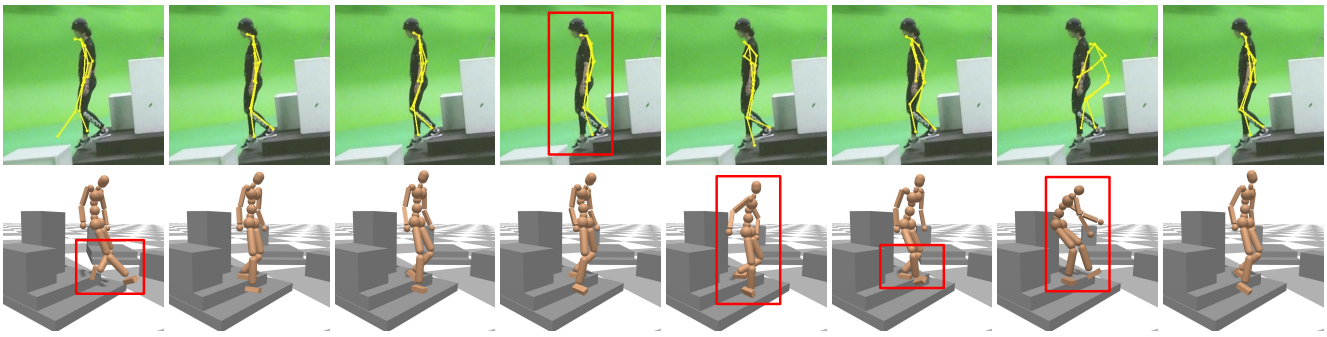

Figure 6. Ablation on different components. (a, d, h) are the results of our method that removes interaction constraint, image-level loss, and two-branch decoder, respectively. (e, f) replace the distribution prior with uniform distribution and gaussian distribution. (c) is the kinematic result from our optimization and (b) is the simulated result with ground-truth reference motion.

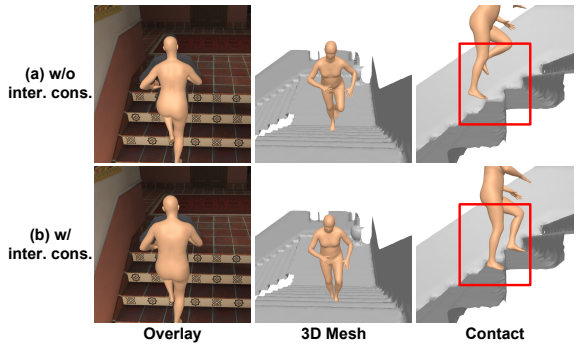

Figure 7. The kinematic reference motion obtained with and without interaction constraint on complex terrain.

Method	*GPA* MPJPE	*GPA* PA-MPJPE	*GPA* $e_{f,z}$	*3DOH* MPJPE	*3DOH* e_S
*DMMR [16]	107.0	87.4	32.8	102.9	16.2
*Kinematic	106.2	87.2	27.3	94.4	16.5
w/o two-branch	142.2	126.7	28.4	136.8	13.2
w/ Uniform Dist.	136.6	119.1	29.6	142.1	10.3
w/o Inter. Cons.	116.4	109.4	24.3	93.4	9.4
w/ Gaussian CMA	103.9	84.4	23.5	95.4	9.8
w/o image-level loss	95.8	84.4	21.3	96.3	9.7
w/ GT reference	93.6	80.0	17.3	89.6	9.2
Neural MoCon	94.8	80.3	21.2	93.4	9.2

Table 3. Quantitative results of ablation studies. w/o denotes to remove corresponding component of our method. w/ Uniform Dist. and w/ Gaussian CMA indicate to replace distribution prior with uniform and gaussian distribution. w/ GT reference uses ground-truth reference motion for neural motion control. * denotes the kinematics-based method.

In addition, an inaccurate contact seriously affects the performance of sampling-based motion control. Fig. 6 (a) shows a reference pose floating on the air can trigger improper simulated pose. The gap between the results of the method with and without this constraint on GPA is greater than that on 3DOH in Tab. 3, which proves its importance for motion capture on complex terrains.

5. Limitation and future work

Although our method can obtain physically plausible human motion via neural motion control, there are some limitations for the current implementation. First, the discrepancy between the geometric primitives of our character and the real human body makes our method unable to reconstruct accurate body contact (*e.g.*, Lying on the sofa). To solve this problem, building a more delicate character model like [63] may be a feasible approach. Second, the cumulative error of an undesirable sample may result in failure to sample a long sequence. Future work can integrate long-term temporal information in the sampling. Finally, due to a lack of ground-truth terrain data, we can only evaluate our method on similar interactions like stairs for motion capture tasks. Therefore, to build a large-scale human-scene interaction dataset for human motion capture in complex scenarios is also worthwhile.

Among Neural MoCon, DRL-based methods, and traditional sampling-based motion control, DRL can obtain highly accurate results for a specific task, and sampling control is more general to unknown scenarios. Neural MoCon is in between these two typical technical approaches. To combine the accuracy of DRL and the generalization ability of sampling control may be a potential direction to promote future physics-based motion capture.

6. Conclusion

In this paper, we propose a framework to capture physically plausible human motion with complex terrain interactions, human shape variations, and diverse behaviors. We first introduce an interaction constraint based on SDF in optimization to estimate accurate human-scene contact. Then, a novel two-branch decoder is designed to train a distribution prior with real physical supervision. With the trained prior and the estimated reference motion, several loss functions are used to select a satisfied sample to consist of a complete human motion. The proposed method has better generalization ability than DRL-based methods and gets more accurate results than conventional sampling-based motion control.

References

[1] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters for on-policy deep actor-critic methods? a large-scale study. In *ICLR*, 2020. 1, 2

[2] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 6

[3] Stelian Coros, Philippe Beaudoin, and Michiel Van de Panne. Generalized biped walking control. *ACM TOG*, 29(4):1–9, 2010. 2

[4] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021. 3

[5] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019. 1, 2

[6] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. Revitalizing optimization for 3d human pose and shape estimation: A sparse constrained formulation. In *ICCV*, 2021. 1

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[8] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM TOG*, 34(6):1–13, 2015. 3

[9] Stuart Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 4:5–21, 1987. 4

[10] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR*, 2011. 3

[11] Nikolaus Hansen. The cma evolution strategy: a comparing review. *Towards a new evolutionary computation*, pages 75–102, 2006. 2, 4

[12] Kensuke Harada, Shuuji Kajita, Kenji Kaneko, and Hirohisa Hirukawa. An analytical method for real-time gait planning for humanoid robots. *International Journal of Humanoid Robotics*, 3(01):1–19, 2006. 2

[13] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 3

[14] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 3, 4

[15] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, 2021. 3

[16] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *3DV*, 2021. 4, 6, 7, 8

[17] 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 TPAMI*, 36(7):1325–1339, jul 2014. 5

[18] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 4

[19] Yun Jiang, Hema Koppula, and Ashutosh Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, 2013. 3

[20] Yun Jiang, Marcus Lim, and Ashutosh Saxena. Learning object arrangements in 3d scenes using human context. *arXiv preprint arXiv:1206.6462*, 2012. 3

[21] Shuuji Kajita and Kazuo Tani. Study of dynamic biped locomotion on rugged terrain-derivation and application of the linear inverted pendulum mode. In *ICRA*, 1991. 2

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

[23] Hyung Joo Kim, Qian Wang, Salam Rahmatalla, Colby C Swan, Jasbir S Arora, Karim Abdel-Malek, and Jose G Assouline. Dynamic motion planning of 3d human locomotion using gradient-based optimization. *Journal of biomechanical engineering*, 130(3), 2008. 2

[24] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2pose: Human-centric shape analysis. *ACM TOG*, 33(4):1–12, 2014. 3

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 6, 7

[26] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1

[27] Arthur D Kuo. A simple model of bipedal walking predicts the preferred speed–step length relationship. *J. Biomech. Eng.*, 123(3):264–269, 2001. 2

[28] Kyungho Lee, Sehee Min, Sunmin Lee, and Jehee Lee. Learning time-critical responses for interactive character control. *ACM TOG*, 40(4):1–11, 2021. 2

[29] Yoonsang Lee, Moon Seok Park, Taesoo Kwon, and Jehee Lee. Locomotion control for many-muscle humanoids. *ACM TOG*, 33(6):1–11, 2014. 2

[30] Sergey Levine and Jovan Popović. Physically plausible simulation for character animation. In *SCA*, 2012. 2

[31] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017. 1, 2

[32] 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 *CVPR*, 2019. 2

[33] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. In *Comput. Graph. Forum*, volume 34, pages 415–423, 2015. 2, 3, 4, 6, 7

[34] Libin Liu, KangKang Yin, Michiel van de Panne, and Baining Guo. Terrain runner: control, parameterization, composition, and planning for highly dynamic motions. *ACM TOG*, 31(6):154–1, 2012. 2

[35] Libin Liu, KangKang Yin, Michiel van de Panne, Tianjia Shao, and Weiwei Xu. Sampling-based contact-rich motion control. In *SIGGRAPH*, 2010. 1, 2, 3, 5

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

[37] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM TOG*, 38(4):1–15, 2019. 3

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1

[39] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM TOG*, 37(4):1–14, 2018. 1, 2

[40] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM TOG*, 37(6):1–14, 2018. 1, 2

[41] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 1, 6, 7

[42] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, 2020. 1, 2

[43] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Scenegrok: Inferring action maps in 3d environments. *ACM TOG*, 33(6):1–10, 2014. 3

[44] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM TOG*, 35(4):1–12, 2016. 3

[45] Dana Sharon and Michiel van de Panne. Synthesis of controllers for stylized planar bipedal walking. In *ICRA*, 2005. 2

[46] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *arXiv preprint arXiv:2105.01057*, 2021. 1, 2, 6

[47] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM TOG*, 39(6):1–16, 2020. 1, 2, 4, 5, 6, 7

[48] Kwang Won Sok, Manmyung Kim, and Jehee Lee. Simulating biped behaviors from human motion data. In *SIGGRAPH*. 2007. 2

[49] Michiel Van de Panne and Eugene Fiume. Sensor-actuator networks. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, 1993. 2

[50] Marek Vondrak, Leonid Sigal, Jessica Hodgins, and Odest Jenkins. Video-based 3d motion capture through biped control. *ACM TOG*, 31(4):1–12, 2012. 2

[51] Yangang Wang, Yebin Liu, Xin Tong, Qionghai Dai, and Ping Tan. Outdoor markerless motion capture with sparse handheld video cameras. *IEEE TVCG*, 24(5):1856–1866, 2017. 1

[52] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 6

[53] Xiaolin Wei and Jinxiang Chai. Videomocap: Modeling physically realistic human motion from monocular video sequences. In *SIGGRAPH*, 2010. 1, 2

[54] Pawel Wrotek, Odest Chadwicke Jenkins, and Morgan McGuire. Dynamo: dynamic, data-driven character control with adjustable balance. In *Proceedings of the 2006 ACM SIGGRAPH symposium on Videogames*, 2006. 2

[55] Yujiang Xiang, Jasbir S Arora, and Karim Abdel-Malek. Physics-based modeling and simulation of human walking: a review of optimization-based and other approaches. *Structural and Multidisciplinary Optimization*, 42(1):1–23, 2010. 2

[56] Yujiang Xiang, Jasbir S Arora, Salam Rahmatalla, and Karim Abdel-Malek. Optimization-based dynamic human walking prediction: One step formulation. *International Journal for Numerical Methods in Engineering*, 79(6):667–695, 2009. 2

[57] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *ICCV*, 2021. 1, 5, 6

[58] Zhaoming Xie, Hung Yu Ling, Nam Hee Kim, and Michiel van de Panne. Allsteps: Curriculum-driven learning of stepping stone skills. In *Comput. Graph. Forum*, volume 39, pages 213–224, 2020. 2, 3

[59] Yuting Ye and C Karen Liu. Optimal feedback control for character animation using an abstract model. In *SIGGRAPH*, 2010. 2

[60] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM TOG*, 40(6), 2021. 1, 2, 4

[61] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *ICCV*, 2019. 6

[62] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. *arXiv preprint arXiv:2006.07364*, 2020. 2

[63] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *CVPR*, 2021. 1, 2, 5, 6, 7, 8

[64] Petrissa Zell, Bastian Wandt, and Bodo Rosenhahn. Joint 3d human motion capture and physical analysis from monocular videos. In *CVPR*, 2017. 2

[65] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *ICCV*, 2021. 3

[66] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 1, 6

[67] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, 2019. 1