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/HB2-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).

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 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 control framework that incorporates physically plausible motion generation. We first introduce a novel method to capture physically plausible human motion from monocular RGB videos via neural motion control.

Figure 1. Our method captures physically plausible human motion from monocular RGB videos via neural motion control.
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 adaptation, 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
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 $s = (\theta, \dot{\theta})$, where $\theta$ and $\dot{\theta}$ are the pose and velocity, respectively.

**Sampling-based motion control.** We briefly review the sampling-based motion control approach [35] to promote understanding of our method. A kinematic pose $\tilde{q}_t$ 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 $\Delta \tilde{q}_t$ for reference pose, thus employing the target pose $\tilde{q}_t = \tilde{q}_t + \Delta \tilde{q}_t$ 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:

$$\arg \min_{(z,R,T)} L = L_{data} + L_{prior} + L_{scene},$$

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

$$L_{data} = \sum_{t=1}^{T} \sigma(t) \| \Pi(\tilde{j}_t) - p_t \|^2,$$

where $p_t$, $\sigma(t)$ are 2D poses and their corresponding confidence. $\tilde{j}_t$ is the model joint position. We further add the regularization term:

$$L_{prior} = \| \beta \|^2 + \sum_{t=1}^{T} \left( \| z_t \|^2 + \| z_{t+1} - 2 z_t + z_{t-1} \|^2 \right).$$

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:

$$L_{scene} = \rho \| \text{SDF}(\tilde{j}) \|^2,$$

where $\tilde{j}$ 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 $\rho$ [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 $(\mu_W, \lambda)$-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.

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:

$$L_{kl} = KL(Q(\Delta \tilde{q}_t | s_{t-1}, \tilde{q}_t) || D(\mu_t, \sigma_t)),$$

where $D(\mu, \sigma)$ is the distribution prepared by $(\mu_W, \lambda)$-CMA-ES method and $Q(\Delta \tilde{q} | s, \tilde{q})$ is the estimated distribution. To improve the generalization ability, we sample...
correction of the reference pose from the estimated distribution, which is denoted as $\Delta \tilde{q}_t$. Thus, the target pose is $q_t = \tilde{q}_t + \Delta \tilde{q}_t$.

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.

$$L_{sim} = \| \tilde{q}_t - q_t \|^2 + \| \tilde{j}_t - j_t \|^2,$$  \hspace{1cm} (6)

where $\tilde{q}_t$, $\tilde{j}_t$ and $q_t$, $j_t$ 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:

$$L_{rec} = \| q_t - \tilde{q}_t \|^2 + \| j_t - \tilde{j}_t \|^2.$$  \hspace{1cm} (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:

$$L_{reg} = \| \phi \|^2_2.$$  \hspace{1cm} (8)

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

$$L_{dist} = L_{sim} + L_{rec} + \lambda L_{kl} + L_{reg}. \hspace{1cm} (9)$$

The $\lambda$ 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.

$$L_{tra} = \| q_t - \tilde{q}_t \|^2 + \| j_t - \tilde{j}_t \|^2.$$  \hspace{1cm} (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:

$$L_{dyn} = \| \tilde{q}_t - \dot{q}_t \|^2 + \| \tilde{j}_t - \dot{j}_t \|^2,$$  \hspace{1cm} (11)

where $\dot{q}$ and $\dot{j}$ 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):

$$L_{ban} = \sum_{m=0}^{M} \left( d_m^m - d_t^m \right)^2 + \| \tilde{\dot{\text{CoM}}} - j_t^\text{CoM} \|^2,$$  \hspace{1cm} (12)

where, $d_m^m = (j_m^m - \dot{j}_m^\text{CoM})|_{z=0}$, which denotes the planar vector from end-effector $m$ to CoM. The $\dot{j}_m^\text{CoM}$ 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:

$$L_{reproj} = \sigma \| \Pi (\tilde{j}_t) - p_t \|^2.$$  \hspace{1cm} (13)

The overall loss function for the sampling procedure is:

$$L_{sam} = L_{tra} + L_{dyn} + L_{ban} + L_{reproj}. \hspace{1cm} (14)$$

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. $e_g$ is the difference in joint velocity magnitude between the ground truth motion and the predicted motion. $e_S$ and its standard deviation $\sigma_S$ 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...
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-
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>3DOH</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MPJPE</td>
<td>PA-MPJPE</td>
</tr>
<tr>
<td>*DMMR [16]</td>
<td>102.9</td>
<td>65.8</td>
</tr>
<tr>
<td>*VIBE [25]</td>
<td>98.1</td>
<td>61.8</td>
</tr>
<tr>
<td>*HuMoR [41]</td>
<td>105.1</td>
<td>60.6</td>
</tr>
<tr>
<td>SamCon [33]</td>
<td>102.4</td>
<td>95.4</td>
</tr>
<tr>
<td>PhysCap [47]</td>
<td>107.8</td>
<td>93.3</td>
</tr>
<tr>
<td>Ours</td>
<td>93.4</td>
<td>86.7</td>
</tr>
</tbody>
</table>
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.

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.

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

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


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


