Learning Motion Priors for 4D Human Body Capture in 3D Scenes

Siwei Zhang¹  Yan Zhang¹  Federica Bogo²  Marc Pollefeys¹,²  Siyu Tang¹

¹ETH Zürich  ²Microsoft
{siwei.zhang, yan.zhang, marc.pollefeys, siyu.tang}@inf.ethz.ch  febogo@microsoft.com

Abstract

Recovering high-quality 3D human motion in complex scenes from monocular videos is important for many applications, ranging from AR/VR to robotics. However, capturing realistic human-scene interactions, while dealing with occlusions and partial views, is challenging; current approaches are still far from achieving compelling results. We address this problem by proposing LEMO: LEarning human MOTion priors for 4D human body capture. By leveraging the large-scale motion capture dataset AMASS [38], we introduce a novel motion smoothness prior, which strongly reduces the jitters exhibited by poses recovered over a sequence. Furthermore, to handle contacts and occlusions occurring frequently in body-scene interactions, we design a contact friction term and a contact-aware motion infiller obtained via per-instance self-supervised training. To prove the effectiveness of the proposed motion priors, we combine them into a novel pipeline for 4D human body capture in 3D scenes. With our pipeline, we demonstrate high-quality 4D human body capture, reconstructing smooth motions and physically plausible body-scene interactions. The code and data are available at https://sanweiliti.github.io/LEMO/LEMO.html.

1. Introduction

Recovering realistic human motions in everyday 3D scenes is essential for human behaviour understanding, human-scene interaction synthesis, and virtual avatar creation. Marker-based optical motion capture systems (mocap) have the proven capability of recovering highly accurate human motions. However, such systems require expert knowledge and expensive setup, making it impractical to capture people in their everyday environments, e.g., recording people in their living rooms, offices or kitchens.

Recently, PROX [19] has been proposed as a lightweight pipeline to capture everyday person-scene interactions from monocular sequences given pre-scanned 3D scene geometries. With affordable commodity sensors like an RGB or RGBD camera, it is quite easy to scan a scene and record how humans move in and interact with it. This shows a promising setup for capturing large-scale human motions in everyday environments. However, as shown in this work¹, the recovered human motions exhibit severe skating and jitters. The reconstruction quality is far behind that obtained with commercial mocap systems. Building a multi-view setup or using additional wearable sensors (e.g., Inertial Measurement Unit (IMUs)) can help improve motion reconstruction quality. However, most multi-view settings

¹please see videos on the project page
require careful calibration and synchronization in a controlled environment and IMUs suffer from heading drift and interference. Furthermore, human motions obtained by IMUs [60] or multi-view setups [23] still exhibit jitters and are less compelling than the ones from mocap systems.

To improve the naturalness and accuracy of human motions reconstructed from monocular RGBD sequences (e.g. the PROX pipeline [19]) and to close the performance gap between the monocular RGBD setup and marker-based mocap systems, we argue that it is essential to leverage data-driven approaches and learn powerful motion priors from high-quality large-scale mocap data (e.g. AMASS [38]). To this end, we propose LEMO (LEarning human MOtion priors), which has two key innovations: a marker-based motion smoothness prior and a contact-aware motion infiller which is fine-tuned per-instance in a self-supervised fashion. As shown in the experiments, LEMO can effectively capture the intrinsic properties of human motions and regularize the noisy and partial observations. As a result, the reconstructed human motions are smooth, physically plausible and robust to occlusions which are inevitable when capturing human motions in everyday 3D scenes.

**Marker-based motion smoothness prior.** 3D human bodies reconstructed by PROX [19] have severe jitters over time. Although some heuristic methods like penalizing joint velocity/acceleration can improve temporal smoothness, they also degrade the motion naturalness. As shown in our experiments, they can introduce foot-ground skating artifacts, and may result in invalid body configurations like joint hyperextension. To capture the holistic full-body dynamics, we use a fully convolutional autoencoder to aggregate local motion cues in a bottom-up manner, and derive latent motion patterns that cover a large spatio-temporal receptive field. Then, we design a motion smoothness constraint which works in this latent space rather than directly on the body. To incorporate body shape information and model important degrees-of-freedom (DoFs), e.g. rotation about limb axes, as in [71] we represent the body in each frame by surface markers instead of body joints. We learn this convolutional motion smoothness prior on the AMASS [38] dataset. As shown in our experiments, the proposed prior not only significantly increases the reconstruction quality on the PROX dataset, but also improves the motion naturalness on the IMU-based 3DPW dataset [60], suggesting its effectiveness and potential usage for other motion capture and reconstruction settings.

**Contact-aware motion infiller via per-instance self-supervised learning.** When capturing humans moving in and interacting with everyday 3D environments (e.g. living rooms or offices), partial body occlusions are almost inevitable. They pose a challenge for reconstruction algorithms, causing invalid poses and foot-ground skating artifacts. By leveraging AMASS [38], we learn a neural motion infiller that is able to infer plausible motions of occluded body parts given partial observations. Our network is inspired by [28], but goes beyond the previous work to jointly predict the foot contact status and body motion. Combined with a contact friction term motivated by intuitive physics, the infilled motion is natural, realistic and has proper foot-ground interaction, eliminating the foot skating artifacts. Furthermore, inspired by [24], we propose a per-instance network fine-tuning scheme. For a test instance which contains partial observations (e.g. only the upper body motion as the lower body parts are occluded by the sofa in a 3D scene), we fine-tune the pre-trained motion infilling network by minimizing a self-supervised loss that is defined on the visible body parts. In this way, we effectively adapt the general motion infilling “prior” to per-test-instance, achieving notable improvements both for AMASS [38] and PROX [19].

We further carefully combine the learned motion priors and the contact friction term into a novel multi-stage optimization pipeline for 4D human body capture in 3D scenes.

**Contributions.** In summary, our contributions are 1) a novel marker-based motion smoothness prior that encodes the “whole-body” motion in a learned latent space, which can be easily plugged into an optimization pipeline; 2) a novel contact-aware motion infiller that can be adapted to per-test-instance via self-supervised learning; 3) a new optimization pipeline that explores both learned motion priors and the physics-inspired contact friction term for scene-aware human motion capture. We extensively evaluate the proposed priors and the optimization pipeline. The results show both the wide applicability of the learned motion priors and the efficacy of the optimization pipeline for monocular RGBD human motion capture in 3D scenes.

2. Related Work

**Human motion recovery from RGB(D) sequences.** Human motion recovery extends the problem of reconstructing per-frame body 3D shape and pose [2, 5, 16, 17, 19, 26, 32, 42, 44, 53, 56, 58, 66] to sequences of frames, requiring temporal consistency between estimates. A number of works tackle the problem adopting skeleton/joint-based representations for the body [7, 8, 11, 13, 14, 29, 35, 39–41, 45, 46, 54, 63, 65, 70, 73]. Working with 3D joints instead of surfaces, these representations cannot adequately model the 3D shape of the body and body-scene interactions. Other works propose to use parametric 3D human models (e.g. SMPL [36]) to obtain complete 3D body meshes from multi-view [12, 15, 22, 25, 50, 64] or monocular RGB(D) sequences [10, 27, 31, 34, 37, 55, 67]. Kanazawa et al. [27] learn a temporal context representation to predict motion in past and future frames. Kocabas et al. [31] use a bi-directional gated recurrent unit (GRU)
to temporally encode per-frame image features, and couple it with an adversarial discriminator to distinguish between real and predicted motions. Choi et al. [10] propose to better integrate past and future frames’ temporal information to increase temporal consistency. Sun et al. [55] introduce a multi-level framework to decouple body skeleton and more detailed shape and pose information. Luo et al. [37] propose a two-step encoding scheme, which first captures the coarse overall motion by a pretrained motion representation, and then refines these estimates. Nevertheless, these methods focus only on human body motion reconstruction, ignoring person-scene interactions.

**Person-scene interaction.** Hasler et al. [18] obtain scene constraints for body pose estimation by reconstructing the scene in 3D with multiple unsynchronized moving cameras. Some works rely on physics-inspired error terms (e.g., contact and collision terms [68]), game physics engines [61], and scene semantic labels [51]. Related to us, PROX [19] captures person-scene interactions at a very detailed level, modelling contacts and collisions between SMPL-X body [44] and 3D scenes. Based on such contact and collision modelling, Zhang et al. [69, 72] generate human body meshes in scenes without people in a physically and semantically plausible manner.

**Human motion priors.** A large number of priors for smooth and natural motion have been proposed in the literature [3, 4, 22, 41, 43, 47, 48, 52, 59]. Some priors directly apply to body joint velocity or acceleration [4, 41]. Akhter et al. [3] propose a bilinear model with discrete cosine transform (DCT) basis to provide motion spatio-temporal regularity. Among this line, Huang et al. [22] introduce a DCT prior to reconstruct body motion from multi-view input. Some recent work exploits physical simulation to regularize human motion. Shimada et al. [52] assume a pre-defined virtual character as input, and fit it to monocular sequences via physics-based optimization. Rempe et al. [47] regress body joints and foot-ground contact from images to conduct physics-based trajectory optimization. Kaufmann et al. [28] design a convolutional autoencoder to infill motion of unobserved body joints and remove noise.

**Ours versus others.** In our work we design a motion smoothness prior and a motion infiller, and use them to recover realistic motions of person-scene interactions from RGB-D videos. Compared to existing smoothness priors, ours is trained with high-quality AMASS sequences and the smoothness regularization is applied in the latent space. Consequently, we can produce smooth motions without degrading per-frame body pose accuracy. Our motion infiller has a similar architecture to Kaufmann et al. [28], but processes body markers and predicts foot-ground contact states. Since body markers better constrain body DoFs, and the contact states are jointly learned with body motions, our method consistently outperforms [28] w.r.t. motion recovery and foot skating (as shown in Sec. 4). Compared with [47, 52] which predict contact states by 2D joints detected from RGB images, our jointly learned contact states are better coupled with body dynamics.

### 3. Method

#### 3.1. Overview

We provide an overview of our approach in Fig. 2. Given a sequence of RGB-D frames \( \{I_t, D_t\}_{t=1}^{T} \), capturing a subject moving in a 3D scene, our goal is to reconstruct a high-quality motion, which is smooth, physically plausible, and natural. To this end, we fit the SMPL-X parametric body model to sequence data by proceeding in three stages.

**SMPL-X.** SMPL-X [44] represents the body as a function \( \mathcal{M}_b(\gamma, \beta, \phi) \), whose output is a triangle mesh with vertices \( V_b \in \mathbb{R}^{10475 \times 3} \). SMPL-X parameters are global translation \( \gamma \in \mathbb{R}^{3} \), body shape \( \beta \in \mathbb{R}^{10} \), body and hand pose \( \phi \), and facial expression \( \Phi \in \mathbb{R}^{10} \). We denote by \( J(\beta) \) the 3D body skeleton joints in the neutral pose, and by \( R_{\phi}(J(\beta_i)) \) the \( i \)-th joint posed according to pose \( \phi \) and translation \( \gamma \).

**Multi-stage pipeline.** Given the complexity of our task, we address it in a multi-stage fashion, as done in previous work [6, 52]. In Stage 1, we fit SMPL-X parameters to each RGB-D frame independently. This gives us a reasonable initialization, but does not ensure motion smoothness, nor deals with body-scene occlusions. We achieve temporally consistent motions in Stage 2 by introducing our smoothness prior and contact-friction term. Finally, in Stage 3 we recover plausible motions even for occluded body parts and alleviate foot skating with our motion infiller.

#### 3.2. Per-frame Fitting

Stage 1 adopts the approach proposed in PROX [19]. Given a RGB-D sequence, PROX fits SMPL-X to each frame separately by minimizing the objective function:

\[
E_{\text{PROX}}(\gamma, \beta, \phi) = E_J + \lambda_D E_D + E_{\text{prior}} + \lambda_{\text{contact}} E_{\text{contact}} + \lambda_{\text{coll}} E_{\text{coll}}.
\]

\( E_J \) penalizes the distance between the 2D joints estimated from the RGB image with OpenPose [9], and the 2D projection of SMPL-X joints onto the image. \( E_D \) penalizes the 3D distance between the human point cloud obtained from the depth frame and SMPL-X surface points visible from the camera. \( E_{\text{prior}} \) combines a set of priors regularizing body pose, shape and facial expression [44]. \( E_{\text{contact}} \) encourages contact between scene vertices and a pre-defined set of body “contact” vertices. \( E_{\text{coll}} \) penalizes scene-body interpenetration. For more details, we refer the reader to [19].

#### 3.3. Temporally Smooth Motion

In Stage 2, we process the output of Stage 1. In order to obtain smooth and realistic motion, we design a mo-
The second term minimizes the 1st order derivative of the discrepancy between the first term is the reconstruction loss minimizing layers. More details can be found in the Supp. Mat.

The input to our network is a sparse set of body surface markers, as in [38, 71]. We represent the body with the locations of 81 markers (see marker placement in Supp. Mat.). Given a sequence of $T$ frames, at each time $t$ we compute the time difference of marker locations and concatenate them to a vector of length $S$. Then the entire sequence is represented by a 2D feature map $X_\Delta \in \mathbb{R}^{S \times (T-1)}$. The network encoder $F_s$ converts $X_\Delta$ to its latent representation $Z$. Here we regard the time series $X_\Delta$ as an image and perform 2D convolutions as in [28]. We do not downsample the input, so $X_\Delta$ and $Z$ have identical temporal resolution. Therefore, the network captures spatio-temporal correlations with a large receptive field in the latent space, which can represent motion of overlapped body parts. The decoder $D_s$ has a symmetric architecture with deconvolution layers. More details can be found in the Supp. Mat.

We train our autoencoder on the AMASS dataset with the following loss:

$$L_s(F_s, D_s) = |X_\Delta - X_\Delta^{rec}| + \alpha_s \frac{1}{S(T-2)} \sum_{t=1}^{T-2} |z_{t+1} - z_t|^2,$$

where the first term is the reconstruction loss minimizing discrepancy between $X_\Delta$ and $X_\Delta^{rec} = D_s(F_s(X_\Delta))$, and the second term minimizes the 1st order derivative of the latent sequence $Z = F_s(X_\Delta) = [z_1, z_2, ..., z_{T-1}]$. $\alpha_s$ weights the contribution of the second term.

With a pretrained autoencoder, we design a smoothness loss to regularize motions over time. Specifically, we take the per-frame bodies obtained from Stage 1, and concatenate their markers into a velocity map $X_\Delta^{opt}$. We feed this map into $F_s$, encoding it into $Z^{opt} = F_s(X_\Delta^{opt}) = [z_1^{opt}, z_2^{opt}, ..., z_{T-1}^{opt}]$. The smoothness loss is given by

$$E_{smooth}(\gamma, \theta, \phi) = \frac{1}{S(T-2)} \sum_{t=1}^{T-2} |z_{t+1}^{opt} - z_t^{opt}|^2. \quad (3)$$

Compared to the methods working in joint space locally, our prior can better capture longer-range correlations between the motion of different body parts, hence encoding full-body dynamics.

**Contact friction modeling.** The contact term used in Eq. 1 only considers body-scene proximity, and hence cannot prevent skating artifacts (e.g. a person slides when sitting on a chair). To overcome this issue, we design a contact term that incorporates stationary frictions. Compared to methods which work with foot joints [47, 52], our contact friction term is based on the body and scene mesh, with a more generic human-scene interaction setting, and considers also other body parts such as gluteus.

Specifically, we pre-define a set of "contact friction" vertices $V_c \subset V_n$, corresponding to 194 foot and 113 gluteus vertices. When contact occurs (i.e. the distance between a body vertex to the scene mesh is smaller than 0.01m), the velocity $v_t$ of the contacted vertex in $V_c$ is regularized: the component $v_t$ along the scene normal $n$ should be non-
our convolutional infilling network. Unlike other motion infilling network.

We design a motion infilling network, train it on AMASS, and exploit it as a prior in an optimization algorithm.

Motivation. Figure 3 shows an overview of our convolutional infilling network. Unlike other motion infilling models e.g. [28], our infiller takes body surface markers as input, and infers motions and contact states jointly.

Algorithm 1: smooth motion recovery in Stage 2.

**Result:** Smooth motion w.r.t. SMPL-X body parameters

**Init:** Fitted meshes from Stage 1, scene mesh, smoothness prior $F_s(\cdot)$;

for $i = 1 : N$ do

$Z^{opt} = F_s(X^{opt}_{\Delta})$;

compute $E_{smooth}$ with Eq. (3);

compute $E_{fric}$ with Eq. (5);

minimize $E_{PROX_M} + E_{smooth} + E_{fric}$

end

Figure 3: Illustration of our motion infilling network. The blue and green color denote the marker local coordinate and the foot-ground contact state, respectively. Yellow, gray and orange color represent the root translational velocity $[t_1, t_2]$ and rotational velocity $\gamma$, respectively. Note that masking is only applied to the local motion. The motion infiller takes global motion as input, and predicts local motion.

negative to prevent interpenetration, and the component tangential to the scene, $v^t_{tan}$, should be small to prevent sliding. Formally, this gives us:

$$v_t \cdot n \geq 0, \quad |v^t_{tan}| \leq \sigma \quad \text{for } t \in T_f,$$

where $T_f$ is the set of frames in which vertex and scene are in contact, and $\sigma$ is a small number as a threshold. Based on this, we formulate our contact friction term as:

$$E_{fric}(\gamma, \theta, \phi) = \sum_{t \in T_f} \sum_{v_t \cdot n < 0} |v_t \cdot n| \cdot \sum_{|v^t_{tan}| \geq \sigma} |v^t_{tan}| - \sigma|.$$

Stage 2 fitting. We combine $E_{smooth}$ and $E_{fric}$ with the error terms in Eq. 1, from which we remove $E_{contact}$ and $E_D$ to obtain a modified function $E_{PROX_M}$. We optimize the resulting objective for $N$ iterations as shown in Alg. 1.

3.4. Recovering Motion under Occlusion

Although Alg. 1 produces smooth motions, it cannot recover realistic body motion under occlusion, which frequently happens in person-scene interactions. Therefore, we design a motion infilling network, train it on AMASS, and exploit it as a prior in an optimization algorithm.

Motion infilling network. Figure 3 shows an overview of our convolutional infilling network. Unlike other motion infilling models e.g. [28], our infiller takes body surface markers as input, and infers motions and contact states jointly.

Here we represent markers in a local coordinate system as in [21, 28]: for each frame $t$, marker locations are relative to the position of the body root, which is the pelvis projected to the ground. In addition, the body global configuration is represented by the root’s translational velocity $t \in \mathbb{R}^2$ and rotational velocity $\gamma \in \mathbb{R}$ around the up-axis. Moreover, we select two markers per foot and check whether they contact with the ground at each frame. The marker is deemed in contact with the ground if its velocity is smaller than 20cm/s and its height above the ground is lower than 10cm. Finally, we arrange the motion sequence into a 3D tensor $Y \in \mathbb{R}^{P \times T \times 4}$ with 4 channels. In the first channel $Y_{local}$, each column denotes a vector concatenating local body marker positions and contact labels of this frame, and $P$ is the vector dimension. The last three channels $Y_{root}$ consist of the repeated entries of global trajectory velocities $t_1, t_2$ and $\gamma$ respectively, which allows us to couple global and local motion more closely than [21, 28].

During training, we set a spatio-temporal visibility mask $M \in \{0, 1\}^{P \times T}$ (1 denotes visible, and 0 otherwise) to corrupt the local motion with $Y_{local} = Y_{local} \odot M$, where $\odot$ denotes element-wise matrix multiplication. Since most (upper) body parts are often visible in practice and it is easy to estimate the root’s motion, we assume the global trajectory is given, and only mask local pose. To generate plausible occlusion masks for training on AMASS, we sample masks computed from the PROX dataset [19], where we leverage depth to reason about occlusions. Note that contact labels are masked whenever feet are not visible. With the masked motion $\hat{Y} = \{\hat{Y}_{local}, \hat{Y}_{root}\}$, we train the infiller autoencoder $G$ to reconstruct the full local motion by minimizing:

$$L_{infill}(G) = h(G(\hat{Y}), Y_{local}),$$

where $h(\cdot)$ is the L1 loss for local marker coordinates, and the binary cross entropy (BCE) loss for the contact labels.

Per-instance self-supervised learning. To better leverage visible body parts during testing, we fine-tune the pre-trained motion infiller on each individual test motion sequence to adapt the learned general prior to per-instance. Unlike [24], our fine-tuning procedure is self-supervised. Specifically, given a partially occluded test sequence $\tilde{Y}$, and occluded markers described by mask $M$, we fine-tune the network parameters by exploiting the visible markers in the sequence via minimizing:

$$L_{finetune}(G) = h(G(\tilde{Y}), Y_{local} \odot M).$$

We show in Sec. 4 that this per-instance self-supervised learning effectively increases prediction accuracy for both visible and invisible body parts.

Stage 3 fitting (Alg. 2). Given results from Stage 2, we combine the global configurations and local markers produced by the fine-tuned infiller, and reconstruct marker
Algorithm 2: occluded motion recovery in Stage 3.

<table>
<thead>
<tr>
<th>Step</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fine-tune $G(\cdot)$ according to Eq. (7);</td>
</tr>
<tr>
<td>2</td>
<td>compute $\hat{X}$, $\hat{C}$ from $G(\cdot)$;</td>
</tr>
<tr>
<td>3</td>
<td>the optimization loop;</td>
</tr>
<tr>
<td>for $i = 1 : N$ do</td>
<td></td>
</tr>
<tr>
<td>$Z^{opt} = F(x^{opt})$;</td>
<td></td>
</tr>
<tr>
<td>compute $E_{smooth}$ with Eq. (3);</td>
<td></td>
</tr>
<tr>
<td>compute $E_{fric}$ with Eq. (5);</td>
<td></td>
</tr>
<tr>
<td>compute $E_{infill}$ with Eq. (8);</td>
<td></td>
</tr>
<tr>
<td>minimize $E_{PRoX} + E_{smooth} + E_{fric} + E_{infill}$</td>
<td></td>
</tr>
<tr>
<td>end</td>
<td></td>
</tr>
</tbody>
</table>

for $i = 1 : N$ do

$E_{infill}(\gamma, \theta, \phi) = |\hat{X} - x^{opt}| \circ (1 - M_b)$

$+ \sum_{t=1}^{T} \sum_{k \in K} \hat{c}^t_k \cdot d(v^t_k, a),$  (8)

where $x^{opt}$ is the marker global location from the SMPL-X body to optimize, $M_b$ is the occlusion mask for the body, and $K$ is the set of foot vertices. For a foot vertex $k$, $\hat{c}^t_k = 1$ if its nearest foot marker contact label is 1, 0 otherwise, and $v^t_k$ is the absolute magnitude of velocity at frame $t$. $d(v^t_k, a)$ corresponds to $|v^t_k - a|$ if $v^t_k > a$, 0 otherwise. We set the foot velocity threshold $a$ as 10cm/s.

4. Experiments

4.1. Datasets

AMASS [38]. AMASS collects 15 high-quality mocap datasets, with 11263 motions from 344 subjects. For each sequence, AMASS provides per-frame SMPL-H [49] parameters obtained via MoSh++ (i.e. fitting SMPL-H to mocap markers). We downsample the sequences to 30fps, and trim them to clips of 120 frames for training. Similar to [71], for each clip we reset the world coordinate to the pelvis joint in the first frame. The $x$-axis is the horizontal component of the direction from the left hip to right hip, the $y$-axis points forward, and the $z$-axis points upward. We exclude TCD_handMocap, TotalCapture, SFU, SSM_synced, KIT and EKUT, and use the rest to train our motion smoothness and infilling models. We exclude TCD_handMocap, TotalCapture, SFU from training since we use them to evaluate our motion infilling method. We do not use EKUT, KIT and SSM_synced due to their inconsistent frame rate.

PROX [19]. We use this dataset to test our models and optimization algorithms in Stage 2 and Stage 3. PROX collects monocular RGB-D sequences from 20 subjects moving in and interacting with 12 different indoor scenes. A Kinect-One sensor [1] is employed to capture the sequences at 30fps, and the 3D reconstructions of the static scenes are provided. SMPL-X parameters are fitted to RGB-D data in each frame (see Sec. 3.2) to reconstruct 3D bodies. Following the same pre-processing procedure for AMASS, we trim the sequences, reset the pelvis coordinate, and obtain 205 clips of 100 frames each for evaluation.

3DPW [60]. As with PROX, we use this dataset for evaluation. 3DPW fits SMPL to IMUs and RGB videos, mostly captured in in-the-wild scenario. Although the provided per-frame SMPL fits are accurate, the motion across frames has jitters and temporal discontinuities. As above, we preprocess the motion sequences and obtain 300 clips of 100 frames for evaluation. Since sequences are captured with a moving camera, global SMPL body configurations are not reconstructed accurately. Hence, for 3DPW we test our priors applying them only on local motions. Namely, the body pelvis joints in different frames are aligned, and joint positions are defined with respect to the local coordinate system of each individual frame.

4.2. Evaluation of Motion Smoothness Prior

We compare our motion smoothness prior (denoted by ‘Ours-SP’) against three optimization-based baselines: the DCT-based prior from [22]; minimizing velocity magnitude (L2-V) [4, 33, 57, 73]; minimizing acceleration magnitude (L2-A) [33, 41, 52]. For all methods, we combine them with $E_J$, $E_{prior}$ in Eq. 1 and minimize the resulting objective function to fit SMPL-X to data. Specifically, the objective function of Ours-SP consists of $E_J$, $E_{prior}$ and $E_{smooth}$ in Eq. 3. We evaluate the fits on both PROX and 3DPW.

4.2.1 Metrics

2D joint accuracy. This metric is only used for PROX. We manually annotate 2D body joints on 542 frames via Amazon Mechanical Turk (AMT). The AMT annotations are in the OpenPose [9] coco-25 format (including 25 body joints), and are converted to the SMPL-X body joint format for evaluation. Following [44], the neck, left and right hip joints are excluded from evaluation due to their definition ambiguity. We report the average L2 norm of 2D joint errors (2DJE) between our results and annotations.

3D accuracy. This metric is only used for 3DPW. Following [60], we report the mean per joint position error (MPJPE) and per vertex error (PVE) with aligned body pelvis between our estimated motions and the ones provided by 3DPW. We expect that an effective motion smooth prior can improve motion temporal consistency while preserving the original body configuration quality. Therefore, the lower these two scores are, the better. However, for an exhaustive evaluation, 3D accuracy should be combined with a metric assessing motion smoothness.
Motion smoothness. Ideally, recovered motions should resemble real ones as much as possible. Translating this into a metric, we use the Power Spectrum KL divergence (PSKL) [20] to measure the distribution distance between our results and AMASS motion sequences. Specifically, we evaluate PSKL w.r.t. the acceleration distribution for both body markers and SMPL-X joints on PROX, and for SMPL joints on 3DPW. Since PSKL is not a symmetric measure, we report the numbers for both directions. Smaller values of PSKL indicate better performance (see Supp. Mat. for more details).

Human-scene interpenetration. We assess the degree of human-scene interpenetration on PROX by using the non-collision score adopted in [69, 72]. It measures the ratio between the number of body vertices with non-negative scene SDF values, and the total number of body vertices, i.e., the ratio of body vertices that do not interpenetrate with the scene mesh. We report the average non-collision scores over all frames, and denote it as ‘NonColl’. A higher value indicates fewer human-scene interpenetration.

4.2.2 Results

Tab. 1 and Tab. 2 show the results of motion smoothness evaluation on PROX and 3DPW, respectively. For both datasets, the originally provided motions have the largest PSKL score w.r.t AMASS, indicating that motions are not natural. Compared to all baselines, our method achieves the lowest PSKL scores in both directions, suggesting that it produces more natural motions. On PROX, all methods achieve comparable non-collision scores. Our method achieves a lower 2D pose error compared to the original PROX data and baseline methods. On 3DPW, our method has small MPJPE/PVE while reporting the best PSKL scores. These results demonstrate that our method can generalize well over different datasets for both global motions and local motions.

Overall, our method consistently outperforms other baselines, by significantly improving motion naturalness while preserving per-frame pose accuracy. This is due to the fact that we learn our smoothness prior from the rich and diverse AMASS data, and apply regularization in latent space. In contrast, baseline methods only encourage motion smoothness of disjoint local body parts, and hence have larger gaps to high-quality AMASS motions. Fig. 4 shows examples of latent sequences obtained from PROX sequences. After fitting using our prior, latent sequences become smoother along the time axis and jitters are removed.

4.3. Evaluation of Motion Infilling Prior

We compare our proposed motion infilling prior (denoted by ‘Ours-IP’) with the infiller from Kaufmann et al. [28] on AMASS. On top of these networks, we addition-ly fit SMPL-X parameters to the infilled markers/joints, so as to perform fair comparison. The fitting function is:

\[ E_{amass} = E_{3D} + E_{prior} + E_{smooth} + E_{foot}, \tag{9} \]

where \( E_{3D} \) is the error between infilled marker positions and the corresponding markers on the SMPL-X body to optimize, and \( E_{prior} \) is the prior term for body and hand pose. For our method, \( E_{foot} \) is the second term in Eq. 8. For the baseline method, we define \( E_{foot} \) using a heuristic: foot-ground contact happens if the foot marker distance from the ground is smaller than 10cm (see Supp. Mat.).
We randomly select 130 sequences from our AMASS test set, in order to remove redundant motions and reduce computational cost. To simulate the occlusions occurring in real person-scene interactions and absent in AMASS, at evaluation time, for the network input of both methods, we mask out all markers belonging the lower part of the body and the contact labels in all frames. Furthermore, we evaluate the proposed motion infilling prior on PROX in terms of 2D joint accuracy, PSKL and non-collision score.

### 4.3.1 Metrics

**3D accuracy.** We report the mean position error for joints (MPJPE), body markers (MPMPE) and body vertices (PVE) in global coordinates between the infilled motions and the motions from AMASS. We compute these three metrics for the full body, and also compute MPJPE and MPMPE for the masked body parts.

**Foot skating.** Following [71], we adopt the “foot skating ratio” as another measure of motion naturalness. We compute it by considering the two markers located on the left and right foot heels. We define skating as happening when the velocity of both foot markers exceeds 10cm/s and their height above the ground is lower than 10cm.

### 4.3.2 Results

The results on AMASS are shown in Tab. 3. Our infiller consistently outperforms the baseline for all metrics. In particular, our model reconstructs more accurate motions with all the three representations (body marker, joint and vertex). Besides, we obtain smaller reconstruction errors for the lower part of the body (MPJPE-L and MPMPE-L).

Compared to the heuristic foot-ground contact rule used for the baseline, our predicted contact labels alleviate foot skating more effectively, and recover foot dynamics during optimization. This is also verified in the ablation study (Ours-IP heuristic contact), where we replace the predicted contact labels by the same heuristic contact rule used in the baseline. A probable reason is that our model learns foot-ground contact and whole body motion jointly, and hence can predict the two more consistently.

In addition, the ablation study suggests that our model performance is consistently improved by the self-supervised fine-tuning (see Sec. 3.4), before SMPL-X fitting (Ours w/o Opt) and after, for both whole body and occluded parts. This indicates that our motion infiller effectively adapts itself to test instances, exploiting more useful information from the unmasked body parts of the input.

The last two rows in Tab. 1 show results of the motion infilling prior on PROX. Compared with results of Stage 2 without motion infilling (Ours-S2), Stage 3 (Ours-S3) has an acceleration closer to AMASS, and lower 2D joint errors. Finally, to assess the stages of our pipeline, we swap Stage 2 and Stage 3, and find that the motion infiller works poorly when taking jittered PROX data as input (see Supp. Mat.). Also, the model overfits to the noisy input when performing the self-supervised test fine-tuning.

### 5. Conclusion

In this paper, we propose a novel motion smoothness prior and a contact-aware motion infilling prior learned from high-quality motion capture data, which effectively learn intrinsic full-body dynamics of smooth motions and recover body parts occluded from the camera view. On top of that, we introduce a new multi-stage optimization pipeline which incorporates the motion priors and a physics-inspired contact friction term, and reconstructs smooth, accurate and occlusion-robust global motions with physically plausible human-scene interactions in complex 3D environments. Nevertheless, there are limitations in the current approach. For instance, human movement is rooted in physics. The current pipeline only incorporates intuitive physics terms (e.g. contact, interpenetration and friction); it is a very promising and challenging research direction to employ more physics inspired motion modeling, in combination with the powerful data-driven motion priors.

**Acknowledgements.** This work was supported by the Microsoft Mixed Reality & AI Zurich Lab PhD scholarship. We sincerely thank Shaofei Wang and Jiahao Wang for proofreading.
References


[53] L. Sigal and M. J. Black. Predicting 3D people from 2D pictures. In *Proc. IV Conf. on Articulated Motion and Deformable Objects (AMDO)*, volume LNCS 4069, pages 185–195, July 2006. 2


[71] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3D bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 2, 4, 6, 8, 1
