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

Siwei Zhang[1] Yan Zhang[1] Federica Bogo[2] Marc Pollefeys[1,2] Siyu Tang[1]
[1]ETH Zürich [2]Microsoft
{siwei.zhang, yan.zhang, marc.pollefeys, siyu.tang}@inf.ethz.ch febogo@microsoft.com

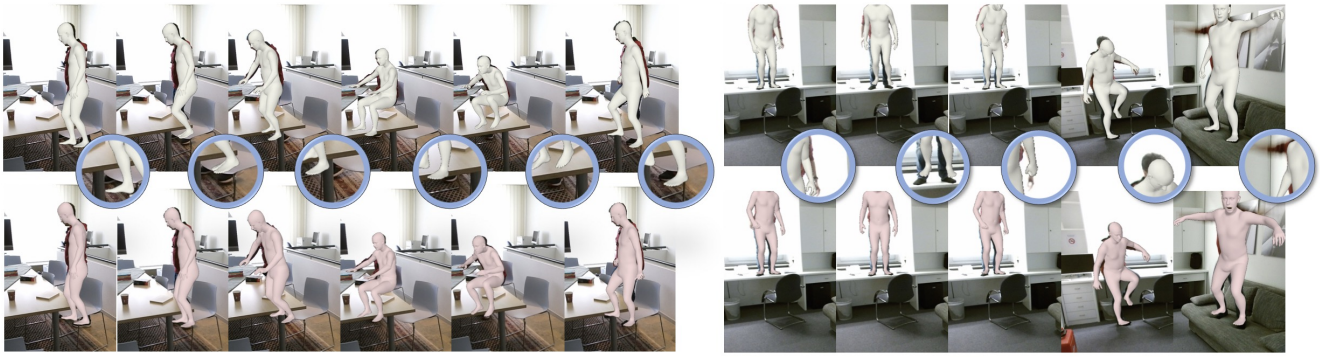

Figure 1: By leveraging data-driven motion priors learned from the large-scale mocap dataset AMASS [38], we reconstruct high-quality human motions in complex 3D scenes from monocular RGB(D) input. Our proposed method (second row) robustly deals with occlusions and achieves more accurate motion reconstructions compared with PROX [19] (first row).

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 work[1], 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

[1]please 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 (**LE**arning human **MO**tion 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 physic 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. Along 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 provides 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 $\mathcal{M}_b(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi})$, whose output is a triangle mesh with vertices $V_b \in \mathbb{R}^{10475\times 3}$. SMPL-X parameters are global translation $\boldsymbol{\gamma} \in \mathbb{R}^3$, body shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, body and hand pose $\boldsymbol{\theta}$, and facial expression $\boldsymbol{\phi} \in \mathbb{R}^{10}$. We denote by $J(\boldsymbol{\beta})$ the 3D body skeleton joints in the neutral pose, and by $R_{\boldsymbol{\theta}\boldsymbol{\gamma}}(J(\boldsymbol{\beta})_i)$ the i-th joint posed according to pose $\boldsymbol{\theta}$ and translation $\boldsymbol{\gamma}$.

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:

$$E_{PROX}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}) = E_J + \lambda_D E_D + E_{prior} + \lambda_{contact} E_{contact} + \lambda_{coll} E_{coll}. \quad (1)$$

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_{prior} combines a set of priors regularizing body pose, shape and facial expression [44]. $E_{contact}$ encourages contact between scene vertices and a pre-defined set of body "contact" vertices. E_{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-

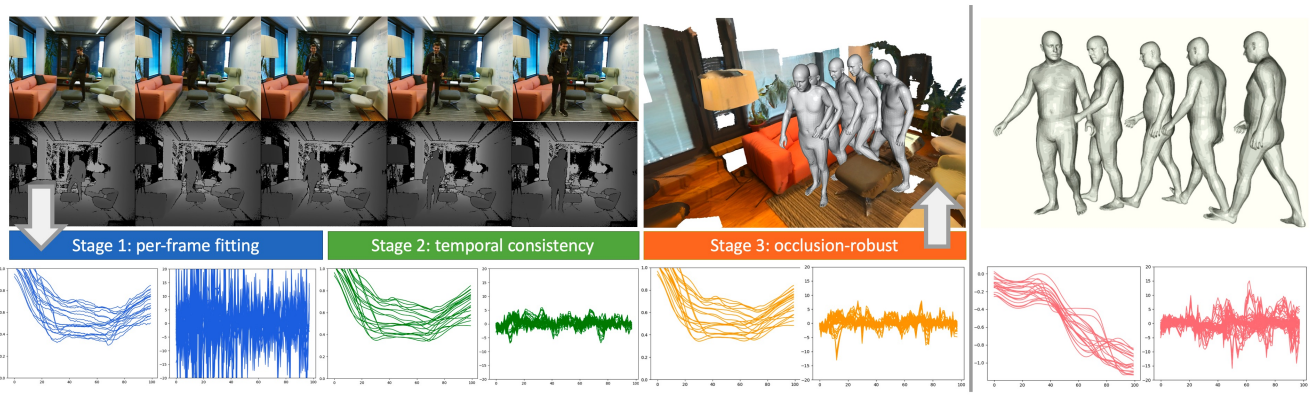

Figure 2: Illustration of our multi-stage pipeline. Provided a scene mesh and an RGBD sequence with body occlusion, our method recovers a realistic global motion, with natural person-scene interactions. The markers trajectories (left) and accelerations (right) of each stage are shown at the bottom, as well as a walking sequence from AMASS [38] (pink). Note that the results from Stage 1 show large and unrealistic motion accelerations (blue). The recovered motion (green) in Stage 2 is significantly smoother. However, it also loses the realistic accelerations (peaks in the acceleration plot) that can happen when the body interacts with the scene (e.g. foot-to-ground contact during walking). Our recovered motion from Stage 3 (orange) is similar to the high-quality AMASS motion w.r.t. both the trajectory smoothness and the acceleration patterns.

tion smoothness prior and a physics-inspired friction term, which are then used in an optimization algorithm.

Motion smoothness prior. Instead of enforcing smoothness explicitly on body joints as in [4, 41, 52], we propose to learn a latent space of smooth motion. To this end, we train an autoencoder with high-quality data from AMASS [38].

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_Δ to its latent representation Z. Here we regard the time series X_Δ as an image and perform 2D convolutions as in [28]. We do not downsample the input, so X_Δ 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:

$$\mathcal{L}_s(F_s, D_s) = |X_\Delta - X_\Delta^{rec}| + \alpha_s \frac{1}{S(T-2)} \sum_{t=1}^{T-2} |\boldsymbol{z}_{t+1} - \boldsymbol{z}_t|^2, \quad (2)$$

where the first term is the reconstruction loss minimizing discrepancy between X_Δ and $X_\Delta^{rec} = D_s(F_s(X_\Delta))$, and the second term minimizes the 1st order derivative of the latent sequence $Z = F_s(X_\Delta) = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_{T-1}]$. α_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_Δ^{opt}. We feed this map into F_s, encoding it into $Z^{opt} = F_s(X_\Delta^{opt}) = [\boldsymbol{z}_1^{opt}, \boldsymbol{z}_2^{opt}, ..., \boldsymbol{z}_{T-1}^{opt}]$. The smoothness loss is given by

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

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 modelling. 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 $V_c \subset V_b$, 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 $\boldsymbol{v}_t$ of the contacted vertex in V_c is regularized: the component $\boldsymbol{v}_t$ along the scene normal $\boldsymbol{n}$ should be non-

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**
  $Z^{opt} = F_s(X_\Delta^{opt})$;
  compute E_{smooth} with Eq. (3);
  compute E_{fric} with Eq. (5);
  minimize $E_{PROX_M} + E_{smooth} + E_{fric}$
end

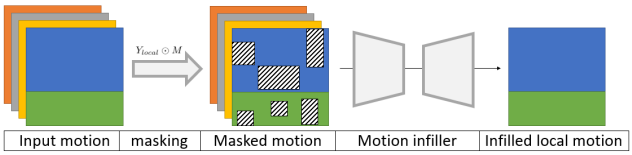

Figure 3: Illustration of our motion infilling network. The blue and green color denote the marker local coordinate and the foot-ground contact state, respectively. Yellow, gray and orange color represent the root translational velocity $[t_1, t_2]$ and rotational velocity γ, respectively. Note that 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:

$$\boldsymbol{v}_t \cdot \boldsymbol{n} \geq 0, \quad |\boldsymbol{v}_t^{tan}| \leq \sigma \quad \text{for} \quad t \in T_f, \tag{4}$$

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

$$E_{fric}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{t \in T_f} \left(\sum_{\boldsymbol{v}_t \cdot \boldsymbol{n} < 0} |\boldsymbol{v}_t \cdot \boldsymbol{n}| + \sum_{|\boldsymbol{v}_t^{tan}| \geq \sigma} ||\boldsymbol{v}_t^{tan}| - \sigma| \right). \tag{5}$$

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 $\boldsymbol{t} \in \mathbb{R}^2$ 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 γ 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 $\tilde{Y}_{local} = Y_{local} \odot M$, 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 $\tilde{Y} = [\tilde{Y}_{local}, Y_{root}]$, we train the infiller autoencoder G to reconstruct the full local motion by minimizing:

$$\mathcal{L}_{infill}(G) = h(G(\tilde{Y}), Y_{local}), \tag{6}$$

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 Y, and occluded markers described by mask M, we fine-tune the network parameters by exploiting the visible markers in the sequence via minimizing

$$\mathcal{L}_{finetune}(G) = h(G(\tilde{Y}), Y_{local}) \odot M. \tag{7}$$

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.

Result: Infilled motion in the presence of occlusions
Init: Results from Stage 2, scene mesh, smoothness prior $F_s(\cdot)$, motion infiller $G(\cdot)$;
Step 1: fine-tune $G(\cdot)$ according to Eq. (7);
Step 2: compute $\hat{X}$, $\hat{C}$ from $G(\cdot)$;
Step 3: the optimization loop;
for $i = 1 : N$ **do**
  $Z^{opt} = F_s(X_{\Delta}^{opt})$;
  compute E_{smooth} with Eq. (3);
  compute E_{fric} with Eq. (5);
  compute E_{infill} with Eq. (8);
  minimize $E_{PROX_M} + E_{smooth} + E_{fric} + E_{infill}$
end

global locations $\hat{X}$ and foot contact labels $\hat{C}$. We define an error term as

$$E_{infill}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}) = |\hat{X} - X^{opt}| \odot (1 - M_b) + \sum_{t=1}^{T} \sum_{k \in K} \hat{c}_k^t \cdot d(v_k^t, a), \quad (8)$$

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 foot vertex k, $\hat{c}_k^t = 1$ if its nearest foot marker contact label is 1, 0 otherwise, and v_k^t is the absolute magnitude of velocity at frame t. $d(v_k^t, a)$ corresponds to $|v_k^t - a|$ if $v_k^t \geq a$, 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 measured 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 additionally fit SMPL-X parameters to the infilled markers/joints, so as to perform fair comparison. The fitting function is:

$$E_{amass} = E_{3D} + E_{prior} + E_{smooth} + E_{foot}, \quad (9)$$

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.).

Table 1: **Evaluation of motion smoothness and infilling priors on PROX.** PSKL-M and PSKL-J denote PSKL computed on markers and joints, respectively. (P, A) denotes PSKL(PROX, AMASS), and (A, P) the reverse direction. For each metric, the best result is in boldface.

Methods	2DJE ↓	PSKL-M ↓		PSKL-J ↓		NonColl ↑
		(P,A)	(A,P)	(P,A)	(A,P)	
PROX [19]	20.94	1.439	2.441	1.464	2.491	**0.955**
DCT [22]	20.96	0.847	1.083	0.937	1.169	**0.955**
L2-A	21.68	0.429	0.396	0.481	0.441	**0.955**
L2-V	21.65	0.551	0.525	0.571	0.536	0.954
Ours-SP	**20.64**	**0.249**	**0.256**	**0.272**	**0.275**	0.954
Ours-S2	20.40	0.273	0.255	0.297	0.275	0.977
Ours-S3	**20.23**	**0.236**	**0.234**	**0.256**	**0.255**	**0.979**

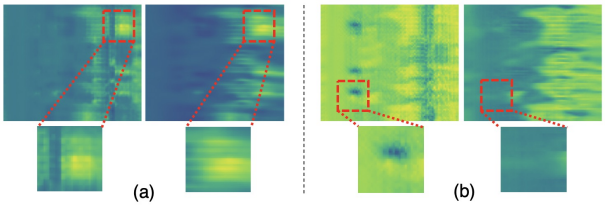

Figure 4: Illustration of two channels ((a) and (b)) of our motion smoothness model latent sequences Z. In each subfigure, the left and right plots show the result before and after Alg. 1, respectively. The row and column in each plot denote feature dimension and time, respectively.

Table 2: **Evaluation of our motion smoothness prior on 3DPW.** PSKL-J denotes PSKL of joints. (3D, A) denotes PSKL(3DPW, AMASS), and (A, 3D) the reverse direction. For each metric, the best result is in boldface.

Methods	MPJPE↓	PVE↓	PSKL-J ↓	
			(3D, A)	(A, 3D)
3DPW [60]	-	-	0.348	0.376
DCT [22]	**0.005**	**0.007**	0.242	0.273
L2-A	0.006	0.009	0.177	0.204
L2-V	0.019	0.025	0.257	0.271
Ours-SP	**0.005**	0.008	**0.173**	**0.197**

Table 3: **Evaluation for motion infilling prior on AMASS.** MPJPE-L / MPMPE-L denotes MPJPE / MPMPE for the masked lower body part. Finetune denotes per-instance self-supervised learning. For each metric, the best result is in boldface.

	Methods	MPJPE ↓	MPMPE ↓	VPE ↓	MPJPE-L ↓	MPMPE-L ↓	Foot Skating ↓
Ours vs baseline	Kaufmann et al. [28]	0.022	0.026	0.025	0.037	0.036	0.237
	Ours-IP	**0.014**	**0.016**	**0.012**	**0.034**	**0.033**	**0.182**
Ablation study	Ours-IP w/o Opt w/o finetune	-	0.025	-	-	0.040	-
	Ours-IP w/o Opt	-	**0.015**	-	-	**0.036**	-
	Ours-IP w/o finetune heuristic contact	0.020	0.024	0.021	0.040	0.038	0.257
	Ours-IP w/o finetune	0.020	0.023	0.021	0.038	0.036	**0.178**
	Ours-IP heuristic contact	**0.014**	**0.017**	**0.013**	**0.036**	**0.035**	0.265

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 Zürich Lab PhD scholarship. We sincerely thank Shaofei Wang and Jiahao Wang for proofreading.

References

[1] Kinect for xbox one. https://en.wikipedia.org/wiki/Kinect#Kinect_for_Xbox_One_(2013). 6

[2] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):44–58, 2005. 2

[3] Ijaz Akhter, Tomas Simon, Sohaib Khan, Iain Matthews, and Yaser Sheikh. Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, 31(2):1–12, 2012. 3

[4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 3, 4, 6

[5] Alexandru O Bălan and Michael J Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008. 2

[6] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 2300–2308, 2015. 3

[7] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3618–3625, 2013. 2

[8] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281, 2019. 2

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 6

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. *arXiv e-prints*, pages arXiv–2011, 2020. 2, 3

[11] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018. 2

[12] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *European Conference on Computer Vision*, pages 210–227. Springer, 2020. 2

[13] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3810–3818, 2015. 2

[14] Ahmed Elhayek, Carsten Stoll, Kwang In Kim, and Christian Theobalt. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In *Computer Graphics Forum*, volume 34, pages 86–98. Wiley Online Library, 2015. 2

[15] Juergen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. Optimization and filtering for human motion capture. *International journal of computer vision*, 87(1-2):75, 2010. 2

[16] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, volume 3, page 641, 2003. 2

[17] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[18] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Jürgen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 224–231. IEEE, 2009. 3

[19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2282–2292, 2019. 1, 2, 3, 5, 6, 7

[20] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 7, 1

[21] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 5, 3

[22] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 2, 3, 6, 7

[23] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[24] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 2, 5

[25] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[26] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[27] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition, pages 5614–5623, 2019. 2

[28] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020. 2, 3, 4, 5, 7, 8

[29] Sena Kiciroglu, Helge Rhodin, Sudipta N Sinha, Mathieu Salzmann, and Pascal Fua. Activemocap: Optimized viewpoint selection for active human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2020. 2

[30] Seong Uk Kim, Hanyoung Jang, and Jongmin Kim. Human motion denoising using attention-based bidirectional recurrent neural network. In *SIGGRAPH Asia 2019 Posters*, pages 1–2. 2019. 3

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 2

[32] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8649, 2019. 6

[34] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. *arXiv preprint arXiv:2011.13341*, 2020. 2

[35] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 2

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[37] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 3

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5442–5451, 2019. 1, 2, 4, 6

[39] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020. 2

[40] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[41] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2, 3, 4, 6

[42] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[43] Dirk Ormoneit, Hedvig Sidenbladh, Michael J Black, and Trevor Hastie. Learning and tracking cyclic human motion. In *Advances in Neural Information Processing Systems*, pages 894–900, 2001. 3

[44] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3, 6

[45] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2

[46] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018. 2

[47] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020. 3, 4

[48] Liu Ren, Alton Patrick, Alexei A Efros, Jessica K Hodgins, and James M Rehg. A data-driven approach to quantifying natural human motion. *ACM Transactions on Graphics (TOG)*, 24(3):1090–1097, 2005. 3

[49] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, Nov. 2017. 6

[50] Nitin Saini, Eric Price, Rahul Tallamraju, Raffi Enficiaud, Roman Ludwig, Igor Martinovic, Aamir Ahmad, and Michael J Black. Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 823–832, 2019. 2

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

[52] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d

motion capture in real time. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 3, 4, 6

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

[54] Leonid Sigal, Michael Isard, Horst Haussecker, and Michael J Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International journal of computer vision*, 98(1):15–48, 2012. 2

[55] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5349–5358, 2019. 2, 3

[56] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. 2017. 2

[57] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. *arXiv preprint arXiv:2003.03473*, 2020. 6

[58] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017. 2

[59] Raquel Urtasun, David J Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 238–245. IEEE, 2006. 3

[60] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 2, 6, 7

[61] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Dynamical simulation priors for human motion tracking. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):52–65, 2012. 3

[62] He Wang, Edmond SL Ho, Hubert PH Shum, and Zhanxing Zhu. Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE transactions on visualization and computer graphics*, 27(1):216–227, 2019. 3

[63] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 2

[64] Yangang Wang, Yebin Liu, Xin Tong, Qionghai Dai, and Ping Tan. Outdoor markerless motion capture with sparse handheld video cameras. *IEEE transactions on visualization and computer graphics*, 24(5):1856–1866, 2017. 2

[65] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2020. 2

[66] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019. 2

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

[68] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 3

[69] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *8th international conference on 3D Vision (3DV 2020)(virtual)*, 2020. 3, 7

[70] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1324–1333, 2020. 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

[72] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020. 3, 7

[73] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 2, 6

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

A. Architecture Details

The model architecture for motion priors is illustrated in Fig. S1. The motion smoothness prior and the motion infilling prior share a similar network architecture, The encoder includes 5 consecutive convolution blocks, with each block containing [conv3x3, LeakyReLU, conv3x3, LeakyReLU, MaxPooling] layers. The motion smoothness prior has the feature channel of [32, 64, 64, 64, 64] for the output of each encoder block. The motion infilling prior has the feature channel of [32, 64, 128, 256, 256] for the output of each encoder block. The decoder includes 5 deconvolution blocks, with each block containing [deconv3x3, LeakyReLU, deconv3x3, LeakyReLU] layers. For the motion smoothness prior, since the smooth constraint (Eq. 3) works on the latent space, we do not downsample the features so that the latent space can preserve the full spatial-temporal resolution the same as the input motion, to model smooth full-body dynamics without losing motion details, thus the MaxPooling layer is not included in the motion smoothness prior.

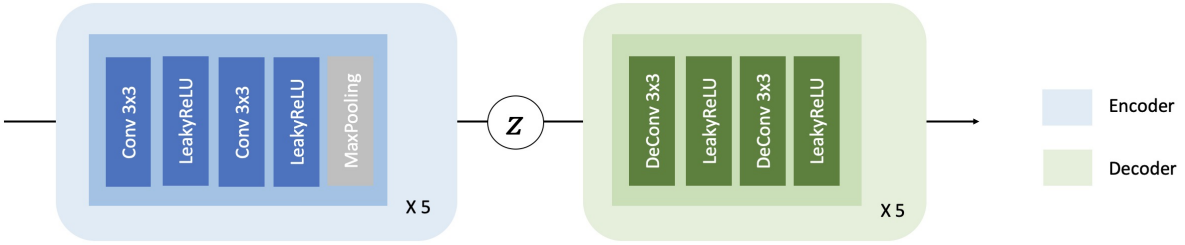

Figure S1: Model architecture for motion priors. The encoder includes 5 consecutive [conv3x3, LeakyReLU, conv3x3, LeakyReLU, MaxPooling] blocks, and the decoder includes 5 consecutive [deconv3x3, LeakyReLU, deconv3x3, LeakyReLU] blocks. The MaxPooling layers are only included in the motion infilling prior network.

B. Experiment Details

B.1. Implementation Details

The proposed algorithm is implemented with PyTorch 1.4.0. We use a single TITAN RTX GPU for training and optimization. For the motion smoothness prior and motion infilling prior training, we use ADAM as the optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with the learning rate 1e-4. The motion smoothness prior is trained for 150 epochs with a batch size of 60, and the motion infilling prior is trained for 900 epochs with a batch size of 120. For the proposed multi-stage optimization pipeline on PROX [19], we use ADAM as the optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with the learning rate 5e-3. In both Stage 2 and Stage 3, we optimize for 900 steps for each motion clip of 100 frames.

B.2. Marker Placement

The body surface marker placement is illustrated in Fig. S2. We choose 67 markers on the SMPL-X body mesh surface, following the SSM2 marker setting in [38]. Furthermore, we select additional 14 markers on the face and fingers to enforce smoothness over hand motions and facial expressions. Note that the additional 14 markers are only utilized in the motion smoothness prior, as we mainly focus on motion infilling for lower part of the body in the motion infilling prior. Compared with body joints, body surface markers can better model degrees-of-freedom (DoFs) and incorporate body shape information [71].

B.3. Evaluation Details

Power spectrum KL divergence (PSKL). We use PSKL to measure the distribution distance between two datasets, as in [20]. Formally, given a motion sequence of T frames, and each frame represented by F features, the power spectrum of each feature sequence s_f is $PS(s_f) = ||FFT(s_f)||^2$. x, y, z accelerations of each frame are used as the features and $F = 3M$, where M denotes the number of body markers or joints. The average power spectrum for feature f over N motion

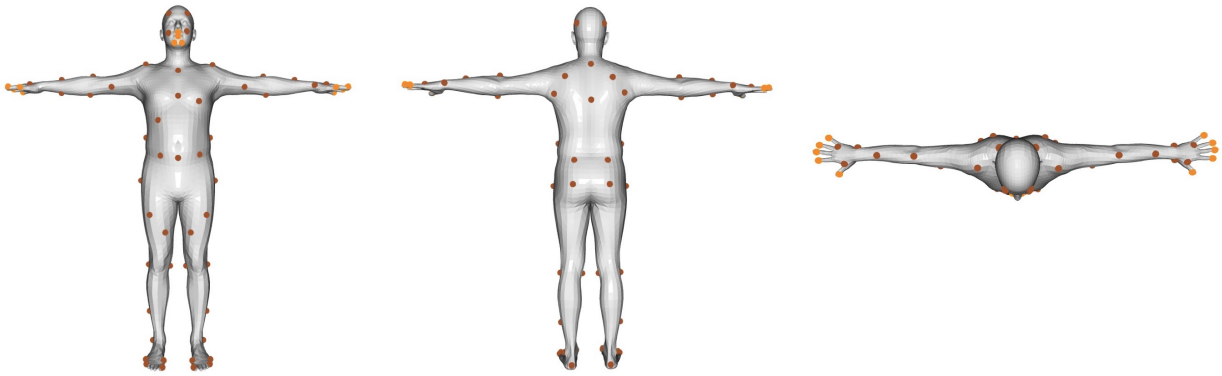

Figure S2: Marker placement for motion priors. The body markers are denoted by spheres over the SMPL-X body surface, following the marker setting (brown) in [38], with 14 additional markers (orange) on the face and fingers. From left to right: the front view, the back view and the top view.

sequences on a dataset C is computed as:

$$PS(C|f) = \frac{1}{N}\sum_{n=1}^{N} PS(s_f), \tag{10}$$

and $PS(C|f)$ is normalized along the frequency dimension. PSKL between datasets C and D is the average power spectrum KL divergence over all feature dimensions:

$$PSKL(C,D) = \frac{1}{F}\sum_{f=1}^{F}\sum_{e=1}^{E} ||PS(C|f)|| * log(\frac{||PS(C|f)||}{||PS(D|f)||}), \tag{11}$$

where e is frequency. KL divergence is asymmetric, thus both directions PSKL(C,D) and PSKL(D,C) are computed.

Motion infilling prior experiments. Here we describe the fitting procedure for the motion infilling prior evaluation on AMASS [38] in detail. For both our proposed method and the baseline method, firstly we implement per-frame fitting with the following objective function:

$$E_{PF} = E_{3D} + E_{prior}, \tag{12}$$

where E_{3D} is the L1 loss between infilled marker positions inferred by the infilling network and marker positions of the SMPL-X body to optimize, and E_{prior} is the prior term for body pose and hand pose. The per-frame fitting aims to provide a good initialization for the temporal fitting. Initialized from per-frame fitting results, the temporal fitting is implemented by minimizing:

$$E_{TF} = E_{3D} + E_{prior} + E_{smooth} + E_{foot}, \tag{13}$$

where E_{smooth} is the proposed smooth prior term in Eq. 3. For our proposed method, E_{foot} is the second term in Eq. 8, which penalizes foot vertex velocity according to the predicted foot-ground contact states. For the baseline method, E_{foot} is defined by a heuristic cue:

$$E_{foot} = \sum_{k,t:z_k^t \leq z_{thres}} d(v_k^t, a), \tag{14}$$

where z_k^t is the distance from the ground of foot joint k at frame t, with z_{thres} set to 10cm. v_k^t is the absolute velocity magnitude of foot joint k at frame t. $d(v_k^t, a)$ corresponds to $|v_k^t - a|$ if $v_k^t \geq a$, 0 otherwise. Foot velocity threshold a is set to 10cm/s. This term penalizes foot joint velocity when its distance to the ground is smaller than 10cm.

Table S1: **Ablation study for swapping the proposed Stage 2 and Stage 3 on PROX.** PSKL-M and PSKL-J denote PSKL computed on markers and joints, respectively. (P, A) denotes PSKL(PROX, AMASS), and (A, P) the reverse direction. For each metric, the better result is in boldface.

	2DJE ↓	PSKL-M ↓		PSKL-J ↓		NonColl ↑
Methods		(P,A)	(A,P)	(P,A)	(A,P)	
Ours-S3	**20.23**	**0.236**	**0.234**	**0.256**	0.255	**0.979**
Ours-S3-S2	20.64	0.273	0.236	0.307	**0.254**	0.972

Table S2: **Comparison with regression-based denoising models on PROX.** 2DJE denotes the 2D joint accuracy. PSKL-J denotes PSKL of joints. (P, A) denotes PSKL(PROX, AMASS), and (A, P) the reverse direction. For each metric, the best result is in boldface.

Methods	2DJE ↓	PSKL-J (P,A) ↓	PSKL-J (A,P) ↓
Wang et al. [62]	140.47	0.294	0.303
Holden et al. [21]	62.97	0.487	0.462
Kim et al. [30]	66.05	0.285	0.278
Ours-SP	**20.64**	**0.272**	**0.275**

Swap Stage 2 & Stage 3. As the motion infilling prior is trained with high-quality data on AMASS, and the proposed self-supervised test fine-tuning relies on the motion of visible body parts, it requires smooth motions as input for good performance. Therefore we first recover temporal consistent motion by the Stage 2, and then include the motion infilling prior in Stage 3. As shown in Tab. S1, if we swap Stage 2 and Stage 3 (denoted by 'Ours-S3-S2'), the overall motion naturalness (PSKL score) will degrade, as well as the pose accuracy (2DJE).

C. Comparison with Regression-based Methods

On PROX, we additionally compare the proposed motion smoothness prior (Ours-SP) with three regression-based denoising methods [21, 30, 62]. These methods directly output smooth motions represented by body joints by taking noisy motion as input. For a fair comparison, we train them on AMASS, adding Gaussian noise to the input motion.

Tab. S2 shows that our smoothness prior (Ours-SP) achieves significantly higher joint accuracy, and produces more realistic motions according to the PSKL scores. The regression-based methods are trained with synthesized noise, which limits their generalizability to different noise distributions, and frequently produces inaccurate global reconstruction, while our motion prior is trained with clean motions, and works very well both on 3DPW captured by IMU sensors and PROX captured by Kinect. Besides, it is unclear how to incorporate the 3D scene constraints directly into the regressors. In contrast, our motion priors and human-scene interaction constraints can be unified in an optimization framework to produce realistic motions that satisfy 3D scene constraints.