# A Structured Latent Space for Human Body Motion Generation

Mathieu Marsot, Stefanie Wuhrer, Jean-Sébastien Franco, Stephane Durocher

# A Structured Latent Space for Human Body Motion Generation

Mathieu Marsot[1] Stefanie Wuhrer[1] Jean-Sébastien Franco[1] Stephane Durocher[2]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France

[2] University of Manitoba 66 Chancellors Cir, Winnipeg MB R3T 2N2, Canada

firstname.lastname@inria.fr, stephane.durocher@umanitoba.ca

## Abstract

*We propose a framework to learn a structured latent space to represent 4D human body motion, where each latent vector encodes a full motion of the whole 3D human shape. On one hand several data-driven skeletal animation models exist proposing motion spaces of temporally dense motion signals, but based on geometrically sparse kinematic representations. On the other hand many methods exist to build shape spaces of dense 3D geometry, but for static frames. We bring together both concepts, proposing a motion space that is dense both temporally and geometrically. Once trained, our model generates a multi-frame sequence of dense 3D meshes based on a single point in a low-dimensional latent space. This latent space is built to be structured, such that similar motions form clusters. It also embeds variations of duration in the latent vector, allowing semantically close sequences that differ only by temporal unfolding to share similar latent vectors. We demonstrate experimentally the structural properties of our latent space, and show it can be used to generate plausible interpolations between different actions. We also apply our model to 4D human motion completion, showing its promising abilities to learn spatio-temporal features of human motion. Code is available at https://github.com/mmarsot/A_structured_latent_space.*

## 1. Introduction

This work investigates learning a structured latent space to represent and generate temporally and spatially dense 4D human body motion, where a single point of a low-dimensional latent space represents a multi-frame sequence of dense 3D meshes. Recently, several works have proposed to learn such motion priors for 4D human body sequences of arbitrary motion by capturing information about pose changes over time [16, 41, 17], in the case of fixed sequence duration. Here, we investigate an orthogonal scenario which models sequences of *varying duration*, by considering motions sufficiently similar to allow temporal alignment.

Learning a generative model of 3D human motion of varying duration with structured latent space is of interest for a wide set of applications in computer vision and graphics, where a lightweight 4D representation translates to gains in information processing. By capturing a spatio-temporal motion prior, the model opens new directions for many completion tasks given temporally, geometrically sparse or incomplete inputs, as it allows to reason within a restricted plausible spatio-temporal solution space.

Learning this space is a difficult task with two major challenges. First, the model needs to capture the intertwined variations of different factors, *e.g.* morphology, global motion, body pose, and temporal evolution of the motion, and do so for motions that differ in duration. In particular, while it is known that morphology impacts the way a motion is performed [35, 39], it remains challenging to take this correlation into account during motion generation. Second, the amount of data that needs to be processed for training is large, as typical acquisition systems for dense human body motions produce $30-50$ frames per second, with each frame containing thousands of geometric primitives.

To address these challenges, we take inspiration from two existing lines of work. The first studies temporally dense skeletal data, with the goal of generating skeletal human motion sequences that capture the temporal evolution of the global motion [37, 33, 21]. These do not address dense surfaces. The second line of work represents realistic 3D human body surfaces in a low-dimensional shape space [3, 26], but do not consider the temporal dimension.

We combine the advantages of both in a data-driven framework that learns a latent motion representation, which allows to simultaneously represent temporal motion information and detailed 3D geometry at every time instant of the motion. The learning uses multi-frame sequences as input and output. Inspired by works on morphable body models [35, 2], we align the training sequences both temporally and spatially, which leads to comparisons at corresponding instances of the motion and anatomically corresponding

*Institute of Engineering Univ. Grenoble Alpes

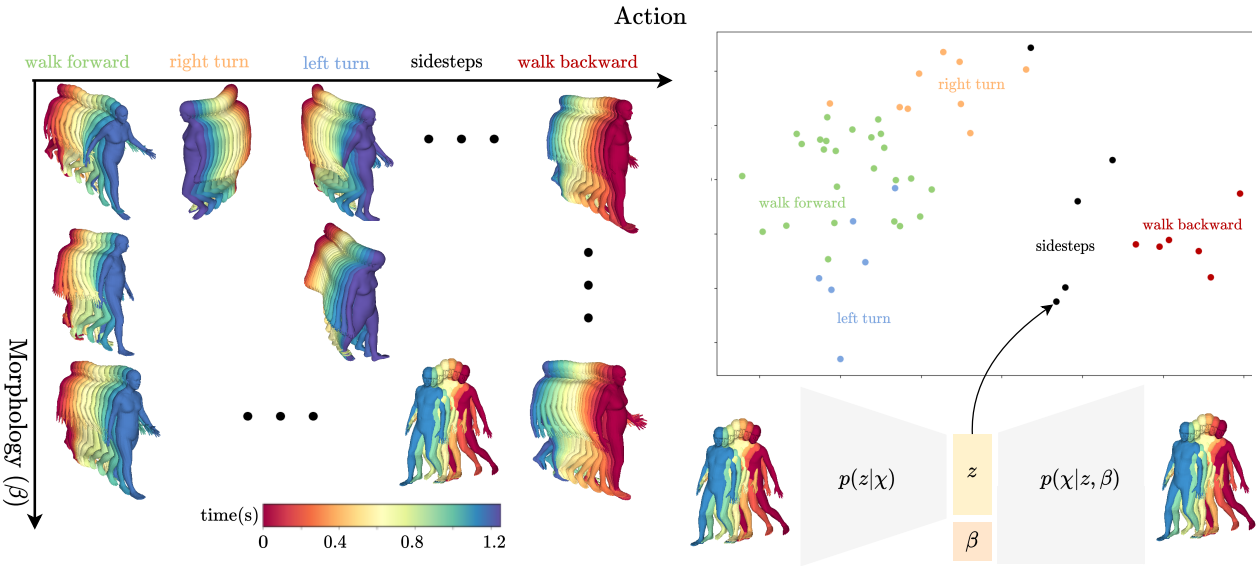


Figure 1: We learn a latent *motion space* from multi-frame 4D sequences. Left: Training sequences consist of different motions performed by different subjects (color-coded as shown in legend). Bottom right: Encoder-decoder architecture learns a latent motion space that encodes motion sequence $\chi$ into latent vector $z$; the decoder conditions $z$ on morphology $\beta$. Top right: Structured latent space. Plot shows subset of 51 motions, manually labelled by action, in 2D projection of latent space. Actions form clusters.

points. In particular, we consider motions whose duration vary significantly while geometrically similar enough to allow for temporal alignment, performed by actors in minimal clothing to allow for effective spatial alignment.

In our experiments, we consider motions during which the hip performs a cycle, as this includes common motions such as walking and running, and generalizes to more complex motions such as dancing or jumping jacks, while imposing no constraints on the arm movements. The resulting latent space is verifiably structured, and allows to generate plausible interpolations between different types of locomotion that outperform linear and per-frame interpolation baselines. As illustrated in Fig. 1, our motion space also learns the interaction between morphology and motion, as generating motions with the same point in latent space conditioned on different representations of morphology leads to motion differences that confirm findings in prior studies conducted on sparse motion data [35].

Our model can serve as prior to complete both spatially and temporally sparse sequences. Given as input unmatched and temporally incoherent point clouds sparsely sampled in space or time, accurate complete 4D reconstructions are obtained. For spatio-temporal completion, our method outperforms state of the art motion priors that encode human motion sequences of fixed duration [16, 41] with sufficient temporal samples, in spite of being trained on significantly less data. It also outperforms a state of the art spatial completion baseline when few samples are available [44].

In summary, we make the following major contributions. First, we present a latent motion space that allows representing and generating multi-frame sequences of dense 3D meshes of varying duration, which accounts for interaction between morphology and motion. Second, we demonstrate that this latent space is structured: similar motions form clusters, and linear interpolation in latent space outperforms baselines. Third, when using our motion space as prior, we outperform state of the art for the application of motion completion from sparsely sampled data in space or time.

## 2. Related Work

The vast literature on generation of human models and motions can be roughly divided into three categories. *Temporally dense* encompasses methods that learn the structure of human motion on a representation that is sparse in 3D space. *Spatially dense* encompasses methods that generate realistic 3D human models without treating long-term motion or dynamic effects. *Full 4D* methods combine long-term motion models with dense 3D shapes per frame.

The first two lines of work have been studied for the past two decades. Studies on temporally dense human motion models proposed different data-driven methods to synthesize motion patterns of skeletal representations or sparse marker positions *e.g.* [30, 36, 37, 10, 17]. These works effectively learn the structure of human motion over durations

of multiple seconds. Studies on spatially dense human models proposed a variety of data-driven methods to synthesize geometrically detailed 3D models *e.g.* [2, 3, 22, 25, 19]. Some models have been extended to learn soft-tissue deformations [26, 18, 31]. Recent works in this area leverage deep learning techniques, and can decouple variations due to different factors *e.g.* [11, 7, 45] or include hands, faces and soft-tissue deformation, *e.g.* [40]. These works generate realistic and geometrically detailed 3D human models.

Over the past few years, a number of works proposed studying 4D human motion data that is densely sampled in space and time. Some work aims to generate dense 3D human motion from sparse MoCap [3, 18, 20, 9] or 2D video data [13, 43]. Given as input marker points or a 2D image per frame of the motion, these works reconstruct dense 4D motion data. Of particular interest for our work is that statistical body models learned on static data have been fitted to MoCap data, providing a large corpus of semi-synthetic dense 4D data [20]. This provides the community access to a large 4D dataset, which we leverage in our work.

The works most related to ours learn spatially and temporally dense 4D motion models of bodies in a data-driven way. The first work to tackle this problem [14] combines two linear models: one capturing dense static 3D shape data and one capturing the motion of MoCap markers. The two linear models are coupled based on semantic parameters including weight and height, which allows generating 4D human motion sequences. Inspired by this idea, our model learns a non-linear model from 4D data, which includes both morphology and motion. We show experimentally that our model generalizes better than a linear one.

With 4D data becoming increasingly available in recent years, a number of studies propose data-driven methods trained on 4D data. First methods including [1, 5, 27, 28] train on either a single motion sequence or multiple sequences showing the same subject performing different motions. A recent work that studies motions of a single subject proposes a deep latent variable model for 4D human motion synthesis [8] to model the probabilistic character of motion.

Recently, 4D motion priors of different subjects performing different motions have received considerable attention. One line of work uses implicitly defined surfaces over time to learn from raw 4D sequences [23, 12], and successfully process human motion data. However, the high dimensionality of the 4D data constrains the sequences to few frames.

To consider longer temporal spans, other works build motion priors from sequences of pose parameters of template aligned meshes. These works include methods that consider a set of labeled actions to learn motion generation based on action labels [24] and methods that model motion as a sequence of transitions between poses [29, 15]. Most similar to our work are methods that build motion priors of unlabeled 4D human motion data [16, 41]. These methods consider motions of a fixed duration and encode them in a motion space, which captures information about pose changes over time. In contrast, we investigate learning a motion space for 4D sequences of varying duration. We demonstrate experimentally that our motion space outperforms [16] and [41] for motion completion.

## 3. Generative model of multi-frame sequences

Two previously identified major challenges need to be tackled in our model: first the very large dimensionality of the problem as is concerns temporally dense sequences of dense 3D meshes; second the modeling of intertwined variations in the generation of 4D sequences, between subject shape, morphology, motion, and temporal unfolding.

To address them, we first need to ensure that we produce a compact and structured motion representation. Our general strategy for this is to extend the static shape space representations (*e.g.* SCAPE) to the spatio-temporal domain, with a similar low-dimensionality characteristic, as detailed in Section 3.1. Second, we articulate our data-driven strategy around an encoder-decoder architecture (Section 3.2). Notably, to explicitly model the interaction between morphology and motion, we choose to condition the motion generation on a representation of morphology. Third, we build our experimental demonstration in a use case that benefits from these choices, focusing our effort on a database of 4D human motion sequences that perform a cyclic motion of the hip joint. This allows to evidence the intended behaviour for this space, which is to group similar locomotions (*e.g.* all walking motions) in clusters. Section 3.3 explains how the model is trained.

### 3.1. Representation of motion sequences

Fig. 2 (top left) shows our representation for 4D sequences. A 4D human motion sequence is parameterized by a single point $z$ in motion space and a identity parameter $\beta$ representing the morphology of the moving person.

**Anchor frames** To represent motion data, we align an unstructured spatio-temporal motion signal. Temporally, we uniformly sample $n$ frames from the motion signal, which we call anchor frames in the following. These anchor frames allow representing motions of various duration with the same number of frames. Spatially, we build on 3D morphable body models to align the frames *e.g.* [22, 25, 19]. These models represent static 3D human body surfaces using a common mesh template. This results in $n$ aligned anchor meshes, making motion comparison practical.

**Representing temporal evolution** The resulting anchor mesh sequence $M = [m_1, \ldots, m_n]$ does not represent the temporal evolution of a motion. The temporal sampling causes an information loss, as it is invariant to similar motions with different temporal unfolding

like walking and running. Therefore, we associate to anchor mesh $m_i$ a timestamp $\tau_i$, and call the timestamp vector $\mathcal{T} = [\tau_1, \ldots, \tau_n]$. The representation $[M, \mathcal{T}]$ is high-dimensional. To simplify processing and disentangle the influence of morphology on motion, we leverage 3D morphable body models that decouple the influence of morphology and pose. By holding morphology constant over $M$, we can represent each $m_i$ using parameter vectors for morphology $\beta$, pose $\theta_i$, and global translation $\gamma_i$. While any decoupled static model can be used, *e.g.* [11, 7, 45], in our implementation we chose the commonly used SMPL model [19] as the AMASS dataset [20] is parameterized by SMPL. We denote the model function by $SMPL$ such that $m_i = SMPL(\theta_i, \gamma_i, \beta)$ and thus $M = [(SMPL(\theta_0, \gamma_0, \beta), \ldots, SMPL(\theta_n, \gamma_n, \beta)]$. By denoting the pose and global translation vectors by $\Theta = [\theta_1, \ldots, \theta_n]$ and $\Gamma = [\gamma_1, \ldots, \gamma_n]$, respectively, $[\Theta, \Gamma, \beta, \mathcal{T}]$ is a low dimensional representation of $[M, \mathcal{T}]$. To retain variation in global displacement (*e.g.*walking backward or forward) and temporal evolution (*e.g.*walking or running), we model $\Gamma$ and $\mathcal{T}$ in the multi-frame sequence representation. $\tau_i$ allow to place freely and on any time span length the anchor meshes, thereby allowing to represent motions with various duration using a constant number of meshes.

**Notation** To emphasize the difference between motion and morphology parameters, we denote $\chi = [\Theta, \Gamma, \mathcal{T}]$ the motion parameters and introduce function $\mathcal{F}$ such that $[M, \mathcal{T}] = \mathcal{F}([\chi, \beta])$. As pre-processing for training, we map a raw motion sequence to the SMPL mesh template using existing solutions [42, 20]. Let $SMPL^{-1}$ denote the mapping function which associates a single raw motion frame to its representation parameters $\theta, \gamma, \beta$.

**Numerical representation** In practice, we represent $\beta$ and $\Gamma$ as in SMPL. Pose features $\Theta$ are joint rotations of a skeleton, represented by a continuous 6D rotation [46] that was shown to outperform other rotation representations when training neural networks.

### 3.2. Architecture

To learn the interaction between morphology and motion patterns, we condition motion generation on $\beta$ using an architecture based on conditional variational auto-encoders (CVAE) [34], as shown in the bottom of Fig. 2. Our architecture encodes motion vector $\chi$ into a low-dimensional latent vector $z$, and $\beta$ is used as condition for the decoder, thereby allowing to capture dependencies between $\chi$ and $\beta$. We assume $z$ and $\beta$ to be independent and learn a disentangled representation. Therefore, the encoder models posterior distribution $p(z|\chi)$, and is not conditioned on $\beta$.

The encoder outputs are interpreted as mean $\mu$ and standard deviation $\sigma$ of the posterior distribution of the latent space. The corresponding latent vector $z$ is sampled as $z = \mu + \epsilon \times \sigma$, with $\epsilon \sim \mathcal{N}(0, 1)$. We denote the probabilistic encoding function by $E : \chi, \epsilon \mapsto z$, and the decoding function as $D : z, \beta \mapsto \hat{\chi}$. The decoder takes $(z, \beta)$ as input, and outputs $\hat{\chi} = [\hat{\Theta}, \hat{\Gamma}, \hat{\mathcal{T}}]$ which are converted back to a sequence of timestamped anchor meshes $[\hat{M}, \hat{\mathcal{T}}] = \mathcal{F}(\hat{\chi}, \beta)$. To go from a reconstructed sequence $\hat{M}$ to a temporally continuous motion, we assume constant motion between anchor meshes.

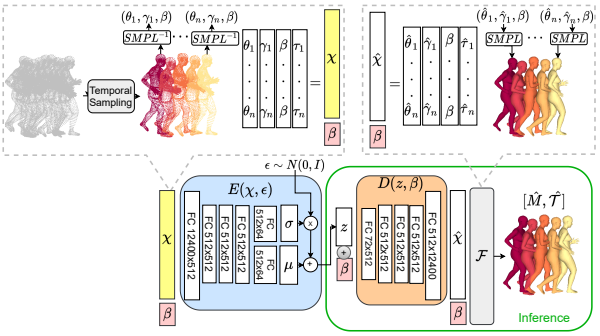


Figure 2: Overview of motion representation and architecture. Top: representation. Left: pre-processing during training samples $n$ anchor frames and extracts per-frame representations of pose $\theta$, translation $\gamma$ and morphology $\beta$ with their timestamp $\tau$ to obtain motion representation $\chi$ and morphology $\beta$. Right: illustration of the function $\mathcal{F}$. Bottom: our architecture consists of a probabilistic encoder $E$ and a decoder $D$, and learns a mapping from $\chi$ to a single latent vector $z$. At inference time, $D$ conditions $z$ on $\beta$ to generate sequence features $\hat{\chi}$ (green box).

### 3.3. Training

The network is trained with a reconstruction term to minimize the difference between the input and output vectors, and a regularization term to constrain the latent variables to follow a known prior distribution. The training is divided into two phases. First, we consider a reconstruction loss on $\chi$ to allow for fast and memory efficient initialization. Second, we replace it by a loss computed directly on the sequence of anchor meshes $M$ in $\mathbb{R}^3$.

**Reconstruction loss on** $\chi$ The standard reconstruction term would be $(\hat{\chi} - \chi)^2$. To balance the influence of the different types of information captured by $\chi$, we divide this loss into three terms operating on pose $\mathcal{L}_{pose} = (\Theta - \hat{\Theta})^2$ translation $\mathcal{L}_{trans} = (\Gamma - \hat{\Gamma})^2$, and time $\mathcal{L}_{time} = (\mathcal{T} - \hat{\mathcal{T}})^2$. This gives a total reconstruction loss

$$\mathcal{L}_{rec} = \omega_{pose}\mathcal{L}_{pose} + \omega_{trans}\mathcal{L}_{trans} + \omega_{time}\mathcal{L}_{time}, \quad (1)$$

where $\omega_{pose}, \omega_{trans}$ and $\omega_{time}$ are the respective weights of the partial reconstruction losses. To minimize $\mathcal{L}_{rec}$, we use adaptive weights to trade off the relative influence of $\mathcal{L}_{pose}, \mathcal{L}_{trans}$ and $\mathcal{L}_{time}$ [6], which do not have the same order of magnitude. Adaptive weights are initialized at 1.0

and updated automatically during training, which ensures that the partial losses are decreasing in similar proportions.

**Reconstruction loss in 4D** The second reconstruction loss is $\mathcal{L}_{spatial} = (M - \hat{M})^2$, where $M$ denotes the 3D coordinate vector or the anchor mesh sequence, resulting in the $4D$ reconstruction term

$$\mathcal{L}_{rec4D} = \omega_{spatial}\mathcal{L}_{spatial} + \omega_{time}\mathcal{L}_{time}, \tag{2}$$

where $\omega_{spatial}$ is an adaptive weight.

**Regularization loss** The regularization term is the squared Kullback-Leibler (KL) divergence between the learned posterior distribution $\mathcal{N}(\mu, \sigma)$ of the latent variable $z$ and a normal prior distribution $\mathcal{N}(0, 1)$, denoted $\mathcal{L}_{KL}$.

**Optimization** A common problem when training VAEs is the weighting of the regularization loss versus the reconstruction loss. We use a fixed weight $\omega_{KL} = 0.01$ to trade off these losses. The training optimizes first

$$\mathcal{L}_{init} = \mathcal{L}_{rec} + \omega_{KL}\mathcal{L}_{KL} \tag{3}$$

and subsequently

$$\mathcal{L} = \mathcal{L}_{rec4D} + \omega_{KL}\mathcal{L}_{KL}. \tag{4}$$

# 4. Evaluation

This section presents comparisons to baselines. We investigate the structure of the learned latent space by visualizing labeled motion sequences in latent space and by linearly interpolating between pairs of input motion sequence. Finally, we demonstrate that the proposed model learns information on the interaction of morphology and motion by visualizing the motion changes caused by changing $\beta$ for a fixed point $z$. Implementation details, a study of the influence of the latent space dimension and regularisation, and video visualizations are in supplementary material.

## 4.1. Data

We automatically extract motion sequences during which the hip performs a cycle from a dataset by comparing all subsequences to a set of 4D template motions using dynamic time warping [4] as distance. Subsequences are considered if this distance is below a threshold. As post-processing, we prune segments with a duration above $3s$ or below $0.3s$. We manually generate two 4D template motions as gait cycles starting with the left and right foot.

We experiment with AMASS [20] and Kinovis [42] datasets. AMASS regroups a set of MoCap recordings and fits SMPL to all data. When splitting AMASS into training and test sets, we treat all sequences of the same MoCap dataset as one entity. For training, our cropping results in 12085 sequences corresponding to $\approx 4.5h$ of motion. We call the extracted test set *AMASS test set*. Kinovis dataset contains 4D motion sequences from a multi-view platform

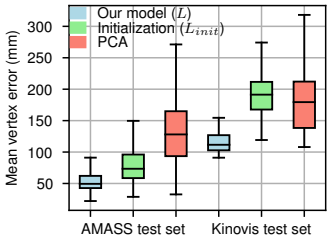


Figure 3: Comparison to baselines w.r.t. reconstruction error. Our model (blue) outperforms a linear PCA baseline (red) and a baseline that considers spatial sampling at skeleton level (green). Boxes follow [38].

and allows to evaluate the generalization of our model to densely captured 4D data. We consider all walking and running sequences, pre-process the data by fitting SMPL before extracting cyclic hip motions, and call this dataset *Kinovis test set*. Details are in supplementary material.

## 4.2. Comparison to baseline models

We compare our model to two baselines w.r.t. the reconstruction error $\frac{1}{nk}(M - \hat{M})^2$, with $\left[\hat{M}, \hat{\mathcal{T}}\right] = \mathcal{F}(\hat{\chi}, \beta) = \mathcal{F}(D(E(\chi, \epsilon), \beta), \beta)$, where $n$ is the number of anchor frames and $k$ the number of vertices per frame. The first baseline applies a linear principal component analysis (PCA) to our representation $[\chi, \beta]$, thereby evaluating the value of using a non-linear model. PCA has access to morphology information when projecting the motion representation to latent space, and reconstructs both $\hat{\chi}$ and $\hat{\beta}$. To provide a fair comparison, we consider the original $\beta$ instead of $\hat{\beta}$ in PCA reconstructions and set the PCA latent dimension to $dim(z) + dim(\beta)$ with $dim(z) = 64$ and $dim(\beta) = 8$. The second baseline considers our model after optimizing $\mathcal{L}_{init}$ only, which operates on skeleton representations, thereby evaluating the value of learning from data that is densely sampled in space.

Fig. 3 shows reconstruction errors for the different models. While PCA provides low reconstruction errors, these are further improved using our model. Our model also improves over its initialization, which shows that considering densely sampled data significantly impacts performance.

## 4.3. Motion space structure and interpolation

Fig. 1 illustrates that our model learns a latent space in which sequences of similar actions are clustered. For the purpose of visualization, we labeled 51 motions by actions and assigned a unique color per action. These motions are then encoded into latent space, which is linearly reduced to two dimensions. Points of the same action form clusters.

This structured latent space can be exploited to generate plausible interpolations between input motions using linear interpolation. Given start and target motion sequences as

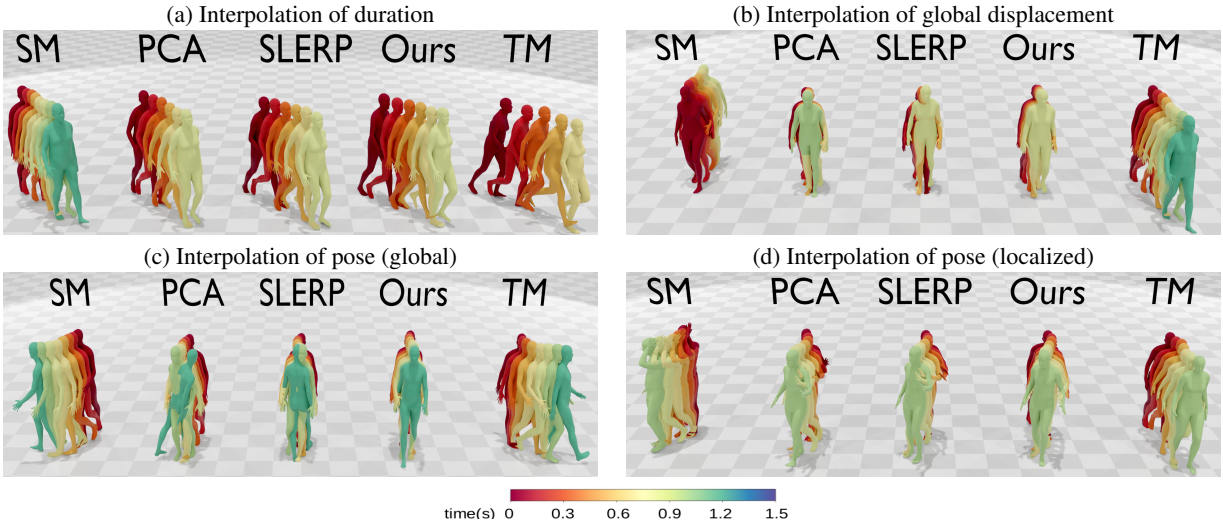


Figure 4: Linear interpolations in latent motion space. Each figure left to right : starting motion, PCA interpolation, SLERP interpolation, our interpolation, and target motion. Sequence models are rendered with a color-coded frame time. **(a)** Running & walking. **(b)** Walking backward & forward. **(c)** Left & right turn. **(d)** Walk & walk carrying an object on the head. All interpolations with our model are plausible, while baselines fail in (b) and (c).

input, we encode them as $(z_s, \beta_s)$ and $(z_t, \beta_t)$, and generate interpolating motion sequences by decoding $((1-k)z_s + kz_t, (1-k)\beta_s + k\beta_t)$ at intermediate position $k \in [0, 1]$.

We compare our results to two baselines. The first uses the PCA model from the previous section and linearly interpolates in PCA space. This comparison, called PCA, evaluates the value of using a non-linear model. The second baseline operates per anchor frame and interpolates linearly between the global displacements, time stamps and morphology parameters, and with spherical linear interpolation [32] (SLERP) between skeletal poses. This comparison, called SLERP, evaluates the value of learning a motion model instead of operating independently per-frame. For all interpolations, visualizations show $k = 0.5$. In the following, we interpolate between sequences that differ in each of the factors encoded in $\chi$.

**Interpolating sequences of different duration** To inspect temporal information learned by our model, we interpolate between a running and a walking motion. For our model, the duration of the intermediate sequences monotonically decreases when going from running to walking, and the intermediate sequences are realistic as shown in Fig. 4(a), showing that our motion space has captured information on the temporal evolution $\tau$. PCA and SLERP baselines also lead to plausible interpolations.

**Interpolating sequences of different global displacement** To inspect global displacement, we interpolate between a forward and a backward walk. Our intermediate sequence corresponds to a really small step, shown in Fig. 4(b). There were no steps this small in the training set. PCA and SLERP baselines fail to interpolate global translation realistically, resulting in foot skating.

**Interpolating sequences of different pose** To inspect the learned information of pose, we consider global and articulated pose separately. First, we interpolate between sequences of turning left and turning right while walking, exhibiting mostly global pose change. The intermediate sequences using our model gradually change from a left to a right turn as shown in Fig. 4(c). PCA and SLERP baselines fail due to the ambiguity when interpolating between opposite rotations, while our model leverages spatio-temporal information to alleviate this ambiguity. Second, we interpolate between walking and walking while carrying an object on the head, exhibiting mostly articulated pose change. The intermediate sequence with our model results in realistic intermediate positions for the arms, gradually elevating them to head level as shown in Fig. 4(d). Both baselines lead to plausible interpolations.

In summary, while our model generates visually plausible interpolations for all parameters encoded in $\chi$, both baselines exhibit failure cases in some scenarios, which shows the value of learning a non-linear 4D motion model.

### 4.4. Interaction between morphology and motion

To examine the influence of morphology $\beta$ on 4D motion $\chi$, we consider a fixed jogging motion represented by $z^*$ in motion space and visualize $\chi$ when setting $\beta$ to $\pm 3$ standard deviations along the first and second principal components. To understand the subtle motion differences, we further visualize the spatio-temporal gradient $\frac{\partial D(z^*, \beta)}{\partial \beta}$ at

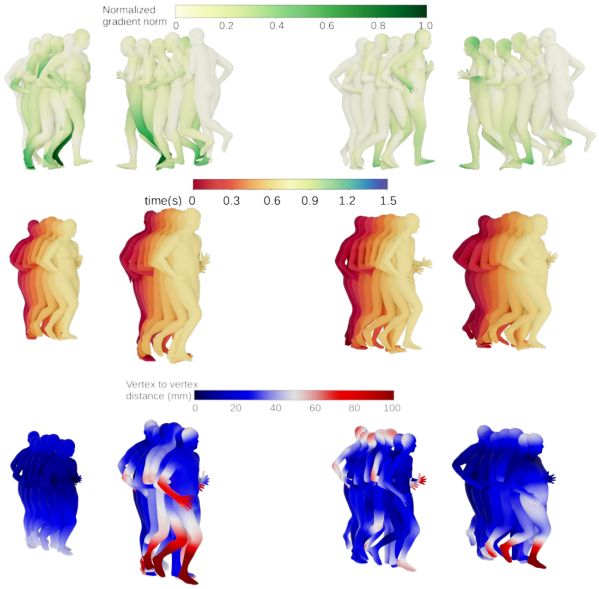


Figure 5: Interaction between morphology and motion on $1^{st}$ (left) and $2^{nd}$ (right) principal components of $\beta$. Top: visualization of our decoder's normalized gradient w.r.t. $\beta$. Middle: our inferences with fixed latent motion vector and $\beta$ taken at $\pm 3$ std. deviations. Bottom: baseline per-frame motion transfer using SMPL for same fixed motion and $\beta$ taken at $\pm 3$ std. deviations, color coded by per-vertex distance to our result. Our learnt correlation has significant impact on motion, which differs up to $10cm$ from baseline.

$\beta = 0$, *i.e.* we look at the gradient learned by the decoder w.r.t. morphology at the mean shape.

We compare our result to a baseline that uses the initial pose parameters and $\beta$ to reconstruct a dense 3D body model using SMPL per frame. This evaluates the influence of learning the interaction between morphology and motion.

Fig. 5 shows the impact of the first (left) and second (right) principal components of $\beta$. The top row shows a color coding of the gradient learned by our decoder w.r.t. $\beta$ on the 4D sequence, and the middle row shows the corresponding 4D motions obtained by our model. The bottom row shows the result of the baseline color-coded by the distance to the result of our model. Changing the first principal component impacts perceived gender. For our model, this changes the 4D motion on the right shoulder and left hip, in agreement with prior studies showing that shoulder sway and hip motion are statistically gender related [35]. Changing the second principal component leads to perceived weight change. For our model, this impacts the 4D motion at the right arm, head and neck. The spatio-temporal areas affected by our motion model are the ones where the baseline leads to significantly different results with up to $10cm$ distance. This shows that our model learns meaningful interactions between morphology and motion.

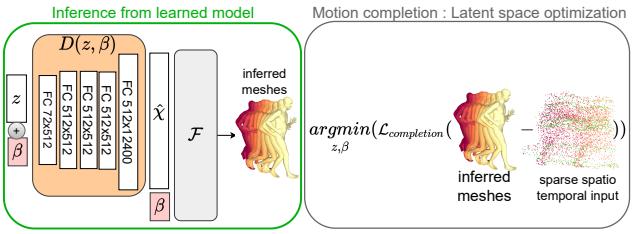


Figure 6: Motion completion. We minimize a loss w.r.t. latent representation $(z, \beta)$. Left: inference pipeline. Right: we optimize $\mathcal{L}_{completion}$ between a sparse 4D point cloud and inferred meshes.

## 5. Application to motion completion from spatio-temporally sparse input

This section applies our model to spatio-temporal completion, which has applications ranging from the registration of a raw spatio-temporally densely scanned 4D sequence over computing realistic in-betweenings for a set of frames sparsely sampled in time to completing full human body motion from a sparse set of MoCap markers.

### 5.1. Completion methodology

We consider as input partial motion sequences of unordered dense 3D scans with possibly additional synchronized MoCap for $k$ landmarks and associated time stamps. Let $S = [s_1, \ldots, s_n]$ denote a sequence of $n$ anchor scans uniformly sampled in time, $L = [l_1, \ldots, l_n]$ the corresponding synchronized sequence of landmarks, and $\mathcal{T} = [\tau_1, \ldots, \tau_n]$ the corresponding time stamps. Some anchor frames are empty, and our input consists of a set $I$ of frame indices $i$ for which $s_i$ or $l_i$ and $\tau_i$ are given.

To compute a sequence of anchor meshes $\hat{M}$ with associated time stamps $\hat{\mathcal{T}}$ that approximate the input, we decode a full sequence of anchor frames $[\hat{M}, \hat{\mathcal{T}}]$ using $\mathcal{F}\left(D(z,\beta),\beta\right)$ and optimize for latent vectors $z^*, \beta^*$ as

$$z^*, \beta^* = \underset{z,\beta}{argmin}(\mathcal{L}_{completion}(\hat{M}(z,\beta), \hat{\mathcal{T}}(z,\beta), S, L, \mathcal{T})), \quad (5)$$

where

$$\begin{aligned} \mathcal{L}_{completion} &= \omega_{dense} \sum_{i \in I} \text{Chamfer}(\hat{m}_i(z,\beta), s_i) \\ &+ \omega_{mocap} \sum_{i \in I} \text{Landmark}(\hat{m}_i(z,\beta), l_i) \\ &+ \omega_{time} \sum_{i \in I} (\hat{\tau}_i(z,\beta) - \tau_i)^2. \end{aligned} \quad (6)$$

The weights $\omega_{dense}$, $\omega_{mocap}$ and $\omega_{time}$ are adaptive [6]. When $s_i = \emptyset, \omega_{dense} = 0$ and when $l_i = \emptyset, \omega_{mocap} = 0$. Varying $\omega_{mocap}$ allows to evaluate the benefit of having tracked input markers. Chamfer is the Chamfer distance between two point clouds and Landmark is the squared Euclidean distance between $k$ vertices of the SMPL template,

Table 1: Comparative evaluation of motion completion. Mean and standard deviation of Chamfer distance in $mm$, computed between completions and ground truth anchor scans from CHUM. N.A. means not applicable.

| | Points per scan $p$ | | | | | Frames ($f$) | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 50 | 100 | 1000 | 10000 | 5 | 20 | 100 |
| Ours (dim(z)=256) | **42±48** | **23±7** | **21 ±9** | **20±10** | 20±10 | 30±14 | **20±10** | **20±10** |
| [44] | N.A. | 58±0.99 | 47 ±0.6 | 21±0.3 | **10±0.46** | N.A. | N.A. | N.A. |
| [41] | 46 ± 52 | 26±8 | 24 ±9 | 22±10 | 22±10 | **22±10** | 22±11 | 22±10 |
| [16] | 216 ± 41 | 88±10 | 69 ±10 | 36±10 | 26±11 | 33±13 | 26±11 | 26±11 |

selected once for all experiments, and the $k$ given landmarks. This optimization is visualized in Fig. 6.

### 5.2. Completion dataset

We introduce a new dataset of cyclic human motion (CHUM), which was captured using a 4D modeling platform with 68 RGB cameras and a Qualisys MoCap system. Data consists of dense scans of approximately 10000 points acquired at 50fps with synchronised MoCap for 16 markers. We recorded 4 actors with different morphologies (2 males and 2 females) performing various cyclic motions like walking, running, side-stepping, skipping, boxing and kicking. For our experiment, we segmented 4 gait cycles manually for each original sequence and found an initial 3D transformation (rotation + translation) to align each segment at $t = 0$. We do not fit SMPL to the dense scans because $\mathcal{L}_{completion}$ does not require correspondence information.

### 5.3. Results

We compare our results to three state of the art approaches. The first performs static 3D completion per frame [44]. Due to its high computational complexity, we apply the static method to a subset of CHUM while other methods are applied to the full dataset. This method is only applicable for spatial completion where observations are available at every frame. The second and third are motion spaces for sequences of fixed duration that can serve as prior [41, 16]. Given a partial motion as input, we optimize a latent motion vector $z$, a morphology $\beta$ and a set of per-frame translation parameters for [41], as global translation is not encoded in this motion space. In case of temporally sparse input, translation parameters are only optimized for frames in $I$ and the remaining are found using linear interpolation between the closest observed frames. For [41] and [16], we optimize for $\mathcal{L}_{completion}$ with $\omega_{time} = 0$, as these motion spaces are designed for sequences of fixed duration and cannot benefit from time stamp information. These methods are applicable for both spatial and temporal completion. [41] uses a latent space of 256 dimensions while [16] uses a total of 36. For fair comparison to the more precise method, we re-train our model with $dim(z) = 256$.

**Spatial completion** We first evaluate the quality of spatial completion by simulating different levels of spatial sparsity by varying the number of points $p$ per scan $s_i$. The sampled points are not in correspondence over time. Table 1 shows the evolution of the reconstruction error in $mm$ when varying $p$. Our method outperforms the static method [44] for very sparse scans ($p < 100$), the two methods are on-par for denser scans ($p = 1000$), and the static method outperforms our method for dense scans ($p = 10000$). This quality on sparse scans is achieved because our model optimizes for all frames simultaneously, so few points per scan suffice to find a plausible solution. The static method deforms a template, and can capture higher levels of geometric detail for dense scans. Our method further outperforms state of the art motion spaces [41, 16], in spite of being trained on significantly less motion data (4.5h for ours vs. 34h for [41, 16]). Qualitative results are shown in supplementary material.

**Temporal completion** Second, we evaluate the quality of temporal completion by varying the number of observed frames. To vary this number for each test sequence, we reduce $I$ to simulate lower frame rates. Table 1 shows the evolution of the reconstruction error. The $f = 100$ frame completion task includes all frames and is given as reference. The model extrapolates with almost no loss of precision with $I_{20} = [5, 10, \ldots, 95, 100]$ (20 frames) and the error is still low with $I_5 = [20, 40, 60, 80, 100]$ (5 frames). While the motion space for sequences of fixed duration [41] is better for sparsely sampled temporal data, we outperform both [41] and [16] for temporally denser data, in spite of using significantly less training data.

## 6. Conclusions and future work

This work presents a latent space that allows to represent and generate multi-frame sequences of human motion in 4D. This latent space contains information on global motion, body pose, temporal evolution of the motion, and morphology. We demonstrated that similar motions tend to form clusters in this latent space and that linear interpolations between pairs of sequences in latent space are plausible. Furthermore, our model to generate 4D motion sequences captures the interaction between morphology and motion. We applied this model to spatio-temporal motion completion, demonstrating state of the art performance. For future work, it would be interesting to explore how to synthesize longer term and more general motion.

## 7. Acknowledgements

We thank Jinlong Yang and Jiabin Chen for the Kinovis test set, Joao Regateiro, Anne-Hélène Olivier and Edmond Boyer for helpful discussions, and Laurence Boissieux, Julien Pansiot, and our volunteer subjects for help with 4D data acquisition. This work was supported by the National French Research Agency under grants 19-CE23-0013-01 (3DMOVE) and ANR-21-ESRE-0030 (CONTINUUM).

## References


[1] 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

[2] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. ACM transactions on graphics (TOG), 22(3):587–594, 2003. 1, 3

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In ACM SIGGRAPH 2005 Papers, pages 408–416. 2005. 1, 3

[4] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In KDD workshop, volume 10, pages 359–370. Seattle, WA, USA:, 1994. 5

[5] Adnane Boukhayma and Edmond Boyer. Surface motion capture animation synthesis. IEEE transactions on visualization and computer graphics, 25(6):2270–2283, 2018. 3

[6] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International conference on machine learning, pages 794–803. PMLR, 2018. 4, 7

[7] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodola. Limp: Learning latent shape representations with metric preservation priors. In European Conference on Computer Vision, pages 19–35. Springer, 2020. 3, 4

[8] Saeed Ghorbani, Calden Wloka, Ali Etemad, Marcus A Brubaker, and Nikolaus F Troje. Probabilistic character motion synthesis using a hierarchical deep latent variable model. In Computer Graphics Forum, volume 39, pages 225–239. Wiley Online Library, 2020. 3

[9] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. ACM Transactions on Graphics (TOG), 40(4):1–16, 2021. 3

[10] 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. 2

[11] Boyi Jiang, Juyong Zhang, Jianfei Cai, and Jianmin Zheng. Disentangled human body embedding based on deep hierarchical neural network. IEEE transactions on visualization and computer graphics, 26(8):2560–2575, 2020. 3, 4

[12] Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, and Yanwei Fu. Learning compositional representation for 4d captures with neural ode. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5340–5350, 2021. 3

[13] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5614–5623, 2019. 3

[14] Alina Kuznetsova, Nikolaus F Troje, and Bodo Rosenhahn. A statistical model for coupled human shape and motion synthesis. In GRAPP/IVAPP, pages 227–236, 2013. 3

[15] Yongjoon Lee, Kevin Wampler, Gilbert Bernstein, Jovan Popović, and Zoran Popović. Motion fields for interactive character locomotion. In ACM SIGGRAPH Asia 2010 papers, pages 1–8. 2010. 3

[16] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In 2021 International Conference on 3D Vision (3DV), pages 771–781. IEEE, 2021. 1, 2, 3, 8

[17] Suhas Lohit, Rushil Anirudh, and Pavan Turaga. Recovering trajectories of unmarked joints in 3d human actions using latent space optimization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2342–2351, 2021. 1, 2

[18] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (ToG), 33(6):1–13, 2014. 3

[19] 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. 3, 4

[20] 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/CVF international conference on computer vision, pages 5442–5451, 2019. 3, 4, 5

[21] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2891–2900, 2017. 1

[22] Alexandros Neophytou and Adrian Hilton. Shape and pose space deformation for subject specific animation. In 2013 International Conference on 3D Vision-3DV 2013, pages 334–341. IEEE, 2013. 3

[23] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5379–5389, 2019. 3

[24] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10985–10995, 2021. 3

[25] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. Pattern Recognition, 67:276–286, 2017. 3

[26] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: A model of dynamic human shape in motion. ACM Transactions on Graphics (TOG), 34(4):1–14, 2015. 1, 3

[27] Joao Regateiro, Adrian Hilton, and Marco Volino. Dynamic surface animation using generative networks. In 2019 International Conference on 3D Vision (3DV), pages 376–385. IEEE, 2019. 3

[28] João Regateiro, Marco Volino, and Adrian Hilton. Deep4d: A compact generative representation for volumetric video. Frontiers in Virtual Reality, 2:739010, 2021. 3

[29] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11488–11499, 2021. 3

[30] Charles Rose, Michael F Cohen, and Bobby Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. IEEE Computer Graphics and Applications, 18(5):32–40, 1998. 2

[31] Igor Santesteban, Elena Garces, Miguel A Otaduy, and Dan Casas. Softsmpl: Data-driven modeling of nonlinear soft-tissue dynamics for parametric humans. In Computer Graphics Forum, volume 39, pages 65–75. Wiley Online Library, 2020. 3

[32] Ken Shoemake. Animating rotation with quaternion curves. In Proceedings of the 12th annual conference on Computer graphics and interactive techniques, pages 245–254, 1985. 6

[33] Leonid Sigal, David J Fleet, Nikolaus F Troje, and Micha Livne. Human attributes from 3d pose tracking. In European conference on computer vision, pages 243–257. Springer, 2010. 1

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 28, 2015. 4

[35] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision, 2(5):2–2, 2002. 1, 2, 7

[36] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision, 2(5):2–2, 2002. 2

[37] Nikolaus F Troje. Retrieving information from human movement patterns. Understanding events: How humans see, represent, and act on events, 1:308–334, 2008. 1, 2

[38] John W Tukey. Box-and-whisker plots. Exploratory data analysis, pages 39–43, 1977. 5

[39] Jungdam Won and Jehee Lee. Learning body shape variation in physics-based characters. ACM Transactions on Graphics (TOG), 38(6):1–12, 2019. 1

[40] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6184–6193, 2020. 3

[41] Jiachen Xu, Min Wang, Jingyu Gong, Wentao Liu, Chen Qian, Yuan Xie, and Lizhuang Ma. Exploring versatile prior for human motion via motion frequency guidance. In 2021 International Conference on 3D Vision (3DV), pages 606–616. IEEE, 2021. 1, 2, 3, 8

[42] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In European Conference on Computer Vision, pages 439–454. Springer, 2016. 4, 5

[43] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7114–7123, 2019. 3

[44] Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, Bugra Tekin, and Edmond Boyer. Reconstructing human body mesh from point clouds by adversarial gp network. In Proceedings of the Asian Conference on Computer Vision, 2020. 2, 8

[45] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In European Conference on Computer Vision, pages 341–357. Springer, 2020. 3, 4

[46] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5745–5753, 2019. 4