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[a], Stefanie Wuhrer[a], Jean-Sébastien Franco[a], Stephane Durocher[b]

[a] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France,* *Institute of Engineering Univ. Grenoble Alpes*
[b] *University of Manitoba,66 Chancellors Cir, Winnipeg MB R3T 2N2,Canada*

Abstract

This work investigates learning a structured latent space to represent and generate temporally and spatially dense 4D human body motion. Once trained, the proposed model generates a multi-frame sequence of dense 3D meshes based on a single point in a low-dimensional latent space. Learning a generative model of human motion with an underlying structured latent space is important for a wide set of applications in computer vision and graphics, including virtual and augmented reality, 3D telepresence, and content generation for entertainment applications. We learn this latent motion representation in a data-driven framework that builds upon two existing lines of works. The first analyzes temporally dense skeletal data to capture the global displacement, poses and temporal evolution of the motion, while the second analyzes static densely captured human scans in 3D to represent realistic 3D human body surfaces in a low-dimensional space. Building upon the respective advantages of these two concepts allows our model to simultaneously represent temporal motion information for sequences of varying duration and detailed 3D geometry at every time instant of the motion. We experimentally demonstrate that the resulting latent space is structured in the sense that similar motions form clusters in this space, and use our latent space to generate plausible interpolations between different actions. We also illustrate the benefits of the approach for 4D human motion completion, showing promising abilities of our model to learn spatio-temporal features of human motion.

1. Introduction

This work investigates learning a structured latent space that allows to represent and generate temporally and spatially dense 4D human body motion. Once trained, the proposed model generates a multi-frame sequence of dense 3D meshes based on a single point in a low-dimensional latent space. Recently, several methods have been proposed to learn such motion priors for 4D human body sequences of arbitrary motion and fixed duration by capturing information about pose changes over time [16, 38, 17]. In this work, we investigate an orthogonal scenario which models sequences of varying duration by considering motions that are sufficiently similar to allow for temporal alignment. We demonstrate experimentally that the resulting latent space is structured in the sense that similar motions form clusters in this space. Fig. 1 visualizes the learned model and latent space.

Learning a generative model of human motion with an underlying 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, *e.g.* virtual and augmented reality, 3D telepresence, and content generation for entertainment applications. Thanks to capturing a spatio-temporal motion prior, the model also opens possibilities for a wide set of completion tasks from temporally sparse, spatially sparse or incomplete inputs, for shape sequence reconstruction, motion transfer and retargeting. A structured latent space has the potential to allow for intuitive control when generating motion, *e.g.* by allowing for meaningful interpolations between pairs of input motions.

Learning a generative model for spatially and temporally dense 3D human motion data of varying duration presents two major challenges. First, the model needs to capture the intertwined variations of different factors, in-

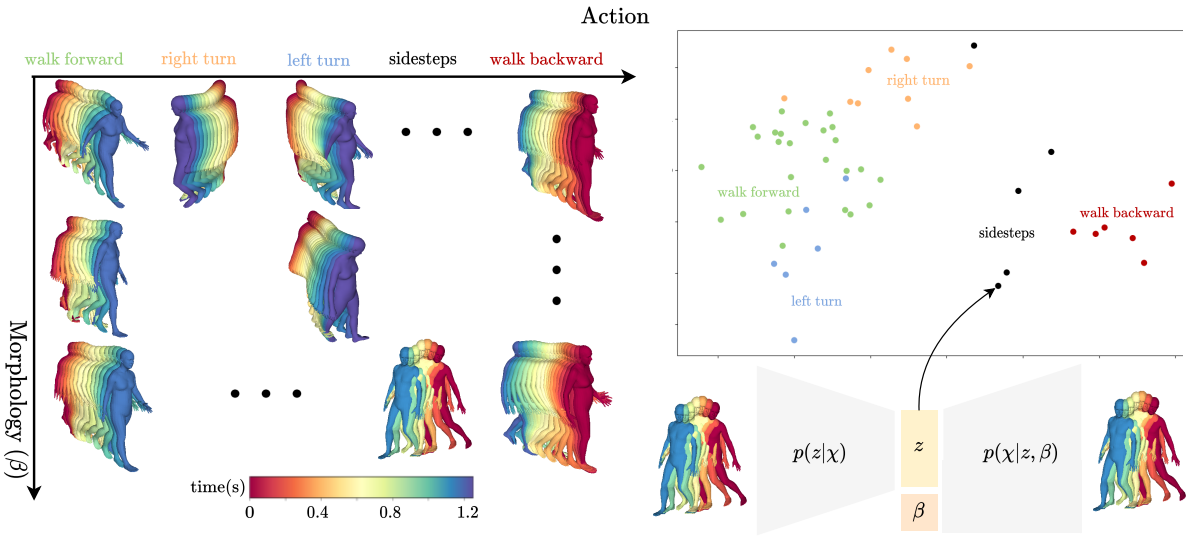


Figure 1: We learn a latent motion space from multi-frame 4D sequences. Left: Training sequences are different motions performed by different subjects (color-coded as shown in legend). Bottom right: Encoder-decoder architecture learns a latent motion space that encodes a motion sequence χ into a latent parameter vector z; the decoder conditions z on morphology β. Top right: Structured latent space. Plot shows a 2D linear projection of 51 motions in latent space, actions form clusters.

cluding 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], 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 global motion and the temporal evolution of the motion [34, 31], and good results can be achieved using deep learning based methods *e.g.* [21]. The second line of work considers spatially dense static data, with the goal of representing realistic 3D human body surfaces in a low-dimensional shape space [3, 26], which allows for detailed static 3D reconstructions of body pose and morphology that may even include soft tissue components learned from dynamic data *e.g.* [37]. Both lines of work follow a data-driven strategy and use aligned training data to allow for meaningful comparison as is common in morphable body models [35, 2].

We combine the advantages of these two concepts 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 models, we align the training sequences both temporally and spatially, which leads to comparisons at corresponding instances of the motion and anatomically corresponding points. In particular, we consider motions that vary significantly in duration while being geometrically sufficiently similar to allow for temporal alignment, performed by actors in minimal clothing to allow for effective spatial alignment. By building on works for modeling temporally dense skeleton data, our method learns a motion space that encodes a full motion sequence in a single latent space vector. Learning from spatio-temporally aligned data results in a structured motion space that can

be viewed as a morphable 4D human motion model, and that captures the intertwined variations of factors contributing to the 4D motion. To address the challenge of the increased complexity caused by spatially dense data, we opt for a low-dimensional shape space parameterization of static human bodies.

In our experiments, we consider motions performed by minimally dressed subjects 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. We demonstrate experimentally that our method learns a structured latent space which allows generating varying motions. We visualize the structure of our latent space to demonstrate that different actions (*e.g.* walking) form clusters, and use our latent space to generate plausible interpolations between different types of locomotion that outperform linear and per-frame interpolation baselines. Our motion space 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 spatial and temporal completion, our method outperforms a state of the art motion prior that encodes human motion sequences of fixed duration [38] as long as the data is sufficiently densely sampled in time, in spite of being trained on significantly less data. For spatial completion, our method outperforms a state of the art method when few samples are available [41].

We make the following major contributions. First, we propose a latent motion space that allows to represent and generate multi-frame sequences of dense 3D meshes of varying duration, while accounting for the interaction between morphology and motion. Second, we demonstrate that this latent space is structured: similar motions form clusters, and linear interpolation in latent space gives intuitive results. Third, when using our motion space as prior we outperform the state of the art for the application of motion completion from sparsely sampled data in space or time.

2. Related Work

There is a vast literature on the generation of human models and motions. Strategies that encode moving human bodies can be divided into temporally dense, spatially dense and full 4D methods. Temporally dense encompasses methods that learn the structure of human motion on a representation that is sparse in 3D space, while 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.* [29, 33, 34, 10]. 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, 30]. Recent works in this area leverage deep learning techniques, and can decouple variations due to different factors *e.g.* [11, 7, 42] or include hands, faces and soft-tissue deformation, *e.g.* [37]. 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, 40]. 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 information. We show experimentally that our model generalizes better than a linear one learned with 4D data.

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] 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]. These works have successfully been applied to human motion data. However, the high dimensionality of the 4D data constrains the sequences to few frames.

To consider longer temporal spans, another line of work 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 [28, 15]. Most similar to our work are methods that build motion priors of unlabeled 4D human motion data [16, 38, 17]. 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 [38] for motion completion.

3. Generative modelling of multi-frame sequences

This section outlines our method to learn a generative model of multi-frame human motion sequences. During training, the method takes as input a database of 4D human motion sequences that perform a cyclic motion of the hip joint and learns a latent motion space. Each point in motion space represents a 4D motion sequence, and we are interested in learning a motion space with structure, where similar locomotions (*e.g.* all walking motions) tend to form clusters. During inference, our model allows to reconstruct a 4D human motion sequence from a single point z in motion space and a parameter vector β representing the morphology of the subject performing the motion.

When building such a model, two major challenges need to be addressed. First, the amount of data that needs to be processed for training is large and unstructured. 4D human motion sequences are produced by acquisition systems that capture hundreds of frames containing thousands of vertices each. The raw capture data is unstructured, which makes it difficult to compare individual frames, let alone multi-frame sequences.

To address this challenge, we propose a new motion representation that is both compact and structured, as detailed in Section 3.1. We build on existing shape spaces of static bodies, which allow for a compact representation of one frame. By enhancing per-frame static shape space representations with information on the global spatial and temporal evolution of the motion, our representation explicitly decouples pose, morphology, global displacement and temporal information.

The second challenge is modeling the intertwined variations of the different factors influencing the motion, and taking their correlation into account when generating 4D sequences. While it is known that different factors including morphology influence the overall 4D motion [34], modeling these interactions explicitly is not straight forward.

To address this challenge, we propose a data-driven framework where the model is learned using an encoder-decoder architecture, as detailed in Section 3.2. The architecture conditions the motion generation on a representation of morphology to explicitly model the interaction between morphology and motion. Section 3.3 outlines how the model is trained.

3.1. Representation of motion sequences

This section introduces a compact and structured representation for 4D sequences, which is illustrated in the top left of Fig. 2. To represent motion data, we need to align an unstructured spatio-temporal motion signal.

Temporally, we uniformly sample n frames from the motion signal, we refer to these n frames as anchor frames in the following. These anchor frames allow to represent motions of various duration with the same number of frames. We experimented with a more complex sampling method using dynamic time warping ([4]) to further temporally align the data, but this leads to similar results as the simpler uniform sampling. Spatially, we build on static shape spaces to align the frames *e.g.* [22, 25, 19]. Shape space models represent static 3D human body surfaces by projecting them on a common mesh template, thereby providing correspondences over time, as well as correspondences between motions. By projecting the anchor frames on a template, we obtain n spatially aligned anchor meshes, making motion comparison practical.

However, the resulting anchor mesh sequence $M = [m_1, \ldots, m_n]$ does not represent the temporal evolution of a motion. The temporal sampling causes a loss of information, as it does not discriminate between similar motions with different temporal unfolding like walking and running. Therefore, we associate each anchor mesh m_i with a timestamp τ_i. We denote the timestamp vector by $\mathcal{T} = [\tau_1, \ldots, \tau_n]$. The representation $[M, \mathcal{T}]$ is a high-dimensional representation of a motion. To make processing easier, and to disentangle the influence of morphology on motion, we further leverage the shape space models. Shape spaces provide a low dimensional representation of the meshes m_i that decouples the influence of morphology and pose for static data. By holding morphology constant over M, we can represent each m_i using parameter vectors for morphology β, pose θ_i, and global translation γ_i. While any decoupled static model can be used, *e.g.* [11, 7, 42], in our implementation we chose the commonly used SMPL model [19] as the AMASS dataset ([20]) is parameterized by this model. We denote the model function by $SMPL$ such as $m_i = SMPL(\theta_i, \gamma_i, \beta)$ 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, $[\Theta, \Gamma, \beta, \mathcal{T}]$ gives 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 Γ and $\mathcal{T}$ in the multi-frame sequence representation. The timestamps allow the network to place freely and on any time span length the anchor meshes, thereby encoding the temporal unfolding of the motion in latent space and allowing to encode motions with various duration using a constant number of meshes.

To simplify notations and emphasize the difference between motion and morphology parameters, we denote by $\chi = [\Theta, \Gamma, \mathcal{T}]$ the motion parameters and we introduce the motion representation function $\mathcal{F}$ such that $[M, \mathcal{T}] = \mathcal{F}([\chi, \beta])$. As pre-processing for training, we require to map a raw motion sequence to the SMPL mesh template, and we use existing solutions to solve this problem [39, 20]. We denote by $SMPL^{-1}$ the mapping function which associates a single raw motion frame to its representation parameters θ, γ, β.

In practice, we represent parameters β and Γ as in SMPL. Pose features Θ are joint rotations of a skeleton, and are represented by a 6D rotation representation [43]. This representation models rotations in a continuous manner and was shown to outperform other representations when training neural networks.

3.2. Architecture

Our goal is to generate multi-frame 4D human motion. One interesting aspect is to learn the relationship between morphology and spatio-temporal motion patterns. Variational autoencoders were shown to be highly effective generative models. Furthermore, the CVAE architecture [32] allows to condition both encoder and decoder on input variables, thereby learning conditional distributions.

Our architecture is close to a CVAE, and shown in the bottom of Fig. 2. Our architecture consumes multi-frame sequences, thereby learning a latent motion space. In particular, the motion vector χ is encoded into a low-dimensional latent vector z, and the morphology representation β is used as condition for the decoder hereby allowing to capture dependencies between χ and β. Unlike in a classic CVAE we make the assumption that our latent variable z and the condition β are independant, hence learning a disentangled representation. Therefore we do not need the encoder to model the posterior distribution $p(z|\chi, \beta)$, but the posterior distribution $p(z|\chi)$. Practically, this is done by removing the condition β from the encoder input of a classic CVAE.

The encoder outputs are interpreted as mean μ and standard deviation σ of the posterior distribution of the latent space, and the corresponding latent vector z is randomly

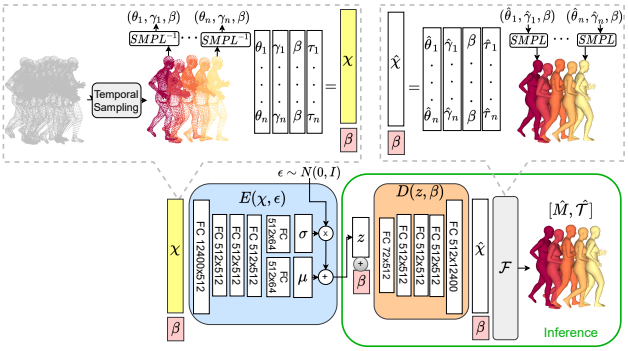


Figure 2: Overview of the motion representation and architecture. Top: motion representation. Left: pre-processing during training samples n anchor frames and extracts per-frame representations of pose θ, translation γ and morphology β with their timestamp τ to obtain motion representation χ and morphology β. 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 χ to a single latent vector z. At inference time, D conditions z on β to generate sequence features $\hat{\chi}$ (green box).

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

3.3. Training

The network is trained as a classical VAE with a two term loss: a reconstruction term which represents 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 the spatio-temporal representation χ and second, we replace it by a loss computed directly on the sequence of anchor meshes M in $\mathbb{R}^3$. Considering a first loss on χ first allows for a fast and memory efficient initialization.

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

To minimize these 3 losses, which do not have the same numerical magnitude, we use adaptive weights to trade off their relative influence [6]. These weights are updated automatically during training based on the norm of the gradient of the partial loss and a learning rate. This ensures that the partial losses are decreasing in similar proportions. 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 ω_{time} are the respective adaptive weights of the partial reconstruction losses.

Reconstruction loss in 4D. The second reconstruction loss is the squared L_2 loss on the 3D coordinates vectors of the anchor mesh sequence M. The contributions of $\mathcal{L}_{pose}$ and $\mathcal{L}_{trans}$ are merged into one spatial loss :

$$\mathcal{L}_{spatial} = (M - \hat{M})^2, \quad (2)$$

which gives the $4D$ reconstruction term

$$\mathcal{L}_{rec4D} = \omega_{spatial}\mathcal{L}_{spatial} + \omega_{time}\mathcal{L}_{time}, \quad (3)$$

where $\omega_{spatial}$ is a new adaptive weight of the spatial loss. Optimizing this loss leads to more accurate reconstruction of the 4D multi-frame sequences because it uses the full surface information, at the cost of higher computation time.

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 $\mathcal{N}(0,1)$, which is denoted by $\mathcal{L}_{KL}$.

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

$$\mathcal{L}_{init} = \mathcal{L}_{rec} + \omega_{KL}\mathcal{L}_{KL} \quad (4)$$

to provide a good initialization and subsequently

$$\mathcal{L} = \mathcal{L}_{rec4D} + \omega_{KL}\mathcal{L}_{KL} \quad (5)$$

to refine the model by using surface information.

4. Evaluation

This section presents an evaluation of the model. After presenting the data and implementation details, we study the influence of the latent space dimension and regularisation, and present a comparison to baselines. Furthermore, we provide experiments to show the structure of the learned latent space by visualizing labeled motion sequences in latent space and by linearly interpolating between pairs of input motion sequences in latent space. Finally, we demonstrate that the proposed model learns information on the interaction of morphology and motion by visualizing the motion changes caused by changing the morphology β for a fixed point z in motion space.

4.1. Data

Our model consumes human motion sequences densely captured in 4D. To learn a structured latent space for unlabeled sequences of different duration, we consider motion sequences for which a temporal alignment can be defined. In our experiments, we focus on motion sequences during which the hip performs a cycle automatically extracted from a dataset by comparing all subsequences of the dataset to a set of 4D template motions using dynamic time warping as distance. Subsequences are considered if this distance is below a threshold. As post-processing, we prune segments with a duration above 3 seconds or below 0.3 seconds. We manually generate two 4D template motions as gait cycles starting with the left and right foot, respectively.

In this work, we experiment with the AMASS dataset [20]. AMASS regroups a large set of MoCap recordings and fits SMPL with additional soft-tissue motions to all data, resulting in a semi-synthetic dataset. As recommended, when splitting the AMASS dataset into training and test set, we split the dataset according to the original MoCap datasets by treating all sequences emanating from the same MoCap dataset as one entity. We leave the data associated with MoCap datasets "MPI_mosh", "SFU"and "TotalCapture" for testing.

Training data. We automatically extract 12085 sequences of motion cycles with various duration and motion types from the AMASS training set which amounts to approximately 4.5 hours of motion data. To allow for efficient learning, the sequences are spatially aligned by zeroing the initial translation and we use the identity rotation as initial rotation of the root joint to be invariant in the ground plane.

Test data. We consider two test datasets. The first one is called AMASS test set in the following and contains 1027 sequences extracted from the AMASS test set which amounts to approximately 20 minutes of motion data. The second one, called Kinovis test set in the following, contains 4D motion sequences captured using a multi-view platform. This dataset is an extension of [39] and allows to evaluate the generalization of the model to densely captured 4D data. We consider all walking and running sequences, and pre-process the data by fitting SMPL to the 4D sequences before extracting cyclic hip motions. This results in 37 test sequences, some of which contain less than 100 frames; we augment shorter sequences to 100 frames using linear interpolation between the 6D rotations.

4.2. Implementation details

In our motion representation χ, we do not consider the SMPL components related to hands or dynamic components available in AMASS. We further discard the two foot joints because they have constant rotation. This leaves a total of 20 joints. Our representation χ consists of 100 timestamped anchor meshes, each of which is represented by 124 parameters (120 for θ, 3 for γ and 1 for τ). 100 anchor meshes are chosen as they provide a good trade-off between the error introduced by the sampling and the dimensionality of χ. To normalize the data, we normalize the translation γ in $[-1, 1]^3$, and the timestamps τ in $[0, 1]$ using minmax scaling over the training set. We remove the identity rotation $[1, 0, 0, 0, 1, 0]$ from the 6D representation, which leads to a significant gain in reconstruction accuracy compared to the classic scaling $\frac{\theta - \mu_\theta}{\sigma_\theta}$.

We train for 5000 epochs with $\mathcal{L}_{init}$, using a learning rate of $1e^{-3}$ and a batch size of 256. Each epoch takes 6 s for a total training time of 8 hours. We train with $\mathcal{L}$ for 200 epochs using a smaller batch size of 16 for memory reasons and a learning rate of $1e^{-4}$. Here epoch time is 8 min for a total training time of one day. The training is done on a NVIDIA Quadro RTX8000 with 48G of GPU

RAM. We use $\omega_{kl} = 0.01$ for both steps and chose a latent dimension for the motion space z of 64, and of 8 for β. Note that mesh vertex positions are in meters during training. We initialize all dynamic weights of Eq. 1 and 3 at 1.0 and GradNorm [6] updates the weights dynamically.

4.3. *Influence of latent space dimension and regularisation*

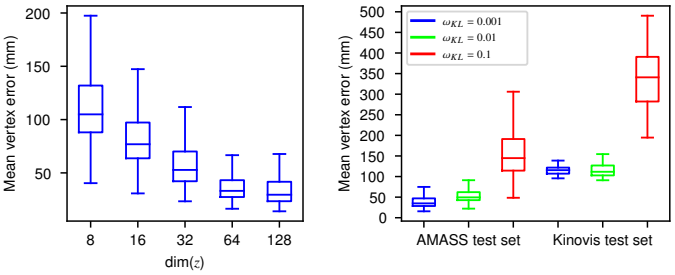
We now investigate the influence of the latent space dimension and regularization on the quality of the model. To evaluate the model's quality, we measure its reconstruction error, which characterizes the model's ability to reconstruct examples unseen during training and is defined as

$$\frac{1}{nk}\left(M - \hat{M}\right)^2, \tag{6}$$

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. As second qualitative error measure, we consider the model's ability to allow for the generation of plausible new sequences by sampling in latent space. In practice, we consider samples that are linearly interpolated between sequences of the test set.

Latent space dimension. We first study the influence of the dimensionality of the latent space on the model quality. Fig. 3a shows the impact of the dimension of z on the reconstruction error on the AMASS test set. As expected, the bigger the latent space dimension, the smaller the error. However, for $dim(z) > 64$, the error starts to stagnate. Therefore we set the dimension of the latent space to 64.

Latent space regularisation. The regularisation of the latent space has a major impact on the model quality. It is controlled by coefficient ω_{KL}, which weighs the influence of latent space regularization at the cost of reconstruction accuracy. Fig. 3b shows the reconstruction error on models trained with different values for ω_{KL}. The smaller ω_{KL}, the smaller the reconstruction error. However, with $\omega_{KL} = 0.001$, the model no longer allows generating plausible new sequences. Therefore, we set $\omega_{KL} = 0.01$ in the following.



(a) Influence of dimension (b) Influence of regularization

Figure 3: Influence of latent space dimension and regularisation on reconstruction error. Left: Increasing dimension of the latent space leads lower reconstruction errors on AMASS test set. Right: Smaller regularization ω_{KL} leads lower reconstruction errors on both test sets. Boxes follow Tukey's method [36].

4.4. *Comparison to baseline models*

We compare our model to two baselines with respect to the reconstruction error defined in Eq. 6. The first baseline applies a linear principal component analysis (PCA) to our motion 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 $\hat{\beta}$ from latent space. To provide a fair comparison, we consider the original β instead of $\hat{\beta}$ in PCA reconstructions and set the PCA latent dimension to $dim(z) + dim(\beta)$. The second baseline considers our model after initialization when only loss $\mathcal{L}_{init}$ is optimized that operates on a skeleton-level representation over time, thereby evaluating the value of learning from data that is densely sampled in space and time.

Fig. 4 shows the reconstruction error for the different models. While PCA already provides low reconstruction errors, these are further improved using our non-linear model. Our model also improves over its initialization, which shows that considering spatio-temporal data that is densely sampled significantly impacts the reconstruction performance of the model.

4.5. *Motion space structure and interpolation*

We now investigate the structure of the latent motion space. One novelty of our method is its ability to learn from sequences of various duration by using as representation a sequence of time-stamped anchor frames. Fig. 1

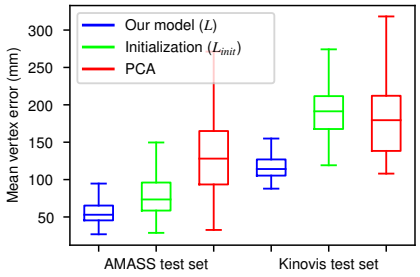


Figure 4: Comparison to baseline models in terms of reconstruction error. Our model outperforms a linear PCA baseline (blue) and a baseline that considers sparse spatial sampling at skeleton level (red). Boxes follow Tukey's method [36].

visualizes the latent motion space after linearly reducing its dimension to two. The space contains 51 motions that are labeled by actions for the purpose of visualization. Each action is assigned a color and points corresponding to the same action tend to form clusters in latent space. This demonstrates that our motion representation allows learning a latent space in which sequences showing similar actions are clustered.

This structure in latent space can be exploited to generate plausible interpolations between two input motion sequences using simple linear interpolation. Given a start and a target motion sequence as input, we encode them as (z_s, β_s) and (z_t, β_t), and generate their interpolating motion sequence by decoding $((1-k)z_s + kz_t, (1-k)\beta_s + k\beta_t)$ at an arbitrary intermediate position $k \in [0, 1]$.

We compare our results to two baselines. The first baseline uses the PCA model from the previous section, where PCA is applied to our motion representation $[\chi, \beta]$, which results in a motion model that allows for linear interpolations in its latent space. This comparison, called PCA in the following, 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 (SLERP) between skeletal poses. This comparison, called SLERP in the following, evaluates the value of learning a motion model instead of operating independently per-frame. For all interpolations, visualizations show the mid-point at $k = 0.5$ for all methods.

In the following, we interpolate between input sequences that differ in terms of their duration, displacement, and (global and local) pose, *i.e.* each of the factors encoded in our motion representation χ.

Interpolating sequences of different duration. To inspect the temporal information learned by our model, we interpolate between a running and a walking motion, which have different duration and dynamics. For our model, the duration, given by τ_n, of the intermediate sequences monotonically decreases when going from running to walking, and the intermediate sequences are realistic as shown in Fig. 5 (top left), proving that our motion space has captured information on the temporal evolution of the motion. Both the 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. We observe that our intermediate sequence corresponds to a really small step as shown in Fig. 5 (top right). There were no steps this small in the training set. Our latent space has captured information on Γ and is able to generate interesting new sequences. The PCA and SLERP baselines fail to interpolate global translation realistically, which results in foot skating.

Interpolating sequences of different pose. To inspect the learned information of pose, we distinguish between global pose and pose articulation of the body. 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 turn to a right turn as shown in Fig. 5 (bottom left), leading to a meaningful interpolation. In this case, both PCA and SLERP baselines fail due to the ambiguity when interpolating between opposite rotations, while our model is able to leverage spatio-temporal information to alleviate this ambiguity.

Second, we interpolate between a walking motion and a walking motion while carrying an object on the head. The intermediate sequence with our model results in a realistic intermediate position for the arms, and gradually elevates them to head level as shown in Fig. 5 (bottom right). This shows that the latent space has captured information on Θ,

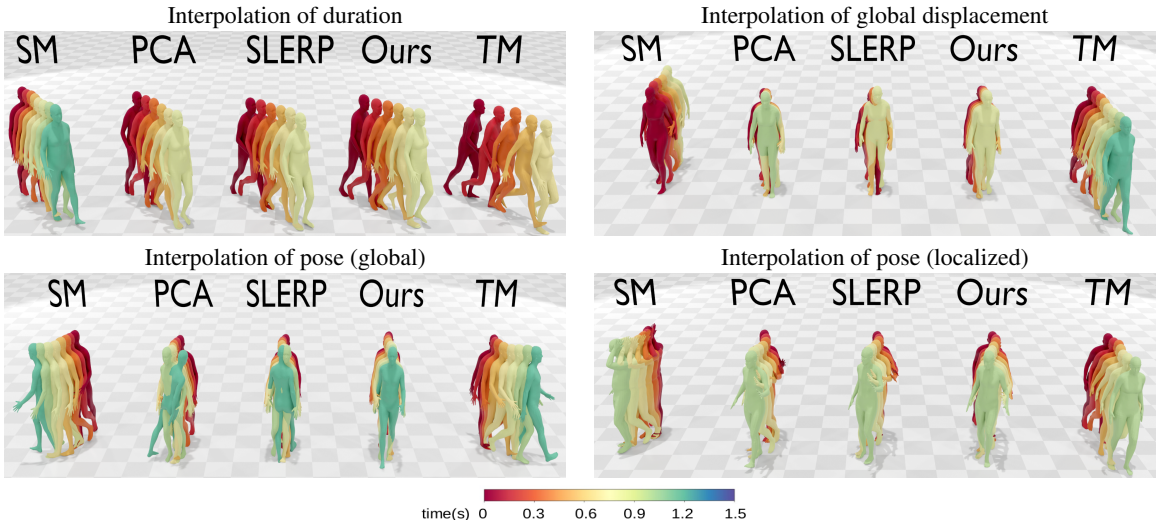


Figure 5: Linear interpolations in latent motion space. Each figure left to right : starting motion, interpolation in PCA space, direct interpolation using SLERP, interpolation with our model and target motion. Sequences are color-coded by the time at which the rendered frames appear in the sequence. **Top-Left** Interpolating running and walking. **Top-Right** Interpolating walking backward and forward. **Bottom-Left** Interpolating left and right turn. **Bottom-Right** Interpolating a walk and a walk carrying an object on the head. All interpolations with our model are plausible, while baselines fail when interpolating global displacement or pose.

and can generate interesting motions unseen during training. Both the PCA and SLERP baselines lead to plausible interpolations.

In summary, while our model generates visually plausible interpolations for all types of parameters encoded in the motion representation, both baselines exhibit failure cases in some scenarios. This demonstrates the value of learning a non-linear 4D human motion model.

4.6. Interaction between morphology and motion

To allow our model to capture the interaction between morphology and motion, the decoder conditions a vector in motion space z with morphology β to generate a 4D motion sequence.

We examine the influence of β on the final output 4D motion χ learned by our model. To this end, we consider a fixed jogging motion represented by point z^* in our motion space and investigate χ when setting β to ± 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 $\beta = 0$, *i.e.* we look at the gradient learned by our decoder with respect to morphology at the mean shape.

We compare our result to a baseline that reconstructs a dense 3D body model at every frame of the jogging motion independently with the initial pose parameters and β using SMPL [19]. This evaluates the influence of learning the interaction between morphology and motion.

Fig. 6 shows the impact of the first (left) and second (right) principal components of β. The top row shows a color coding of the gradient learned by our decoder with respect to β 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 per-frame baseline color-coded by the distance to the result of our model.

Changing the first principal component impacts the body shape of the subject to change the perceived gender. For our model, this impacts the 4D motion on the right shoulder and the left hip (see top row of Fig. 6), which is in line with prior studies that show that shoulder sway is statistically gender related and that the movement of the hips tend to be more pronounced for women [35]. Note that the spatio-temporal areas affected by our motion model are the ones where the baseline leads to a significantly different result with up to $10cm$ distance.

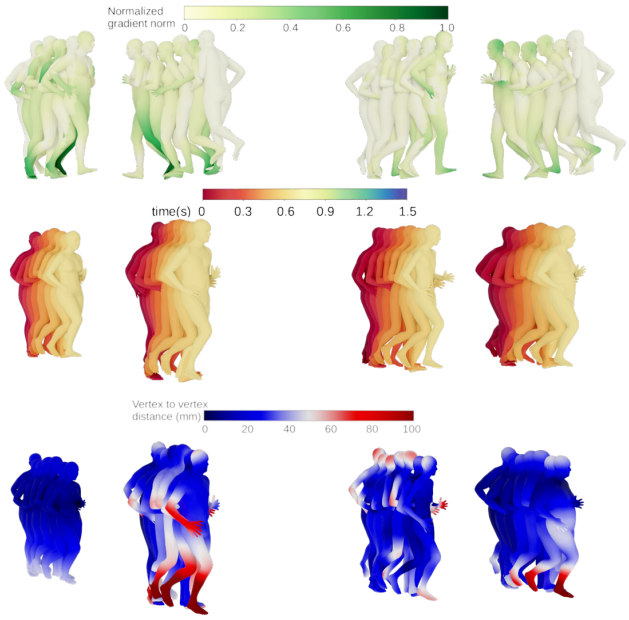


Figure 6: Interaction between morphology and motion demonstrated *w.r.t.* first (left) and second (right) principal components of morphology. Top: for our method, visualization of the normalized gradient of the 4D motion *w.r.t.* the morphology vector. Middle: for our method, inferences with fixed latent motion vector and morphologies taken at ±3 standard deviations. Bottom: baseline per-frame motion transfer using SMPL for same fixed motion and morphologies taken at ±3 standard deviations, color coded by per-vertex distance to result of our method. Our learnt correlation has a significant impact on the motion, which differs up to 10cm from baseline.

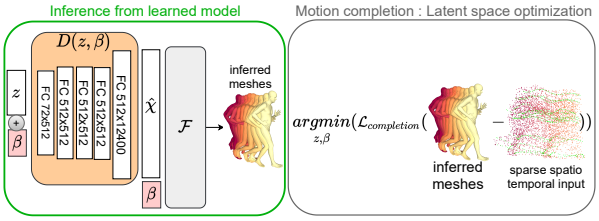


Figure 7: Motion completion. We minimize a loss *w.r.t.* the latent representation (z, β). Left: inference pipeline. Right: we optimize $\mathcal{L}_{completion}$ between a sparse 4D point cloud and the inferred meshes.

Changing the second principal component leads to a perceived weight change of the subject. For our model, this impacts the 4D motion at the right arm and head/neck poses (see top row of Fig. 6). Again, the spatio-temporal areas affected by our motion model are the ones where the baseline leads to a significantly different result with up to 10cm distance.

This allows to conclude that our model learns meaningful interactions between morphology and motion, which leads to 4D motion sequences that differ significantly from a baseline that applies $SMPL$ independently per frame.

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

This section demonstrates our model's performance for spatio-temporal completion. Given as input a set of sparse, unmatched and temporally incoherent point clouds, our model can retrieve a spatio-temporally aligned sequence of meshes by leveraging its learned priors of human surface motion. This is interesting in 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 the full human body in motion from a space 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, let $L = [l_1, \ldots, l_n]$ denote

the corresponding synchronized sequence of landmarks, and let $\mathcal{T} = [\tau_1, \ldots, \tau_n]$ denote the corresponding time stamps. As we consider temporally sparse input, 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 τ_i are given.

Our goal is to complete this data, *i.e.* to compute a sequence of anchor meshes $\hat{M}$ with associated time stamps $\hat{\mathcal{T}}$ that approximate the input. To achieve this, we decode a full sequence of anchor frames $[\hat{M}, \hat{\mathcal{T}}]$ using $\mathcal{F}(D(z,\beta),\beta)$ and optimize for latent vectors z^*, β^* as

$$z^*, \beta^* = \underset{z,\beta}{argmin}(\mathcal{L}_{completion}([\hat{M}(z,\beta), \hat{\mathcal{T}}(z,\beta)], [S, L, \mathcal{T}])), \tag{7}$$

where

$$\mathcal{L}_{completion} = \omega_{dense}\mathcal{L}_{dense} + \omega_{mocap}\mathcal{L}_{mocap} + \omega_{time}\mathcal{L}_{time} \tag{8}$$

$$\mathcal{L}_{dense} = \sum_{i \in I} \text{Chamfer}(\hat{m}_i(z,\beta), s_i) \tag{9}$$

$$\mathcal{L}_{mocap} = \sum_{i \in I} \text{Landmark}(\hat{m}_i(z,\beta), l_i) \tag{10}$$

$$\mathcal{L}_{time} = \sum_{i \in I} (\hat{\tau}_i(z,\beta) - \tau_i)^2. \tag{11}$$

The weights ω_{dense}, ω_{mocap} and ω_{time} are adaptive [6]. We set $\omega_{mocap} = 0$ when no landmarks are given as input and $\omega_{dense} = 0$ when no dense scan is given as input. Varying ω_{mocap} allows to evaluate the benefit of having corresponding points over time for the completion task. Chamfer is the Chamfer distance between two point clouds and Landmark is the squared Euclidean distance between k vertices of the SMPL template, selected once for all experiments, and the K given landmarks. This optimization is visualized in Fig. 7.

5.2. Completion dataset

To evaluate the motion completion, we introduce a new dataset of cyclic human motion (CHUM), which was captured using a 4D modeling platform with 68 RGB cameras and a standard Qualisys MoCap system. Data consists of dense scans of approximately 10000 points acquired at 50fps with synchronised MoCap data 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

In the following, we evaluate the accuracy of our model when reconstructing dense 4D data from sparse input.

We compare our results to two state of the art approaches. The first one is a static 3D completion method [41] applied per-frame. Note that due to its high computational complexity, we apply the static method to a subset of CHUM while our method is applied to the full dataset. This method is only applicable in case of spatial completion where observations are available at every frame. The second one is a motion space for sequences of fixed duration that can serve as prior [38] trained on approximately 34hours of motion data from the AMASS dataset. Given a partial motion as input, we optimize a latent vector in motion space along with a morphology parameter and a set of per-frame translation parameters, as morphology and global translation are not encoded in this motion space. The goal is to approximate the input, and we optimize for $\mathcal{L}_{completion}$ with $\omega_{time} = 0$, as the motion space is designed for sequences of fixed duration and cannot benefit from time stamp information. In case of temporally sparse input, per-frame translation parameters are only optimized for frames in I and the remaining translations are found using linear interpolation between the closest observed frames. This method is applicable for both spatial and temporal completion The motion prior we compare to uses a latent space of 256 dimensions. For fair comparison, we re-train our model with 256 latent dimensions for this application.

Spatial completion. We first evaluate the quality of spatial completion by simulating different levels of spatial sparsity of the data. To do so, we vary the number of points p per scan s_i. The sampled points are not in correspondence across time.

Table 1 shows the evolution of the reconstruction error in *mm* when varying p. Our method clearly outperforms

Table 1: Comparative evaluation of motion completion. Mean and standard deviation of Chamfer distance in *mm* (the lower the better), 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**
Zhou *et al.* [41]	N.A.	58±0.99	47 ±0.6	21±0.3	**10±0.46**	N.A.	N.A.	N.A.
Xu *et al.* [38]	46 ± 52	26±8	24 ±9	22±10	22±10	**22±10**	22±11	22±10

the static method [41] 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 input data ($p = 10000$). 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. For dense scans however, the static method, which deforms a template, can capture higher levels of geometric detail. Our method further outperforms the motion space for sequences of fixed duration [38], in spite of being trained on significantly less motion data (4.5h for ours vs. 34h for Xu *et al.* [38]).

A qualitative comparison for the completion task with $p = 100$ and available landmark data is shown in Fig. 8. Note that our method leads to more plausible wrist and hand motion than Xu *et al.* [38] and better temporal coherence and leg motion than Zhou *et al.* [41].

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 [38] is better for sparsely sampled temporal data, we outperform this method 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. This study focuses on locomotions with cyclic hip movements. One possible avenue is to investigate the inclusion of a wider variety of actions with attention-based architectures, which have shown good results in sequence processing.

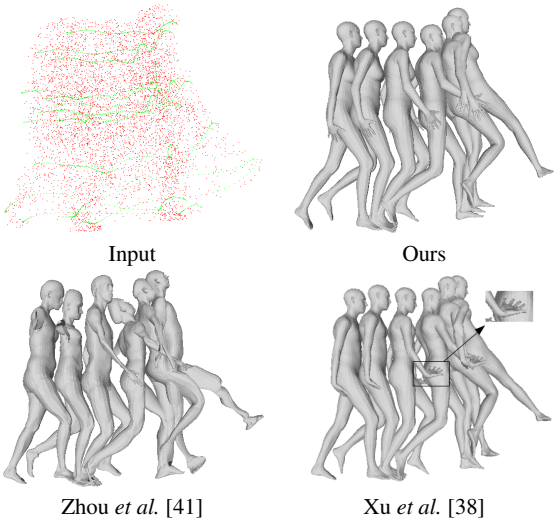


Figure 8: Qualitative comparison of spatial completion on kick sequence from CHUM with $p = 100$. Input scans shown in red, landmarks in green. Visualization shows 6 of 100 completed frames. Note that our motion completion is plausible and coherent with input.

7. Acknowledgements

We thank Jinlong Yang and Jiabin Chen for providing us the Kinovis test set, Joao Regateiro, Anne-Hélène Olivier and Edmond Boyer for helpful discussions, and the Kinovis platform at Inria Grenoble, the engineers Laurence Boissieux and Julien Pansiot, and our volunteer sub-

jects for help with the 4D data acquisition. This work was partially funded by ANR grant 3DMOVE - 19-CE23-0013-01.

References

[1] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Skeikh. Bilinear spatiotemporal basis models. ToG, 31:#17:1–12, 2012.

[2] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: reconstruction and parameterization from range scans. ToG, 22(3):587–594, 2003.

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. ToG, 24(3):408–416, 2005.

[4] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In DMKD, pages 359–370, 1994.

[5] A. Boukhayma and E. Boyer. Surface motion capture animation synthesis. TVCG, 2018.

[6] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In PMLR, volume 80, pages 794–803, 2018.

[7] L. Cosmo, A. Norelli, O. Halimi, R. Kimmel, and E. Rodolà. Limp: Learning latent shape representations with metric preservation priors. In ECCV, 2020.

[8] S. Ghorbani, C. Wloka, A. Etemad, M. Brubaker, and N. Troje. Probabilistic character motion synthesis using a hierarchical deep latent variable model. In SCA, 2020.

[9] M. Habermann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Real-time deep dynamic characters. TOG, 40(4):94:1–16, 2021.

[10] D. Holden, J. Saito, and T. Komura. A deep learning framework for character motion synthesis and editing. ToG, 35(4):#138, 2016.

[11] B. Jiang, J. Zhang, J. Cai, and J. Zheng. Disentangled human body embedding based on deep hierarchical neural network. TVCG, 2020.

[12] B. Jiang, Y. Zhang, X. Wei, X. Xue, and Y. Fu. Learning compositional representation for 4d captures with neural ode. In CVPR, pages 5340–5350, 2021.

[13] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In CVPR, 2019.

[14] A. Kuznetsova, N. Troje, and B. Rosenhahn. A statistical model for coupled human shape and motion synthesis. In Conf. Comput. Graph. Theory App., 2013.

[15] Y. Lee, K. Wampler, G. Bernstein, J. Popović, and Z. Popović. Motion fields for interactive character locomotion. In ACM SIGGRAPH Asia 2010 papers, pages 1–8. 2010.

[16] J. Li, R. Villegas, D. Ceylan, J. Yang, Z. Kuang, H. Li, and Y. Zhao. Task-generic hierarchical human motion prior using VAEs. In 3DV, 2021.

[17] S. Lohit, R. Anirudh, and P. Turaga. Recovering trajectories of unmarked joints in 3d human actions using latent space optimization. In WACV, pages 2342–2351, January 2021.

[18] M. Loper, N. Mahmood, and M. Black. MoSh: motion and shape capture from sparse markers. ToG, 33, 2014.

[19] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. ToG, 34(6):1–16, 2015.

[20] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: archive of motion capture as surface shapes. In ICCV, pages 5442–5451, 2019.

[21] J. Martinez, M. Black, and J. Romero. On human motion prediction using recurrent neural networks. In CVPR, 2017.

[22] A. Neophytou and A. Hilton. Shape and pose space deformation for subject specific animation. In 3DV, 2013.

[23] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In ICCV, pages 5379 – 5389, 2019.

[24] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In ICCV, pages 10985–10995, Oct. 2021.

[25] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. Pattern Recognit., 67:276–286, 2017.

[26] G. Pons-Moll, J. Romero, N. Mahmood, and M. Black. DYNA: a model of dynamic human shape in motion. ToG, 34(4):120:1–14, 2015.

[27] J. Regateiro, A. Hilton, and M. Volino. Dynamic surface animation using generative networks. In 3DV, 2019.

[28] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas. Humor: 3d human motion model for robust pose estimation. In ICCV, pages 11488–11499, 2021.

[29] C. Rose, M. F. Cohen, and B. Bodenheimer. Verbs and ad-verbs: multidimensional motion interpolation. Comput. Graph. Appl., 18:32–40, 1998.

[30] I. Santesteban, E. Garces, M. A. Otaduy, and D. Casas. SoftSMPL: Data-driven Modeling of Nonlinear Soft-tissue Dynamics for Parametric Humans. CGF, 2020.

[31] L. Sigal, D. Fleet, N. Troje, and M. Livne. Human attributes from 3d pose tracking. In ECCV, 2010.

[32] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In NIPS, pages 3483–3491, 2015.

[33] N. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. J. Vis., 2:371–387, 2002.

[34] N. Troje. Retrieving information from human movement patterns. In Understanding Events: From Perception to Action, pages 308–334. Oxford Univ. Press, 2008.

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

[36] J. W. Tukey. Box-and-whisker plots. EDA, pages 39–43, 1977.

[37] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In CVPR, 2020.

[38] J. Xu, M. Wang, J. Gong, W. Liu, C. Qian, Y. Xie, and L. Ma. Exploring versatile prior for human motion via motion frequency guidance. In 3DV, 2021.

[39] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In ECCV, 2016.

[40] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik. Predicting 3d human dynamics from video. In ICCV, 2019.

[41] B. Zhou, J.-S. Franco, F. Bogo, B. Tekin, and E. Boyer. Reconstructing human body mesh from point clouds by adversarial gp network. In ACCV, 2020.

[42] K. Zhou, B. L. Bhatnagar, and G. Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In ECCV, pages 341–357, 2020.

[43] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the Continuity of Rotation Representations in Neural Networks. In CVPR, page 9, 2019.