Supplementary Material for *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

1. Data

We start by providing more details on the training and test data used from the AMASS [3] and Kinovis [6] datasets. AMASS regroups a large set of MoCap recordings and fits a parametric body model to all data. When splitting the AMASS dataset into training and test sets, we treat all sequences emanating from the same MoCap dataset as one entity. We leave the MoCap datasets "MPI_mosh", "SFU", and "TotalCapture" for testing, and call this dataset *AMASS test set*. The Kinovis dataset contains 4D motion sequences captured using a multi-view platform and allows to evaluate the generalization of the model to densely captured 4D data. We consider all walking and running sequences, preprocess the data by fitting SMPL before extracting cyclic hip motions, and call this dataset *Kinovis test set*.

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 [6] and allows to evaluate the generalization of the model to densely captured 4D data. We consider all walking and running sequences, and preprocess 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.

*Institute of Engineering Univ. Grenoble Alpes

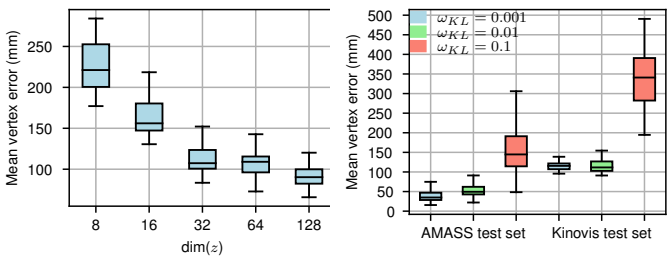
2. Implementation Details

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 to 1.0 and GradNorm [1] updates the weights dynamically.

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



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

Figure 1: 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 [4].

error, which characterizes the model's ability to reconstruct examples unseen during training and is defined as

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

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. 1a 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. 1b 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 and a problematic interpolation is shown in supplementary material. Therefore, we set $\omega_{KL} = 0.01$ in the following.

4. Motion completion from spatially sparse input

A qualitative comparison for the completion task with $p = 100$ and available landmark data is shown in Fig. 2. Note that our method leads to more plausible wrist and hand motion than [5] , better temporal coherence than [7], better leg motion than [7] and [2] and better global translation than [2].

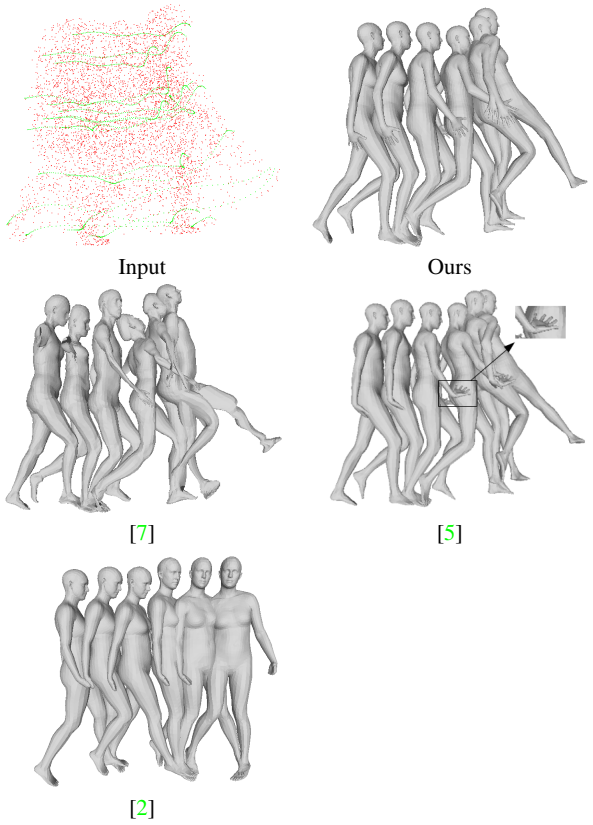


Figure 2: Qualitative comparison of spatial completion on kick sequence from CHUM with $p = 100$. Input scans shown in red, landmarks in green. Visualization shows 6 of 100 completed frames. Note that our motion completion is plausible and coherent with input.

References

[1] 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. 1

[2] 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. 2

[3] 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. 1

[4] John W Tukey. Box-and-whisker plots. Exploratory data analysis, pages 39–43, 1977. 2

[5] 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. 2

[6] 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. 1

[7] 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