Generating Smooth Pose Sequences for Diverse Human Motion Prediction —Supplementary Material—

Wei Mao[1], Miaomiao Liu[1], Mathieu Salzmann[2,3]
[1]Australian National University; [2]CVLab, EPFL; [3]ClearSpace, Switzerland
{wei.mao, miaomiao.liu}@anu.edu.au, mathieu.salzmann@epfl.ch

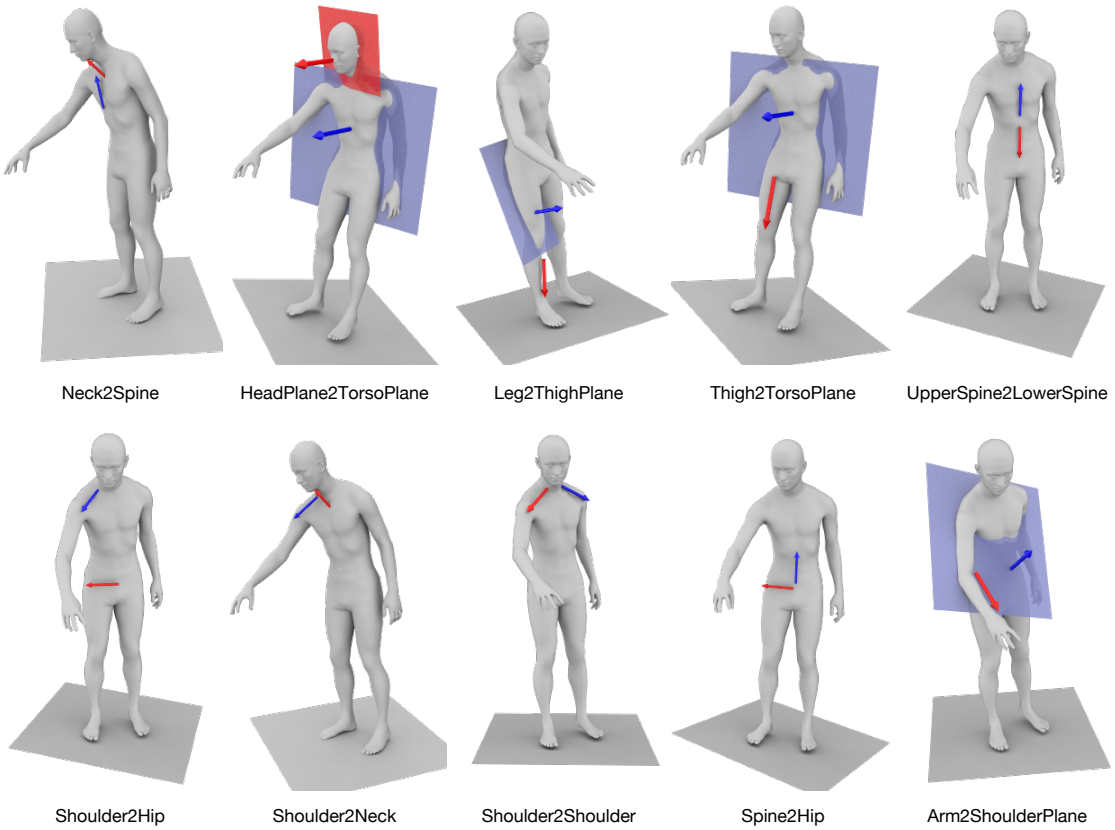

Figure 1: **Definition of additional angles.**

1. Details of Our Model

1.1. Pose Prior

Our pose prior aims to model the distribution of valid human poses. As the validity of a pose mostly depends on the kinematics of its joint angles, instead of modeling the distribution of 3D joint coordinates, which couple the limb directions with their lengths, we propose to learn the distribution of limb directions only. In particular, given the i-th joint coordinate $\mathbf{J}_i \in \mathbb{R}^3$ and the coordinates of its parent joint $\mathbf{J}_{p_i}$, the limb direction can then be computed as

$$\mathbf{d}_i = \frac{\mathbf{J}_i - \mathbf{J}_{p_i}}{\|\mathbf{J}_i - \mathbf{J}_{p_i}\|_2} \ . \tag{1}$$

We then represent a human pose as the directions of all limbs $\mathbf{d} = [\mathbf{d}_1^T, \mathbf{d}_2^T, \cdots, \mathbf{d}_J^T]^T \in \mathbb{R}^{3J}$, where J is the number of limbs, which, in our case, is equal to the number of joints.

As mentioned in the main paper, we choose a simple network with 3 fully-connected layers to model our pose prior. To ensure the invertiblity of the network, we formulate each fully-connected layer as

$$\mathbf{f}^{'} = \sigma(\mathbf{f}\mathbf{Q}\mathbf{R} + \mathbf{b}) \ , \tag{2}$$

where $\mathbf{f}^{'} \in \mathbb{R}^{3J}$ and $\mathbf{f} \in \mathbb{R}^{3J}$ are the feature vectors of the output and input, respectively; $\mathbf{Q} \in \mathbb{R}^{3J \times 3J}$ is an orthogonal matrix; $\mathbf{R} \in \mathbb{R}^{3J \times 3J}$ is an upper triangular matrix with

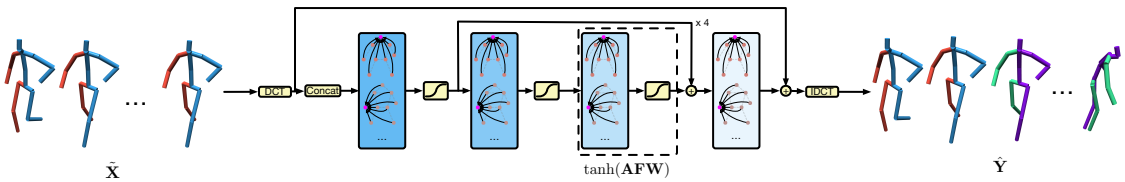

Figure 2: **Overview of our generator.** Note that, here, we show a generator that predicts the whole body motion ($N = 1$). In our experiments, however, we use the same architecture to predict the motion of human body parts.

positive diagonal elements; $\mathbf{b} \in \mathbb{R}^{3J}$ is the bias; $\sigma(\cdot)$ is the PReLU function.

1.2. Angle Loss

In Fig. 1, we visualise additional angles used to defined our kinematics constraints. In Table 1, we provide the valid motion range for most angles used.

Angles (in Degree)	Human3.6M [1]		HumanEva-I [4]	
	LowerBound	UpperBound	LowerBound	UpperBound
Neck2Spine	0	124	0	50
HeadPlane2TorsoPlane	0	120	-	-
Leg2ThighPlane	80	180	72	166
Thigh2TorsoPlane	0	140	0	58
UpperSpine2LowerSpine	110	180	-	-
Shoulder2Hip	0	84	-	-
Shoulder2Neck	32	134	-	-
Shoulder2Shoulder	83	180	128	180
Spine2Hip	60	120	-	-
Arm2ShoulderPlane	-	-	0	91

Table 1: **Ranges of different angles.** Depending on the skeleton model, some angles are undefined in some dataset.

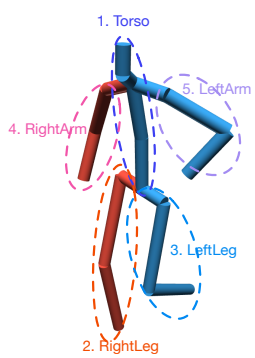

Figure 3: **Body parts.** We divide a pose into 5 parts: 1. torso, 2. right leg, 3. left leg, 4. right arm and 5. left arm.

1.3. Generator

Recall that the input to the generator is $\tilde{\mathbf{X}} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_H, \mathbf{x}_H, \cdots, \mathbf{x}_H]$ of length $H+T$. The first H frames are the motion history, and the remaining T frames are replications of last observed frame $\mathbf{x}_H$. The goal of the generator is to predict the DCT coefficients of the future motion $\hat{\mathbf{Y}} \in \mathbb{R}^{D\times(H+T)}$ given those of the replicated motion sequence, which we translate to learning motion residuals. Note that, to encourage a smooth transition between past poses and future ones, our generator not only predicts the future motion (corresponding to the last T frames of $\hat{\mathbf{Y}}$), but also recovers the past motion (corresponding to the first H frames of $\hat{\mathbf{Y}}$). Therefore, we define a reconstruction loss on the past motion as

$$\mathcal{L}_{past} = \|\hat{\mathbf{Y}}[1:H] - \tilde{\mathbf{X}}[1:H]\|_2^2 \,, \tag{3}$$

where $\hat{\mathbf{Y}}[1:H]$ and $\tilde{\mathbf{X}}[1:H]$ are the first H frames of $\hat{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$, respectively. Furthermore, to enforce the limb length of the past poses to be the same as those of the future poses, we include the limb length loss

$$\mathcal{L}_{limb} = \sum_{t=1}^{T}\sum_{i=1}^{J} \|\hat{l}_{t,i} - l_i\|_2^2 \,, \tag{4}$$

where $\hat{l}_{t,i}$ is the i-th limb length of t-th future pose, l_i is the ground truth of i-th limb length obtained from the history poses and J is the number of limbs. The loss weights of $(\mathcal{L}_{past}, \mathcal{L}_{limb})$ for Human3.6M and HumanEva-I are both $(100, 500)$.

The detailed architecture of our generator is shown in Fig. 2. It consists of 4 residual blocks. Each block comprises 2 graph convolutional layers. It also contains two additional layers, one at the beginning, to bring the input to feature space, and the other at the end, to decode the feature to the residuals of the DCT coefficients.

2. Implementation Details

We implemented our network using Pytorch [3] and used ADAM [2] to train it. The learning rate was set to 0.001 with a decay rate defined by the function

$$lr_decay = 1.0 - \frac{\max(0, e - 100)}{400} \,, \tag{5}$$

where e is the current epoch.

For Human3.6M, our model observes the past 25 frames to predict the future 100 frames, and we use the first 20 DCT coefficients. For HumanEva-I, our model predicts the

Figure 4: **Results of controllable motion prediction with $N = 5$.** In each row, we show the end poses of 10 samples predicted with the same motion for certain body parts. The controlled body parts are shown in gray.

future 60 frames given the past 15 frames, and we use the first 8 DCT coefficients.

To generate the pseudo ground truth for the multi-modal reconstruction error ($\mathcal{L}_{mm}$), we follow the official DLow [5] implementation of the multi-modal version of FDE (MMFDE) and use the distance between the last pose of the history to choose the pseudo ground truth. That is, for a training sample, any other training sequence with similar last pose in terms of Euclidean distance is chosen to be a possible future. We set the distance threshold to 0.05 for both Human3.6M and HumanEva-I.

3. Additional Qualitative Results

3.1. Controllable Motion Prediction with $N > 2$

In the main paper, we divided a human pose into 2 parts ($N = 2$): lower body and upper body, and show the results of predicting motions with same lower body motion but diverse upper body motions. Here, we provide qualitative results with $N = 5$. In particular, as shown in Fig. 3, we split a pose into: 1. torso, 2. right leg, 3. left leg, 4. right arm and 5. left arm, following this order for prediction. As shown in Fig. 4, our model is able to perform different levels of controllable motion prediction given the detailed body parts segmentation.

3.2. Ablation Study

We show a qualitative comparison on the results of our model without either the pose prior loss $\mathcal{L}_{nf}$ or the angle loss $\mathcal{L}_{ang}$ in Fig. 5. This clearly demonstrates the dramatic decrease in pose quality in both cases.

References

[1] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, jul 2014. 2

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2

[3] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 2

[4] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 2

[5] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, pages 346–364. Springer, 2020. 3

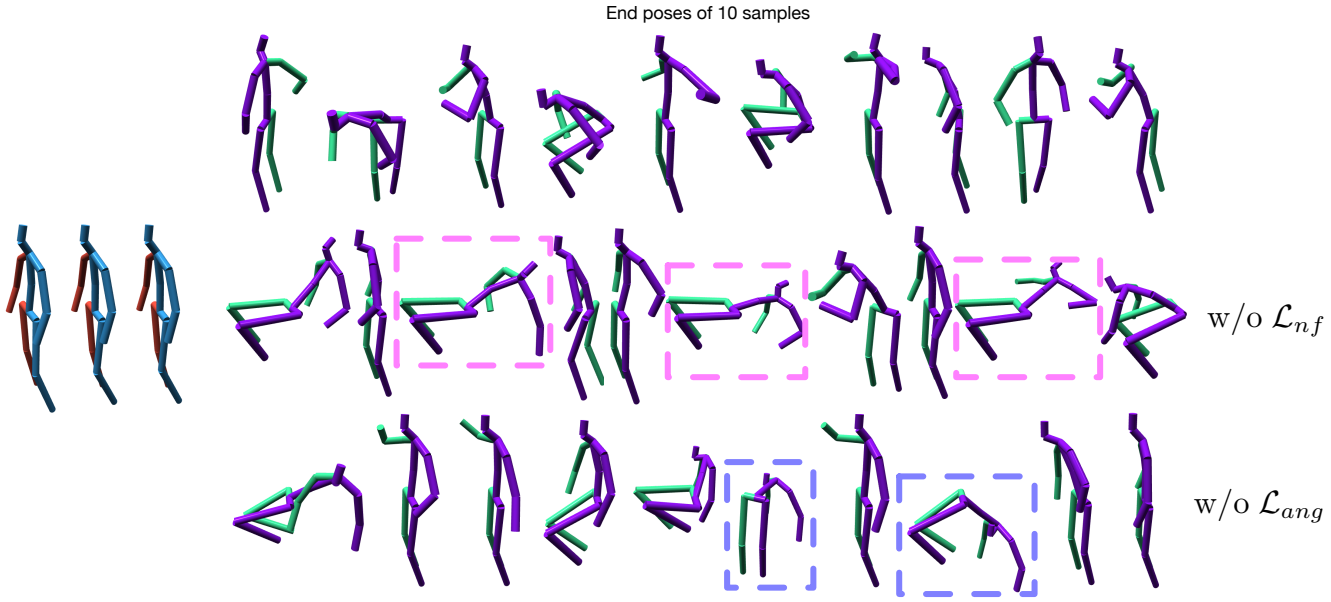

Figure 5: **Qualitative results of ablation study.** From top to bottom, we show the end pose of 10 samples of our model with all proposed losses, without the pose prior loss $\mathcal{L}_{nf}$ and without the angle loss $\mathcal{L}_{ang}$. Without the pose prior, the model predicts unlikely poses, as highlighted by the magenta boxes. When our model is trained without the angle loss, it produces invalid poses, highlighted by blue boxes.