DancingPose: A Simple Model for Generating Dancing Pose Sequence from Music

Yanyu Xu[1], Zhixin Piao[1], Peiyao Wang[1], Jiashi Feng[2], and Shenghua Gao[1]

[1] ShanghaiTech University, Shanghai, China
{xuyy2, piaozhx, wangpy, gaoshh}@shanghaitech.edu.cn
[2] National University of Singapore
elefjia@nus.edu.sg

The organization of our supplementary material is summarized as follows:

1 The implementation details about transferring 3D human skeletons into the 3D mesh.
2 The tatistics of music and dance dataset and our music dataset.

1 Implementation Details

In this section, we introduce the implementation details for transferring 3D human skeletons into 3D mesh used in user study and avatar renderings, such as the network architecture, loss function, training details, and some remarks.

Network Architecture. We use a simple model called SmplRegressor to regress SMPL[4] (a well known 3D parametric human body model) parameters θ from 3D human body key-points x, as shown in Fig. 1. The generator consists of the encoder, two residual blocks, and decoder. Encoder and decoder are both linear layers which have 1024 hidden dimension, 2 residual blocks stacked with linear layer(1024x1024), Batch Normalization layer, and RELU layer are used. Following HMR[1], we use the same discriminator, which includes two 1x1x32 Conv2d layers to encode SMPL parameters and two different modules to regress classification results (real or fake).

Loss Functions. We use 3d reconstruction loss L_{3d} and LSGAN[5] loss L_{adv} to regress better and more realistic results:

$$L_{3d} = \sum ||P_{3d}(\theta) - x||_2$$
$$L_{adv}^{G} = \sum (D(\theta) - 1)^2$$

Here, θ is SMPL parameter. P_{3d} is projection function which can render SMPL mesh by SMPL parameter θ and generate 3D human key-points. D is discriminator. The full objective function of generator is shown in the following:

$$L^{G} = \lambda_{3d} L_{3d} + \lambda_{adv} L_{adv}^{G}$$

, where λ_{3d} and λ_{adv} are hyper-parameters. In discriminator, the loss function is:

$$L_{adv}^{D} = \sum D(\theta)^2 + \sum (D(\theta') - 1)^2$$

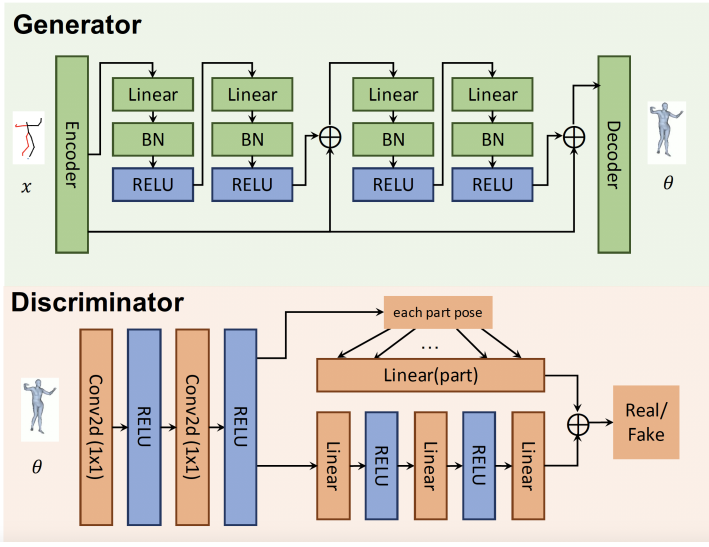

Fig. 1. SmplRegressor Model

, where θ is SMPL parameter predicted by generator and θ' is random sampled from nature dataset.

Training Details. To train this network, all 3D key-points are normalized to [-1, 1] and the center of mass is moved to $(0, 0, 0)$. We sample 4k key-points from music & dance dataset [6] and use Mosh dataset[3] as corresponding nature dataset in the discriminator. The mini-batch is 256 in our experiments. λ_{3d} and λ_{adv} are set to 1.0 and 0.07 respectively. Adam[2] is used for parameter optimization of both generator and discriminator.

2 Statistics of music and dance dataset and our music dataset

Table 1 shows the statistics of training and test set of music & dance (M&D) dataset [6] and our music dataset. Since only the original files are provided without training and testing splits in [6], we randomly select 51 music-dance pairs for training and use the rest for testing as shown in Table 1. Due to its limited data volume and to evaluate model generalization ability, we also collect 88 music to form a new evaluation dataset as shown in Table 1.

References

1. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (2018)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

Table 1. Statistics of training and test set of music & dance (M&D) dataset [6] and our collected dataset.

	M&D Training		M&D Test		Ours	
	# Files	Duration (s)	# Files	Duration (s)	# Files	Duration
Cha-cha	6	546.0	2	271.0	22	4741.3
Rumba	8	700.0	2	138.0	22	4386.9
Tango	7	1441.0	2	524.0	22	3752.8
Waltz	30	377.4	4	55.24	22	4421.2
Total	51	3064.4	10	988.24	88	17302.2

3. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) **33**(6), 220 (2014)
4. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 248 (2015)
5. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2017)
6. Tang, T., Jia, J., Mao, H.: Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In: 2018 ACM Multimedia Conference on Multimedia Conference. pp. 1598–1606. ACM (2018)