Spatial-Temporal Parallel Transformer for Arm-Hand Dynamic Estimation (Supplementary Material)

Shuying Liu, Wenbin Wu, Jiaxian Wu, Yue Lin
NetEase Games AI Lab, Guangzhou, China
{liushuying, wuwenbin02, wujiaxian, gzlinyue}@corp.netease.com

Table 1. Network architecture of 2D hand key-point estimator. Each line represents a group of identical layers, repeating n times. All layers in the same group have the same number of output channels c. The first layer of each group has a stride s. And t is the expansion factor of the bottleneck.

Input	Operator	t	c	n	s
224*224*3	Conv3x3	-	32	1	2
112*112*32	DWConv3x3	-	32	1	1
112*112*32	Bottleneck	2	32	4	2
56*56*32	Bottleneck	2	64	5	2
28*28*64	Bottleneck	4	128	14	2
14*14*128	Conv1x1	-	21	1	1

1. Architecture of 2D Hand Key-points Estimator

The architecture of 2D hand key-point estimator is searched using SPOS-NAS [2], Tabel 1 shows the detailed configuration.

2. Dataset

Fig 1 shows sample frames of our motion capture dataset. We retarget the mocap data to the MIXAMO ybot character [1]. Fig 2 shows sample frames of the rendered dataset. We choose 3 human-look characters [1](Leonard, Stefani, Pete) and some in-the-wild background pictures to simulate in-the-wild videos of human motions.

3. Architecture Parameter Analysis

We investigate different parameters of network architecture to search for a optimal setting. Experiments involve parameters of t_depth, t_head and s_depth, and results can be found in Table 2. The best combination of these parameters for our task is t_depth=6, t_head=8, s_depth=2.

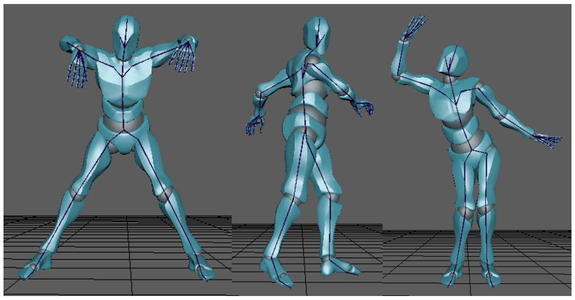

Figure 1. Samples of our motion capture dataset.

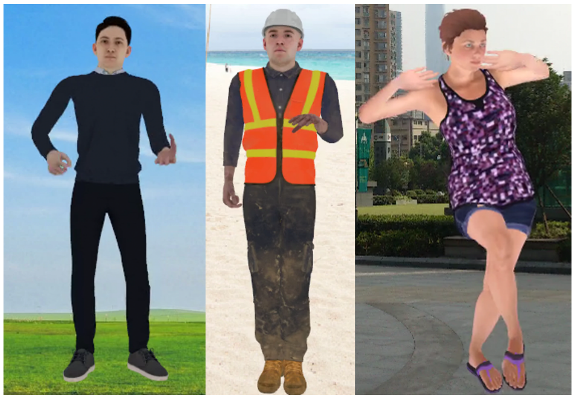

Figure 2. Samples of our rendered dataset.

References

[1] Mixamo. animate 3d characters for games, film, and more. http://https://www.mixamo.com. Accessed: 2021-09-30. 1

[2] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *European Conference on Computer Vision*, pages 544–560. Springer, 2020. 1

Table 2. Ablation study on different model parameters in constructing PAHMT. Evaluation is conducted on our motion capture dataset and MPJPE is reported. t_depth and t_head is the layer number and multi-head self-attention head number of the temporal transformer encoder; s_depth is the layer number of spatial transformer encoder.

t_depth	t_head	s_depth	MPJPE↓
6	8	2	**0.0274**
4	8	2	0.0295
8	8	2	0.0486
6	4	2	0.0333
6	16	2	0.0305
6	8	1	0.0338