Hierarchical Style-based Networks for Motion Synthesis

Supplementary Material

1 Architecture details of short-range generation model.

Model Architecture. Both two encoders consist of 1D convolution layers operating along the temporal axis to capture motion information. The content encoder consists of two 1D convolution layers. The first layer is responsible for mapping motion vector to high dimensional space, while the second layer produces content feature. The style encoder consists of four 1D convolution layers. The activation function is ReLU for intermediate layers and that of the final layer is Tanh.

Design Consideration. Note that the motion content changes along with the specific state at each time step, while the style is kept relatively constant throughout the subsequence. Therefore, to capture instantly changing content feature the kernel width of the two convolutional layers in content encoder are 1 and 3 respectively, i.e., expanding a temporal window with length of 3. As for the style encoder (4 layers), the kernel width are 1, 3, 3 and 5 respectively, i.e., the overall receptive field is 45. With common input frame-rate at 30 fps the style encoder observes about 1.5s of real-world motion, which is sufficient to capture the style feature of one motion sequence.