Dance In the Wild: Monocular Human Animation with Neural Dynamic Appearance Synthesis (Supplementary material)

Tuanfeng Y. Wang, Duygu Ceylan, Krishna Kumar Singh
Adobe Research
{yangtwan,ceylan,krishsin}@adobe.com

Niloy J. Mitra
Adobe Research, University College London
nimitra@adobe.com

1. Network architecture

Our network consists of five sub-modules: 1) pose encoder $E_P$; 2) motion encoder $E_M$; 3) pose refinement network $E_{Refine}$; 4) style generator $T$; and 5) discriminator $D$. The basic building block of each of the modules is a convolutional layer $\text{Conv}^{\text{out},k,s}(x)$, where $x$ is the input tensor, $\text{out}$ is the number of output channels, $k$ is the kernel size, $s$ is the stride. The convolutional layer is followed by a LeakyRelu activation by default. We further define a residual block:

$$\text{Res}^{\text{out}}(x) = \frac{1}{\sqrt{2}} (\text{Conv}^{\text{out},3,2}(\text{Conv}^{\text{in},3,1}(x)) + \text{Conv}^{\text{out},1,2}(x)),$$

where $\text{in}$ is the number of input channels from $x$. For $T$ and $D$, we adopt the exact architecture from StyleGAN2. Please refer to [4] for more details.

**Pose encoder $E_P$.** Starting from a pose signature $P_i(6 \times 512 \times 512)$, we apply $\text{Conv}^{32,3,1}(\cdot) \rightarrow \text{Res}^{64}(\cdot) \rightarrow \text{Res}^{128}(\cdot) \rightarrow \text{Res}^{256}(\cdot) \rightarrow \text{Res}^{512}(\cdot)$ to generate our pose feature $P_i(512 \times 32 \times 32)$.

**Motion encoder $E_M$.** Similar to $E_P$, starting from a motion signature $M_i(60 \times 512 \times 512)$, we apply $\text{Conv}^{64,3,1}(\cdot) \rightarrow \text{Res}^{256}(\cdot) \rightarrow \text{Res}^{512}(\cdot) \rightarrow \text{Res}^{1024}(\cdot) \rightarrow \text{Res}^{2048}(\cdot)$ to generate our motion feature $M_i(2048 \times 32 \times 32)$.

**Pose refinement network $E_{Refine}$.** We first concatenate $P_i$ and $M_i$ to get $P^{\text{init}}(2560 \times 32 \times 32)$. We then apply $\text{Conv}^{1024,3,1}(\cdot) \rightarrow \text{Conv}^{512,3,1}(\cdot) \rightarrow \text{Conv}^{512,3,1}(\cdot)$ to produce $P_i(512 \times 32 \times 32)$.

2. Comparison with Savitzky-Golay filter [5]

Inspired by the previous efforts on pose stabilization in motion capture literature (e.g., [3]), we apply the Savitzky–Golay filter [5] on the learned pose feature sequence to produce ghost effects around thin structures, i.e., hands.

<table>
<thead>
<tr>
<th>$P_i$</th>
<th>MSE $\downarrow$</th>
<th>SSIM $\uparrow$</th>
<th>LPIPS $\downarrow$</th>
<th>FID $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w/$ Savitzky–Golay filter</td>
<td>0.0377</td>
<td>0.7909</td>
<td>0.1014</td>
<td>50.5325</td>
</tr>
<tr>
<td>Ours</td>
<td>0.0347</td>
<td>0.9724</td>
<td>0.0913</td>
<td>48.8647</td>
</tr>
</tbody>
</table>

Table 1. We conduct quantitative evaluation around the hand region of each frame.
3. Shape prior from dense body UV prediction.

Since our method starts from the dense body UV map generated by Densepose [2] detector, the difference between the shape of the bare body and the shape of the dressed body may affect the performance of the proposed method, especially when the target pose is not covered in the train set. In this experiment, we show an example with the extreme pose that is very different from the dancing sequence used to train our network in Figure 3. We observe that our approach can still synthesis a reasonable appearance of the character for the target extreme pose, but such shape prior of the dense body UV prevents us from generating more plausible details. This indicates the exploration of clothes-aware dense UV (i.e., [6]) can be a promising direction for future work.

References


