Supplementary Material of Diff3DHPE: A Diffusion Model for 3D Human Pose Estimation

Jieming Zhou[1], Tong Zhang[2], Zeeshan Hayder[3], Lars Petersson[3], Mehrtash Harandi[4]
[1]Australian National University, [2]EPFL, [3]CSIRO, [4]Monash University
jieming.zhou@anu.edu.au, tong.zhang@epfl.ch,
{zeeshan.hayder, Lars.Petersson}@data61.csiro.au, mehrtash.harandi@monash.edu

1. Proof of Iteration Steps Required by DDIM

The reverse diffusion process proposed by DDIM [2] is:

$$\hat{\boldsymbol{y}}_{\tau_{i-1}} = \sqrt{\bar{\alpha}_{\tau_{i-1}}}\left(\frac{\hat{\boldsymbol{y}}_{\tau_i} - \sqrt{1-\bar{\alpha}_{\tau_i}}\hat{\boldsymbol{\epsilon}}_{\tau_i}}{\sqrt{\bar{\alpha}_{\tau_i}}}\right) + \sqrt{1-\bar{\alpha}_{\tau_{i-1}}}\hat{\boldsymbol{\epsilon}}_{\tau_i}, \tag{1}$$

$$\hat{\boldsymbol{y}}_0 = \frac{\hat{\boldsymbol{y}}_{\tau_1} - \sqrt{1-\bar{\alpha}_{\tau_1}}\hat{\boldsymbol{\epsilon}}_{\tau_1}}{\sqrt{\bar{\alpha}_{\tau_1}}}, \tag{2}$$

where τ_i is sampled every $\lceil T/S \rceil$ steps from $\{t_1, t_2, ..., t_T\}$, $\tau_1 < \tau_2 < ... < \tau_S \in [1, T]$, $S < T$, $\hat{\boldsymbol{y}}_t$ is the estimated 3D coordinates at step t, $\hat{\boldsymbol{y}}_{\tau_S} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $\bar{\alpha}_t$ is a predefined noise schedule. In this paper, we select cos schedule for $\bar{\alpha}_t$ proposed by [1]:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, f(t) = cos\left(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2}\right)^2, s = 0.008. \tag{3}$$

We assume the 3D coordinate value of a human joint is between $[-1000, 1000]$ mm after centralizing the body. Then, we normalize the coordinate value to $[-1, 1]$, which is required by the diffusion model. Since $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we have 95% probability that $|\epsilon| < 2\sigma = 2$. σ is the standard deviation of $\boldsymbol{\epsilon}$. Therefore, we shall have

$$\frac{\sqrt{1-\bar{\alpha}_{\tau_1}}}{\sqrt{\bar{\alpha}_{\tau_1}}} < 10^{-3}\cdot\frac{1}{|\epsilon|} = 5\times 10^{-4} \tag{4}$$

in Eq. 2, which ensures impact introduced by noise value to the final prediction has 95% probability smaller than 1 mm. To achieve this, the minimum $\tau_1 = 1$. Then, we derive

$$\bar{\alpha}_1 = \frac{cos\left(\frac{1/T+s}{1+s}\cdot\frac{\pi}{2}\right)^2}{cos\left(\frac{s}{1+s}\cdot\frac{\pi}{2}\right)^2} > \frac{1}{1+(5\times 10^{-4})^2}, \tag{5}$$

$$\frac{cos\left(\frac{1/T+s}{1+s}\cdot\frac{\pi}{2}\right)}{cos\left(\frac{s}{1+s}\cdot\frac{\pi}{2}\right)} > \sqrt{\frac{1}{1+(5\times 10^{-4})^2}} \tag{6}$$

from Eq. 4. According to small-angle approximations, we can have

$$\frac{1-\frac{(\frac{1/T+s}{1+s}\cdot\frac{\pi}{2})^2}{2}}{1-\frac{(\frac{s}{1+s}\cdot\frac{\pi}{2})^2}{2}} > \sqrt{\frac{1}{1+(5\times 10^{-4})^2}}, \tag{7}$$

when $T \gg 1$ and $s = 0.008$. Thus, we obtain:

$$T > \frac{1}{\frac{2(1+s)}{\pi}\sqrt{2-\sqrt{\frac{1}{1+(5\times 10^{-4})^2}}\left(2-(\frac{s}{1+s}\cdot\frac{\pi}{2})^2\right)}-s} \approx 1.55\times 10^5, \tag{8}$$

which can meet the target.

2. Hyper-parameter settings

The hyper-parameter search space and the final choice in our experiments are listed in Table 1 and 2.

Table 1. Hyper-parameter search space.lr: learning rate. $StepEmb$: whether or not using step embedding. S: the number of reverse diffusion steps.

Param.	Search Space
lr	1E-4,4E-4,1E-3,4E-3
$StepEmb$	T, F
S	1,3,4,5,6,7,8,10,15,20,40,80,160,320

References

[1] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

Table 2. Final hyper-parameters of each model. Diff3DHPE-M: Diff3DHPE with MixSTE backbone. Diff3DHPE-P: Diff3DHPE with PoseFormer backbone. DDIM-M: Diffusion model with DDIM reverse diffusion method and MixSTE backbone. $*$: we train the baselines with only L2 loss of 3D pose prediction error and normalize the training target 3D pose ground truth to $[-1, 1]$. F: the number of frames. bs: batch size. lr: learning rate. dim: embedding dimension. $depth$: the number of Transformer blocks. $StepEmb$: whether or not using step embedding. S: the number of reverse diffusion steps. dr: dropout rate. wd: weight decay. lrd: learning rate decay factor.

<table>
<tr><th>Model</th><th>Dataset</th><th>F</th><th>bs</th><th>lr</th><th>dim</th><th>$depth$</th><th>$StepEmb$</th><th>S</th><th>dr</th><th>wd</th><th>lrd</th></tr>
<tr><td>Diff3DHPE-P</td><td>H3.6M CPN</td><td>81</td><td>1024</td><td>4E-3</td><td rowspan="4">32</td><td rowspan="4">8</td><td rowspan="2">T</td><td>5</td><td rowspan="17">0.1</td><td rowspan="17">0.1</td><td rowspan="17">0.1</td></tr>
<tr><td>Diff3DHPE-P</td><td>H3.6M GT</td><td>81</td><td>1024</td><td>4E-3</td><td>5</td></tr>
<tr><td>PoseFormer</td><td>H3.6M CPN</td><td>81</td><td>1024</td><td>1e-4</td><td rowspan="2">N/A</td><td rowspan="2">N/A</td></tr>
<tr><td>PoseFormer</td><td>H3.6M GT</td><td>81</td><td>1024</td><td>1e-4</td></tr>
<tr><td>Diff3DHPE-M</td><td>H3.6M CPN</td><td>81</td><td>64</td><td>4E-4</td><td rowspan="13">512</td><td rowspan="13">16</td><td rowspan="8">T</td><td>9</td></tr>
<tr><td>Diff3DHPE-M</td><td>H3.6M CPN</td><td>243</td><td>24</td><td>4E-4</td><td>5</td></tr>
<tr><td>Diff3DHPE-M</td><td>H3.6M GT</td><td>81</td><td>64</td><td>4E-4</td><td>5</td></tr>
<tr><td>Diff3DHPE-M</td><td>H3.6M GT</td><td>243</td><td>24</td><td>4E-4</td><td>6</td></tr>
<tr><td>Diff3DHPE-M w/o PDE</td><td>H3.6M CPN</td><td>81</td><td>64</td><td>1E-4</td><td>6</td></tr>
<tr><td>Diff3DHPE-M w/o PDE</td><td>H3.6M CPN</td><td>243</td><td>24</td><td>1E-4</td><td>5</td></tr>
<tr><td>DDIM-M</td><td>H3.6M CPN</td><td>81</td><td>64</td><td>4E-4</td><td>40</td></tr>
<tr><td>DDIM-M</td><td>H3.6M CPN</td><td>243</td><td>24</td><td>4E-4</td><td>80</td></tr>
<tr><td>MixSTE</td><td>H3.6M CPN</td><td>81</td><td>64</td><td>1E-4</td><td rowspan="4">N/A</td><td rowspan="4">N/A</td></tr>
<tr><td>MixSTE</td><td>H3.6M CPN</td><td>243</td><td>24</td><td>1E-4</td></tr>
<tr><td>MixSTE</td><td>H3.6M GT</td><td>81</td><td>64</td><td>1E-4</td></tr>
<tr><td>MixSTE</td><td>H3.6M GT</td><td>243</td><td>24</td><td>1E-4</td></tr>
<tr><td>Diff3DHPE-M</td><td>3DHP GT</td><td>27</td><td>64</td><td>4E-4</td><td>F</td><td>7</td></tr>
</table>