1. Additional Details of GMM Forward Diffusion

In Section 4.2 of the main paper, we describe the GMM-based forward diffusion process. Here, we explain it in more detail, particularly about how it can be framed in a step-wise formulation. We first re-state Eq. 7 in the main paper as follows:

$$h_k = \mu^G + \sqrt{\alpha_k}(h_0 - \mu^G) + \sqrt{(1 - \alpha_k)} \cdot \epsilon^G.$$  (1)

where $\mu^G = \sum_{m=1}^{M} 1_m \mu_m$, $\epsilon^G \sim \mathcal{N}(0, \sum_{m=1}^{M} (1_m \Sigma_m))$, and $1_m \in \{0, 1\}$ is a binary indicator for the $m_{th}$ component such that $\sum_{m=1}^{M} 1_m = 1$, and Prob$(1_m = 1) = \pi_m$.

We remark that Eq. 1 directly formulates $\tilde{h}_k$ as a function of $h_0$ instead of $h_{k-1}$, because this clearly expresses the aim of our GMM-based forward diffusion design, i.e., such that the generated $\tilde{h}_1, ..., \tilde{h}_K$ can converge to the fitted GMM model $\phi_{GMM}$. Yet, we note that the step-wise formulation of $\tilde{h}_k$ in terms of $\tilde{h}_{k-1}$ can still be defined, if necessary. First, we sample according to probabilities $\{\pi_m\}_{m=1}^{M}$, and select a Gaussian component $\tilde{m}$, i.e., $\tilde{h}_k = 1$. Next, we first calculate $\tilde{h}_0$, a “centered” version of $h_0$, using $\tilde{h}_0 = h_0 - \mu^G$, where $\mu^G = \sum_{m=1}^{M} (1_m \mu_m) = \mu_{\tilde{m}}$. Then, we follow the step-wise formulation as follows:

$$\tilde{h}_k = \sqrt{\alpha_k} \cdot \tilde{h}_{k-1} + \sqrt{(1 - \alpha_k)} \cdot \epsilon^G,$$  (2)

where $\epsilon^G \sim \mathcal{N}(0, \sum_{m=1}^{M} (1_m \Sigma_m))$, which is equivalent to $\epsilon^G \sim \mathcal{N}(0, \Sigma_m)$. After taking $k$ steps of Eq. 2 starting from $\tilde{h}_0$, we can get:

$$\tilde{h}_k = \sqrt{\alpha_k}(h_0) + \sqrt{(1 - \alpha_k)} \cdot \epsilon^G.$$  (3)

We observe that the result of the stepwise formulation is thus equivalent to Eq. 1, as we can simply “de-center” our $\hat{h}_0$ and $\tilde{h}_k$ by substituting $h_0 = h_0 - \mu^G$ and $\tilde{h}_k = \tilde{h}_k - \mu^G$.

2. Additional Details of Diffusion Network $g$

In order to provide information to the model regarding the current step number $k$, we generate a diffusion step embedding $f_D^k \in \mathbb{R}^{J \times 256}$ using the sinusoidal function. Specifically, at each even $(2j)$ index of $f_D^k$, we set the element $f_D^k[2j]$ to $\sin(k/10000^{(2j)/256})$, while at each odd $(2j + 1)$ index, we set the element $f_D^k[2j + 1]$ to $\cos(k/10000^{(2j)/256})$.

3. More Implementation Details

In the forward diffusion process, we generate the decreasing sequence $\alpha_{1:k}$ via the formula: $\alpha_k = \prod_{i=1}^{k} (1 - \beta_i)$, where $\beta_{i:k}$ is a sequence from $1e-4$ to $2e-3$, which is interpolated by the linear function. To optimize the GMM parameters $\phi_{GMM}$, we sample 1000 poses from $\hat{H}_K$ (i.e., $N_{GMM} = 1000$) and then model $H_K$ via a GMM model.

During model pre-training, the Context Encoder $\phi_{ST}$ is first pre-trained on the training set to predict 3D poses from 2D poses. Then we adopt the Adam optimizer [7] to train our diffusion model $g$, where the initial learning rate is set to $1e-4$ with a decay rate of 0.9 after ten epochs, and the batch size is set to 4096. Our DiffPose is implemented using PyTorch, and can be trained on a single GeForce RTX 3090 GPU within 96 hours.

4. Experiment Results on Human3.6M under P-MPJPE (Protocol 2)

Tab. 1 and Tab. 2 present the video-based and frame-based results of our DiffPose on Human3.6M under P-MPJPE, where the input 2D poses are detected by CPN [1]. As shown in Tab. 1, our DiffPose can significantly outperform the state-of-the-art methods [8, 21] on all actions with a large margin. Moreover, from Tab. 2, we observe that our method can achieve promising performance on the challenging frame-based setting.
Table 1. Video-based results on Human3.6M with detected 2D poses in millimeters under P-MPJPE.

<table>
<thead>
<tr>
<th>P-MPJPE</th>
<th>Dir. Disc. Eat Greet Phone Photo Pose Put Sit SitD. Smoke Wait WalkD. Walk WalkT.</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin [9]</td>
<td>32.5 35.3 34.3 36.2 37.8 43.0 33.0 32.2 45.7 51.8 38.4 32.8 37.5 25.8 28.9</td>
<td>36.8</td>
</tr>
<tr>
<td>Pavlo [14]</td>
<td>34.1 36.1 34.4 37.2 36.4 42.2 34.4 33.6 45.0 52.5 37.4 33.8 37.8 25.6 27.3</td>
<td>36.5</td>
</tr>
<tr>
<td>Liu [12]</td>
<td>32.3 35.2 33.3 35.8 35.9 41.5 33.1 32.1 32.0 42.8 48.5 34.8 32.4 35.3 24.5 26.0</td>
<td>34.8</td>
</tr>
<tr>
<td>Zheng et al. [24]</td>
<td>32.5 34.8 32.6 34.6 35.3 39.5 32.1 32.0 42.8 48.5 34.8 32.4 35.3 24.5 26.0</td>
<td>34.6</td>
</tr>
<tr>
<td>Li [8]</td>
<td>31.5 34.9 32.8 33.6 35.3 39.6 32.0 32.2 43.5 48.7 36.4 32.6 34.3 23.9 25.1</td>
<td>34.4</td>
</tr>
<tr>
<td>Zhang [21]</td>
<td>28.0 30.9 28.6 30.7 30.4 34.6 28.6 28.1 37.1 47.3 36.5 32.7 30.5 23.6 20.0</td>
<td>30.4</td>
</tr>
<tr>
<td>Ours</td>
<td>26.3 29.0 26.1 27.8 28.4 34.6 26.9 26.5 36.8 39.2 29.4 26.8 28.4 18.6 19.2</td>
<td>28.7</td>
</tr>
</tbody>
</table>

Table 2. Frame-based results on Human3.6M with detected 2D poses in millimeters under P-MPJPE.

<table>
<thead>
<tr>
<th>P-MPJPE</th>
<th>Dir. Disc. Eat Greet Phone Photo Pose Put Sit SitD. Smoke Wait WalkD. Walk WalkT.</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun [17]</td>
<td>42.1 44.3 45.0 45.4 51.5 53.0 43.2 41.3 59.3 73.3 51.0 44.0 48.0 38.3 44.8</td>
<td>48.3</td>
</tr>
<tr>
<td>Martinez [13]</td>
<td>39.5 43.2 46.4 47.0 51.0 56.0 41.4 40.6 56.5 69.4 49.2 45.0 49.5 38.0 43.1</td>
<td>47.7</td>
</tr>
<tr>
<td>Pavlakos [14]</td>
<td>34.7 39.8 41.8 38.6 42.5 47.5 38.0 36.6 50.7 56.8 42.6 39.6 43.9 32.1 36.5</td>
<td>41.8</td>
</tr>
<tr>
<td>Liu [11]</td>
<td>35.9 40.7 38.0 41.5 42.3 47.5 37.8 36.0 46.6 56.6 41.8 38.3 43.7 31.7 26.2</td>
<td>41.2</td>
</tr>
<tr>
<td>Ours</td>
<td>33.9 38.2 30.9 32.1 39.2 40.2 38.3 38.0 40.9 46.9 32.5 41.2 39.7 30.9 30.5</td>
<td>35.8</td>
</tr>
</tbody>
</table>

5. Additional Results

In this section, we further investigate the performance of our method on the frame-based scenario, by conducting experiments on Human3.6M [6].

3D Pose visualization. First, we qualitatively compare our method with state-of-the-art method [22] in this setting, and present results in Fig. 1. We observe that our method can predict more reliable and accurate poses, especially for novel human gestures (e.g., the first and second rows in Fig. 1) and occluded body parts (e.g., the third and fourth rows in Fig. 1).

Forward diffusion process visualization. Extending from our results in Tab. 5 of the main paper, here we qualitatively compare our GMM-based forward diffusion process with the standard diffusion process (as described in Sec. 3 of our main paper). As shown in Fig. 2, the standard diffusion process recurrently adds noise to the source sample and tends to spread the joints’ positions to the whole space. However, our GMM-based diffusion process can add noise according to pose-specific information (obtained from heatmaps) and the data distribution, which generates noise in a more constrained manner. Thus, during training, the GMM-based diffusion process allows us to initialize a $H_K$ that captures the uncertainty of the 3D pose, which boosts the performance of DiffPose.

Reverse diffusion process visualization. We visualize the poses reconstructed by our diffusion model with/without the context information $f_{ST}$. Note that the model without $f_{ST}$ means that no context decoder is used. From the last column of Fig. 3, we observe that both methods can reconstruct realistic human poses while the model with $f_{ST}$ can predict more accurate poses. Moreover, compared to the unconditioned reverse diffusion process (i.e., the model without $f_{ST}$), the model conditioned by $f_{ST}$ can converge to the desired pose faster.

6. Future Work

In this work, we explore a novel diffusion-based framework to tackle monocular 3D pose estimation. Future work includes more investigations into the architecture of the diffusion network, as well as extending to the online setting [2, 5, 19], the few-shot setting [18, 23] and other pose-based tasks [3, 4, 10, 16, 20].
Figure 2. Qualitative comparison between standard diffusion forward process and our GMM-based forward diffusion process.

Figure 3. Qualitative comparison between our reverse diffusion process conditioned on context information $f_{ST}$ (bottom), against a reverse diffusion process without using $f_{ST}$ (top).

References


