3DHumanGAN: 3D-Aware Human Image Generation with 3D Pose Mapping

Supplementary Material

Zhuoqian Yang[1,2†] Shikai Li[1] Wayne Wu[1 ✉] Bo Dai[1]
[1] Shanghai AI Laboratory [2] School of Computer and Communication Sciences, EPFL
zhuoqian.yang@epfl.ch lishikai@pjlab.org.cn
wuwenyan0503@gmail.com daibo@pjlab.org.cn

Abstract

This document provides supplementary information which is not elaborated in our manuscript due to space limits. Section 1 discusses the relation between consistency and equivariance and the reasons for using spatial adpative batch normalization from a theortical angle. Section 2 presents qualitative results for the ablation study. Section 3 presents additional qualitative results for our method and comparison against prior art. We also present a video demo which includes a brief introduction of our method and animated qualitative results.

1. Consistency and Equivariance

We have mentioned in the main text that equivariance is essential for producing consistent outcomes. Here we elaborate on this point. We employ a convolutional backbone that consists only of 1×1 convolutions. The backbone takes a 2D grid of image coordinates as input. In this setting, we hypothesize that the 3D pose mapping network performs certain non-linear spatial transformations on the 2D coordinates based on the SMPL mesh that conditions it. To ensure that these transformations are reflected in the generated image, we need a CNN backbone that is equivariant. Mathematically, a function is equivariant if applying a transformation to its input leads to the same result as applying the transformation to its output. We think this is vital for the network to learn features that are not tied to absolute coordinates but follow the coordinates transformed under 3D guidance.

In the main text we compared between spatial adaptive instance normalization (SAIN) and spatial adaptive batch normalization (SABN) for injecting the style maps rendered by the 3D pose mapping network into the 2D backbone. Both of these operations contain a normalization step and a denormalization step. Here we inspect the normalization step in detail. The former uses instance statistics for this step while the latter uses batch statistics. We will use the following notation: $\mathbf{x}$ is an input vector of size n, $\mathbf{y}$ is an output vector of size n, $\mathbf{W}$ is a weight matrix of size $m \times n$, and $\mathbf{b}$ is a bias vector of size m.

First, let us consider instance normalization. For each instance $\mathbf{x}_i$ in the batch, we have

$$\mathbf{y}_i = \frac{\mathbf{W}\mathbf{x}_i + \mathbf{b} - \boldsymbol{mu}_i}{\boldsymbol{\sigma}_i},$$

where $\boldsymbol{mu}_i$ and $\boldsymbol{\sigma}_i$ are the mean and standard deviation of $\mathbf{W}\mathbf{x}_i + \mathbf{b}$. This means that IN maps all instances to have zero mean and unit variance. However, this also means that IN discards some information about the relative magnitude and scale of each instance. When $\mathbf{x}$ store coordinates, instance normalization effectively scales and shifts $\mathbf{x}$ which makes the network unequivariant.

Next, let us consider normalization with batch statistics. For a batch of size B, we have

$$\mathbf{y} = \frac{\mathbf{W}\mathbf{x} + \mathbf{b} - \boldsymbol{\mu}}{\boldsymbol{\sigma}},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and standard deviation vectors across the batch. When the batch size approaches infinity, we have

$$\lim_{B\to\infty} \boldsymbol{\mu} = \mathbb{E}[\mathbf{W}\mathbf{x} + \mathbf{b}],$$
$$\lim_{B\to\infty} \boldsymbol{\sigma}^2 = \mathrm{diag}(\mathbb{V}[\mathbf{W}\mathbf{x} + \mathbf{b}]),$$

where $\mathbb{E}[\cdot]$ denotes expectation and $\mathbb{V}[\cdot]$ denotes covariance. These limits are independent of any particular instance in the batch. Assuming that the entries in $\mathbf{x}$ are independent

† Work donw as research engineer at Shanghai AI Laboratory.

and normally distributed, *i.e.* $\mathbf{x}_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$,

$$\begin{aligned}\lim_{B\to\infty} \boldsymbol{\mu} &= \mathbb{E}[\mathbf{W}] \cdot \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{b}] \\ &= \mu_x \mathbf{W} \cdot \mathbf{1} + \mathbf{b}, \\ \lim_{B\to\infty} \boldsymbol{\sigma}^2 &= \text{diag}(\mathbb{V}[\mathbf{W}\mathbf{x}]) \\ &= \text{diag}(\mathbf{W}^T \mathbb{V}[\mathbf{x}] \mathbf{W}) \\ &= \sigma_x^2 \text{diag}(\mathbf{W}^T \mathbf{W})\end{aligned}$$

This normalization maps all instances to have the same mean and variance vectors as the whole batch. However, this does not mean that it discards any information about the relative magnitude and scale of each instance, since these limits only depend on the global statistics of $\mathbf{x}$. When $\mu_x = 0$ and $\sigma_x = 1$ the normalization becomes equivariant.

The reasoning above provide mathematical intuitions in preferring SABN over SAIN for better equivariance. However, the equivatiance of SABN still depends on several assumptions. Devloping an operation that do not rely on these assumptions could be a promising direction for future work.

2. Ablation Study (Qualitative Results)

We present qualitative results of the ablation study in Fig. 1. Note that all ablation models are trained at 256×128 resolution due to limited computational resources. Replacing the segmentation-based GAN loss with traditional binary GAN loss causes the model to lose the ability of pose conditioning, as shown by the result of the *VAE* configuration. The model that combines 2D and 3D networks in a feed-forward manner is functional, but less desirable in image quality and consistency. Using an upsampling convolutional backbone instead of pixel-wise independent one results in impaired consistency. When passing the 3D style maps into the 2D backbone, using instance normalization instead of batch normalization has similar effects.

3. Additional Qualitative Results

We show additional qualitative comparison results in Fig. 2. For our method, we provide additional qualitative results in Fig. 3, appearance interpolation results in Fig. 4 and pose interpolation results in Fig. 5. We also provide animated results in the video demo.

Figure 1: **Ablation Study Qualitative Results.** We show four cases separated by dotted lines. For each case we show one identity in two poses and three view-angles. The conditioning mesh is shown on the left of each case.

Figure 2: **Additional Qualitative Comparison.** We show four cases separated by the dotted line. For each case, the first row shows unconditional generation results, the second row shows pose-conditional generation results. We show three view angles, from -30° to 30°

Figure 3: **Additional Qualitative Results.** Each row shows two cases separated by the dotted line. For each case we show one identity in two poses and three view-angles. The conditioning mesh is shown on the left of each case.

Figure 4: **Additional Appearance Interpolation Results.** Each row shows two cases separated by the dotted line.

Figure 5: **Additional Pose Interpolation Results.** Each row shows two cases separated by the dotted line.