Supplemental Materials of "BodyGAN: General-purpose Controllable Neural Human Body Generation"

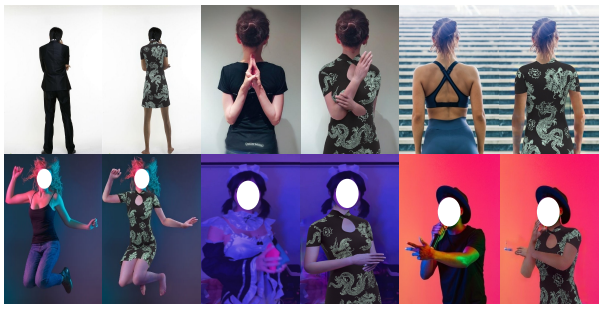
Chaojie Yang[1,*], Hanhui Li[2,*], Shengjie Wu[1], Shengkai Zhang[1], Haonan Yan[1],
Nianhong Jiao[1], Jie Tang[1], Runnan Zhou[1], Xiaodan Liang[2,†], Tianxiang Zheng[1,†]
[1]Beijing Momo Technology Co., Ltd. [2]Shenzhen Campus of Sun Yat-sen University
{1360546528,1421901449}@qq.com, {wu.shengjie24,double4tar,xdliang328,zhengtianxiang1128}@gmail.com,
lihanhui@mail3.sysu.edu.cn, songkey@pku.edu.cn, jnhrhythm@tju.edu.cn,chinatszrn@163.com

(a) (b) (c) (d)

Figure 1. Visual comparison with VITON methods. (a) Original images. (b) Synthesized images generated by CP-VTON-PLUS [2]. (c) Synthesized images generated by PF-AFN [1]. (d) Synthesized images generated by the BodyGAN.

1. Network Structure

The proposed BodyGAN consists of a pose encoding branch, an appearance encoding branch, and a generator. The pose encoding branch and the appearance encoding branch are responsible for condition map generation. The three subnetworks in the pose encoding branch can be replaced with state-of-the-art semantic segmentation networks, image to 3D surface networks, and key point estimation networks. The generator is further composed of two encoders and one decoder. The two encoders are designed for extracting pose and appearance features, which follow the common practice of convolutional encoders. In our implementation, we use the encoder of Pix2PixHD [4] for the pose condition maps, of which the input size is $9 \times 768 \times 576$. Here we follow the notation of tensor shape in PyTorch, i.e., $C \times H \times W$, where C, H, W denote the channel, the height, and the width of a tensor. The input size of the encoder extracting appearance features is $10 \times 768 \times 576$. The pose features and the appearance features are merged via the decoder, which is composed of five sequential SPADE modules [3] (for feature transformation and upsampling). Features fed into these five modules are of size $1024 \times 24 \times 18$, $1024 \times 48 \times 36$, $512 \times 96 \times 72$, $256 \times 192 \times 144$, and $128 \times 384 \times 288$, respectively. The synthesized image is of size $3 \times 768 \times 576$.



(a) (b) (c) (d) (e) (f)

Figure 2. Failure cases. (a), (c) and (e) are the original images. (b), (d) and (f) are the synthesized images generated by the BodyGAN.

2. Qualitative Results

We provide more visual examples of our synthesized images. Fig. 1 demonstrates the results of our methods and other two state-of-the-art virtual try-on methods (CP-VTON-PLUS [2] and PF-AFN [1]). Fig. 3 shows our

*: These authors contribute equally.
†: Corresponding authors.

Figure 3. More visual examples of our synthesized images. The first column are the original images, while the other columns are the synthesized images generated by the BodyGAN with various rendered clothes.

synthesized images with various rendered clothes. In fact, our BodyGAN can generate realistic faces and poses under large changes, as shown in Fig. 4

Our BodyGAN is robust with different subnetwork configurations. We implement three variants, i.e., replacing the backbone for key point detection with HRNet (denoted as Ours-HRNet-kp), enhancing the segmentation network by considering more classes (Ours-enh-sg), and constructing a lightweight DensePose model (Ours-lw-dp). As reported in Tab. 1, these variants achieve similar results on DeepFashion and outperform other methods.

3. Failure Cases

As we have discussed in the paper, there are a few limitations of our BodyGAN, which we propose to handle in the

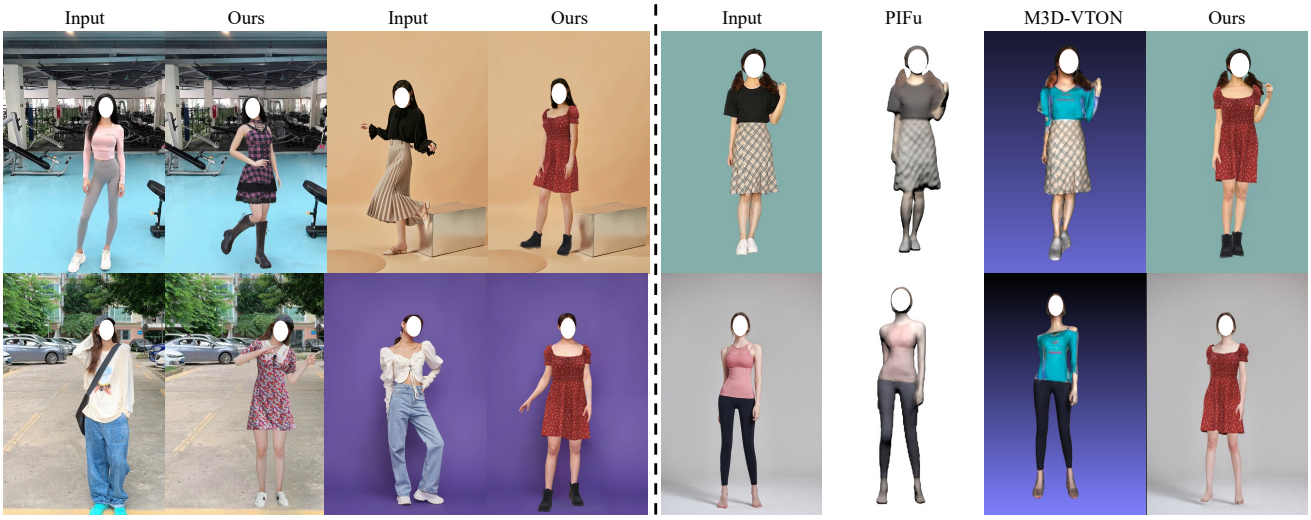


Figure 4. Left: Try-on results of our method under large pose changes and head rotations. Right: Visual comparison with 3D model based methods. The results of our BodyGAN have realistic, rich, and complex textures/details. Please zoom in for the details.

Methods/Results	SSIM ↑	FID↓	LPIPS↓
SPADE*	0.5601	37.2646	0.0234
Pix2PixHD*	0.6205	29.7461	0.0202
Ours-HRNet-kp	0.8471	6.1655	0.0070
Ours-enh-sg	0.8483	6.0487	0.0070
Ours-lw-dp	0.8002	8.5279	0.0097
Ours-original	0.8470	6.1654	0.0070

Table 1. Performance of BodyGAN with different subnetworks.

future. Fig. 2 shows a few failure cases of the BodyGAN. The source images in the first row are captured from behind the person, which are rare (especially for applications like try-on) and such data are not available in our training set. Furthermore, it is hard to detect the faces in this case, which hinders the performance of pose encoding. The source images in the second row are affected by non-uniform illumination distributions and color distortions. In this case, recovering the original skin color of the person is difficult. Therefore, the synthesized images of our BodyGAN are not satisfactory under these two circumstances. Possible solutions for these problems include anomaly detection and ad-hoc procedures/modules. For example, we may consider certain priors or constraints of the skin color to detect the anomalous image, and recover the normal/natural skin color with those from other body parts.

4. Mitigation Strategies for Negative Social Impacts

Our method focuses on human body synthesis, and its results are realistic and be controlled conveniently over multiple factors, which might be used for generating fake images and videos. Therefore, we will not publish the code of our BodyGAN. However, we will consider building a platform for demonstration, or providing encrypted APIs for authenticated downstream applications. Interested Readers can also contact us via email for technical and implementation details.

References

[1] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8481–8489, 2021. 1

[2] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul L. Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1

[3] Taesung Park, Ming-Yu Liu, T. Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2332–2341, 2019. 1

[4] T. Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, J. Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 1