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

Abstract

Recent advances in generative adversarial networks (GANs) have provided potential solutions for photo-realistic human image synthesis. However, the explicit and individual control of synthesis over multiple factors, such as poses, body shapes, and skin colors, remains difficult for existing methods. This is because current methods mainly rely on a single pose/appearance model, which is limited in disentangling various poses and appearance in human images. In addition, such a unimodal strategy is prone to causing severe artifacts in the generated images like color distortions and unrealistic textures. To tackle these issues, this paper proposes a multi-factor conditioned method dubbed BodyGAN. Specifically, given a source image, our BodyGAN aims at capturing the characteristics of the human body from multiple aspects: (i) A pose encoding branch consisting of three hybrid subnetworks is adopted, to generate the semantic segmentation based representation, the 3D surface based representation, and the key point based representation of the human body, respectively. (ii) Based on the segmentation results, an appearance encoding branch is used to obtain the appearance information of the human body parts. (iii) The outputs of these two branches are represented by user-editable condition maps, which are then processed by a generator to predict the synthesized image. In this way, our BodyGAN can achieve the fine-grained disentanglement of pose, body shape, and appearance, and consequently enable the explicit and effective control of synthesis with diverse conditions. Extensive experiments on multiple datasets and a comprehensive user study show that our BodyGAN achieves the state-of-the-art performance.

1. Introduction

Realistic human body image generation is a challenging problem, due to the complex textures, diverse poses and illumination distributions. As there are a large number of application scenarios that rely on the high-quality and controllable human image generation, such as virtual try-on and virtual reality, extensive research has been conducted to tackle this problem. For instance, in recent research on virtual try-on [33, 47], both mask based methods [35, 37] and mask-free methods [25, 42] have been explored to generate images of the human body with target clothes. However, these methods consider the human body and clothes together as the whole target, and adopt the end-to-end generation manner, which leads to poor results and artifacts on human body parts, such as unnatural texture patterns, loss of details and color distortions. Meanwhile, a few skin completion methods [23, 43] target at the generation of face skin. There are also algorithms [5, 18, 28, 30] generating faces based on contour sketches. Nevertheless, considering that the human

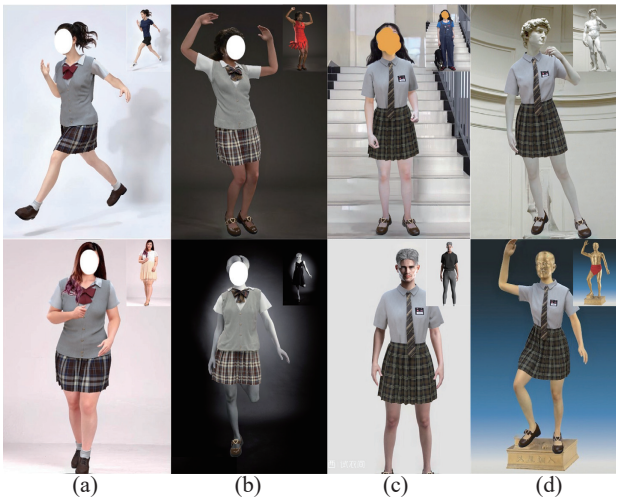


Figure 1. We propose BodyGAN, i.e., a general approach for synthesizing photo-realistic human body with explicit control over multiple factors: (a) Body pose and shape. (b) Skin colors. (c) Hidden face and cartoon. (d) Marble and bronze sculptures. The source pictures are shown in the top right corners.

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

body is a flexible articulated structure with a higher degree of freedom, the above solutions are not suitable for generating human body images.

Recently, a few GAN based methods have been proposed for human body image synthesis, such as StyleGAN [15], StylePoseGAN [34] and StyleRig [36]. The key idea of these methods is to utilize a single model to disentangle pose and appearance, so that different synthesized images can be obtained via changing the pose or the appearance. However, in real-life applications, a single model is not robust enough to disentangle images with various poses and appearances. Furthermore, existing methods may need paired training images (e.g., StylePoseGAN) to ensure the performance of their models. These issues of existing methods make the explicit control of the synthesis difficult and restrict their applications.

To tackle the above limitations of existing methods, we propose a highly controllable framework for human image synthesis called BodyGAN. Unlike previous methods aiming at a particular synthesis task with a single end-to-end network, our BodyGAN focuses on the generation of pure human body images with controllable poses, body shapes and skin colors. The generated human body images are then modified based on the requirements of the downstream task (e.g., merged with a garment to complete try-on). Specifically, given a source image, our BodyGAN adopts a pose encoding branch and an appearance encoding branch, to generate seven types of condition maps for image synthesis. In the pose encoding branch, we utilize three subnetworks to obtain the semantic segmentation based condition map, the 3D surface based condition map, and the key point based condition map of the human body. While in the appearance encoding branch, four appearance condition maps are used to encode the head, the hand, the upper body, and the lower body of the human body. In this way, we achieve the disentanglement of pose and appearance that is more fine-grained, compared with existing methods. Besides, our condition maps encode pose information via both 2D representations (semantic segmentation and key points) and 3D surface faces, which help to generate more natural poses. The condition maps are fed into a generator that is optimized via adversarial learning (without the need for paired training images), to produce the synthesized images.

The contributions of this paper are summarized as follows:

- We propose BodyGAN, which is a general model for human body image generation with explicit controls over multiple factors. Our BodyGAN can disentangle pose and appearance from a single source image, and adopt condition maps, which are a convenient and editable representation of pose and appearance information, to generate realistic human images.
- The proposed condition maps are generated by three subnetworks that model the person from both 2D and 3D perspectives, which enhance the robustness of our BodyGAN effectively. In addition, utilizing pose and appearance condition maps allows us to train and apply the BodyGAN with unpaired images.
- Compared with the existing methods, our BodyGAN can generate more realistic and visually-pleasant results, even with extremely difficult poses such as crossed arms/legs. Extensive experiments on three datasets validate the effectiveness of our BodyGAN.
- We present the downstream applications of BodyGAN in image manipulation, such as virtual try-on and digital humans. In summary, BodyGAN provides a novel diagram for human image generation in computer animation.

2. Related Work

Human rendering. Early human body rendering methods rely on high-quality mannequins, well-calibrated textures, and expensive hardware [41] to achieve realistic results. With the recent advances in deep learning, neural rendering techniques [16, 20, 24] which combine learnable networks with physical knowledge (e.g., illumination model and geometry), have been proposed to speed up the rendering process. In spite of their convenience and efficiency, neural rendering methods are still limited in controllable image synthesis.

Conditional GAN Generation. GANs [32] are one of the most popular deep learning techniques for data generation, which have been applied to a large number of applications in various fields. In the past few years, extensive improvements on GANs [1, 14, 44] have been proposed, which make the training of GAN more stable and generate better results. To realize controllable generation, Mehdi Mirza and Simon Osindero proposed the conditional GAN (CGAN) [26], and its variants have been utilized in image-to-image translation [11,39] and conditioned image generation [27]. Current CGANs [11,26,29,32] require paired data to guide the training process, which limit their applications because of the difficulty of collecting paired data. Furthermore, it is hard for existing methods to introduce multiple controllable factors. For example, [34] aims at separating the pose and appearance to control body synthesis. Nevertheless, [34] adopts the image-to-image translation framework, and hence both the source image and the target image are required. On the contrary, our BodyGAN is self-supervised and does not require paired data. Besides, our BodyGAN can manipulate the skin colors and use a more flexible 3D human body model to complete synthesis conditioning on pose and shape.

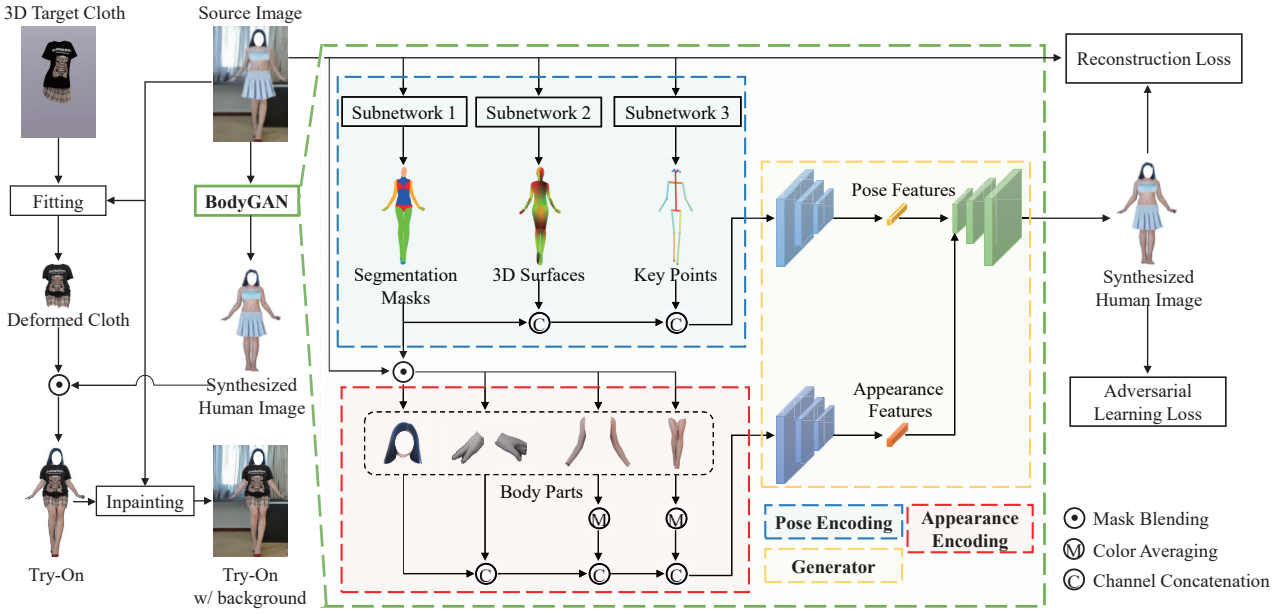


Figure 2. Architecture of the BodyGAN, with virtual try-on as the application. The BodyGAN consists of three components, i.e., the pose encoding branch, the appearance encoding branch and the generator. The two encoding branches obtain the condition maps from the source image, which are fed into the generator to produce the synthesized image.

Motion Transfer/Virtual Clothes Try-On. Motion transfer and virtual clothes try-on are the major applications of human image synthesis. For example, [13] consider the clothes and bodies generation as a concurrent task for synthesis. To obtain satisfactory results, most methods utilize the vid2vid framework [3,31,38] to accomplish these tasks. However, the generalization ability of existing methods is insufficient, and the drop of performance is unavoidable on unknown persons or clothes with new styles. On the other hand, our BodyGAN focuses on the synthesis of human bodies, i.e., the training process of our BodyGAN does not involve downstream applications like try-on. Such a strategy not only ensures the flexibility of our BobyGAN, but also improves its generalization ability.

3. The Proposed Method

In this section we introduce the details of the proposed BodyGAN. We begin by formulating the problem of human image synthesis conditioned on multiple factors in Section 3.1, and then introduce the architecture of BodyGAN for try-on in Section 3.2. The key components of BodyGAN, including two branches for encoding pose and appearance, and a generator for synthesizing human images, are presented in Section 3.3 and Section 3.4, respectively. The objective functions for optimizing the BodyGAN are introduced in Section 3.5.

3.1. Problem Formulation

Given a source human image, our goal is to synthesize a realistic image of the same person, with the explicit control of multiple factors, such as pose, shape and skin color. Formally, let I denote the source image, and $C = \{c_1, c_2, ..., c_M\}$ denote the set of M factors that we are interested in, we aim at learning a transformation f as follows:

$$I_G = f(I|C), \tag{1}$$

so that we can obtain the desired synthesized image (denoted as I_G) via changing the factors in C. In real-world applications, the representations of factors can be various, e.g., c_m can be a key point based representation of pose, or a feature vector depicting certain styles/textures. Without loss of generality, we assume an arbitrary factor $c_m \in C$ can be represented by a condition map of size $H \times W \times D_m$, where H and W are the height and width of the source image, and D_m is the number of user-defined channels of c_m. With such a condition map based representation, local/spatial controls of factors become feasible. Our target now can be further divided into (i) finding a way to obtain condition maps conveniently, as well as (ii) realizing f, and both of them are completed by our BodyGAN introduced in the next section.

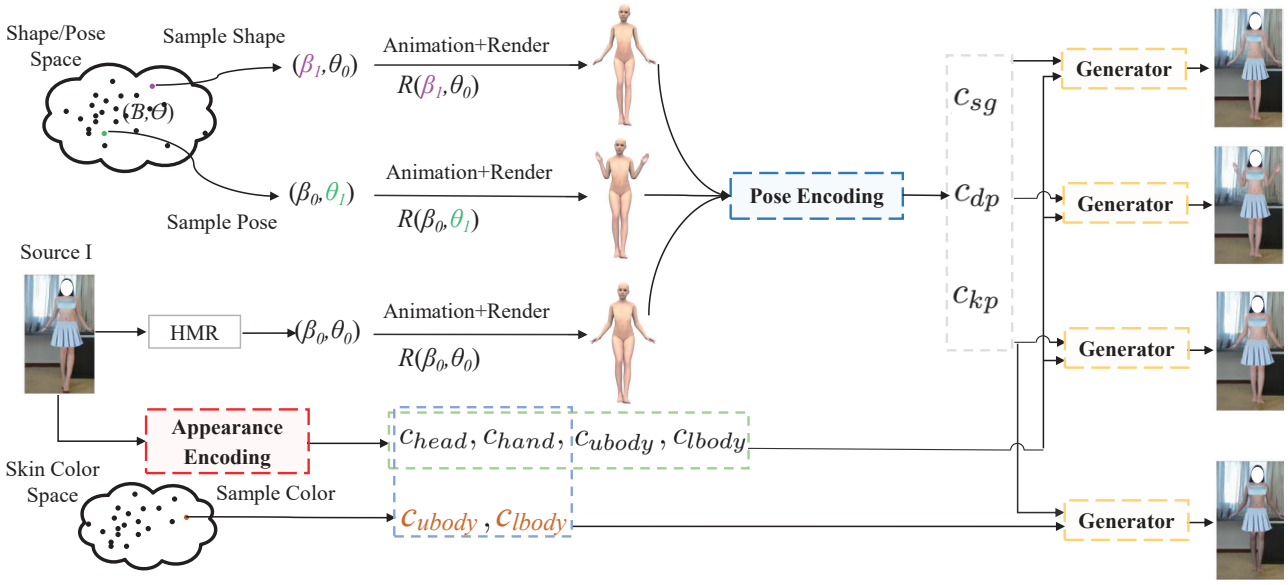


Figure 3. Demonstration of the inference stage of the BodyGAN. We can obtain the pose condition maps of a source image through a 3D model estimated by HMR, in which parameters β and θ control the shape and the pose of the 3D model respectively. The pose condition maps are combined with the appearance condition maps to generate various synthesized images.

3.2. Network Architecture

The proposed BodyGAN is a general framework for controllable human body image generation. For the purpose of better understanding, here we consider the application of virtual try-on as an example to explain the workflow of BodyGAN, which is shown in Fig. 2. Given a 3D garment and a source image, we first apply the BodyGAN to crop out the person from the source image and generate the synthesized image with desired properties. After that, the 3D garment is deformed to fit the person, and then the rendered garment and the synthesized image are merged via mask blending. At last, we apply inpainting [19] to transfer the background and obtain the try-on result.

The BodyGAN is composed of three major components: the pose encoding branch and the appearance encoding branch are responsible for generating the condition maps from the source image, while the generator converts the condition maps into the synthesized image. To ensure the fine-grained disentanglement of pose and appearance, we consider seven types of condition maps in the BodyGAN and all of them can be extracted efficiently: In the pose encoding branch, three pose condition maps are obtained to encode the pose with different representations; While in the appearance encoding branch, four appearance condition maps are used to describe the appearance information of multiple parts of the target person. Details of these condition maps will be introduced in the following section.

The generator adopts an encoder-decoder architecture, and is optimized via adversarial learning and reconstructing the source image. Thanks to the above disentanglement of pose and appearance via the condition maps, our BodyGAN does not require the costly collection of paired training images. This provides our BodyGAN with the great flexibility for various applications.

3.3. Pose and Appearance Encoding

Pose Encoding. Unlike previous methods that mainly depend on a single pose model, our BodyGAN utilizes three subnetworks to encode the pose condition maps. Specifically, we propose to utilize the results of semantic segmentation, 3D surface mapping, and key point estimation, to obtain the pose condition maps. To begin with, we construct our own human parsing network based on DeepLab-V3+ [4], to segment the source image into eleven semantic classes (head, hand, etc.). We draw the segmentation results with a predefined color for each class, so that we obtain the first pose condition map $c_{sg} \in [0,1]^{H\times W\times 3}$. After that, we map the 2D pixels of the source image to the surface of the 3D SMPL model [22] via DensePose [8]. The resulted mapping $c_{dp} \in [0,1]^{H\times W\times 3}$ is used as the second pose condition map directly. At last, we estimate key points from the source image via OpenPose [2], and plot the stick figure with predefined connection orders and colors. The stick figure $c_{kp} \in [0,1]^{H\times W\times 3}$ is used as the third condition map. We concatenate all the three pose condition maps

as the output of the pose encoding branch:

$$c_p = concat(c_{sg}, c_{dp}, c_{kp}). \quad (2)$$

Our three pose condition maps cooperate with each other closely to enrich the information extracted both from 2D source images and 3D articulated models, and overcome the limitations of existing methods in pose modeling. For example, the difficult issue of arms/legs crossing in generating human body can be addressed by the 3D surface maps easily, because they offer a depth-map like prior for distinguishing each body part. Besides, the 2D semantic segmentation results are robust and reliable, which not only improves the performance of estimating the body shape but also helps to locate the regions for appearance encoding.

Appearance Encoding. The semantic segmentation results can be utilized as the masks to obtain the appearance information of each body part efficiently. Particularly, our appearance condition maps $c_a \in [0,1]^{H\times W\times 10}$ are defined as follows:

$$c_a = concat(c_{head}, c_{hand}, c_{ubody}, c_{lbody}), \quad (3)$$

where c_{head}, c_{hand}, c_{ubody}, and c_{lbody} denote the masked image of the head, the hands, the upper body, and the lower body, respectively. c_{head}, c_{ubody}, and c_{lbody} are in $[0,1]^{H\times W\times 3}$. For the hand part, we use the gray-scale image instead of the original RGB image (i.e., $c_{hand} \in [0,1]^{H\times W\times 1}$), because in practice we found that the textures and the illumination distributions on hands are more complex than the other parts, and using the gray-scale image helps to alleviate the produced artifacts. For the similar purpose, we set the pixel values in c_{ubody}/c_{lbody} to the spatial mean of the skin color in the upper/lower body.

In the inference stage, it is convenient to prepare the condition maps to generate various synthesized images, as shown in Fig. 3. For instance, to transfer a pose, we can either render the pose condition maps from a given SMPL model, or estimate the SMPL model from a given reference image via existing methods (e.g., Human Mesh Recovery (HMR) [12]). Unseen poses and shapes also can be achieved via changing the parameters of the SMPL model. With the condition maps, the heavy cost of maneuver manipulations can be reduced considerably.

3.4. Generator

We construct an encoder-decoder based generator to synthesize realistic images based on the condition maps. Following the common configuration of encoder, we adopt four convolutional layers (each followed by a downsampling layer) to embed the pose condition maps c_p into a compact pose feature representation, while two convolution-downsampling layers for embedding the appearance condition maps c_a. For the decoder part, inspired by the semantic layout based image synthesis method [29], we consider the pose feature as the major feature (as it also represents the spatial layout of the person), and use the appearance feature as the conditions for normalizing the pose feature. Therefore, the same SPADE layer proposed in [29] is used to build our decoder. We use four SPADE layers in total and each of them is followed by an upsampling layer.

In the training stage, two discriminators (each for one branch in BodyGAN) of the same architecture as in Pix2PixHD [39] are utilized to realize adversarial learning. For the pose branch, the discriminator takes the concatenation of c_p and I/I_G as its input, while that for the appearance branch takes the concatenation of c_a and I/I_G. Through the pose-appearance separation, each discriminator can focus on its own task and help to improve the generator, and consequently the generator can produce realistic high-quality images.

3.5. Objective Function

The objective functions for optimizing our BodyGAN are defined as follows:

$$L = \lambda_{REC} L_{REC} + \lambda_{GAN} L_{GAN}, \quad (4)$$

where L_{REC} denotes the reconstruction loss and L_{GAN} denotes the improved adversarial loss [39]. λ_{REC} and λ_{GAN} are the scaling factors for balancing the losses. The reconstruction loss can be further divided as follows:

$$L_{REC} = \lambda_1 L_1 + \lambda_{SSIM} L_{SSIM} + \lambda_{VGG} L_{VGG}, \quad (5)$$

where L_1 is the absolute error as follows:

$$L_1 = \|I_G - I\|_1. \quad (6)$$

L_{SSIM} is used to emphasize the structural information in images, which is defined as follows:

$$L_{SSIM} = 1 - SSIM(I_G, I), \quad (7)$$

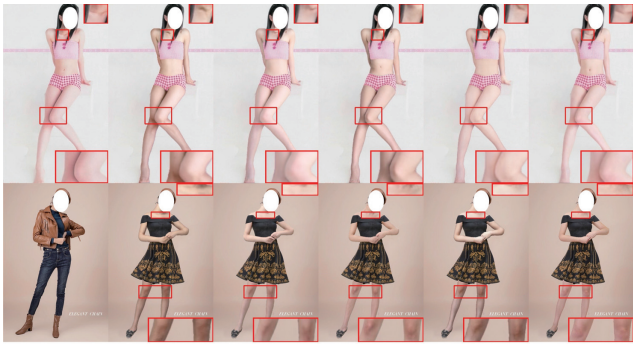
where $SSIM(I_G, I)$ denotes the structural similarity index measure between I_G and I [40]. L_{VGG} is the widely-used perceptual loss that measures the distance between the hidden features of I_G and I extracted by VGG Network [6]. Again, λ_1, λ_{SSIM}, and λ_{VGG} are the scaling factors of these reconstruction loss terms.

4. Experiments

In this section, we conduct extensive experiments to validate the proposed BodyGAN. Due to the page limitation, interested readers can refer to the supplemental materials for the complete network implementation, more experimental results and analysis (e.g., failure cases).

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

Figure 4. Qualitative comparison among different methods. (a) Source image, (b) SPADE [29], (c) SPADE*, (d) Pix2PixHD [39], (e) Pix2PixHD*, (f) Ours, (g) PF-AFN [7]. '*' denotes methods with the same input as our method.



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

Figure 5. Qualitative examples of our method with different condition maps. (a) source image, (b) c_{sg}, (c) c_{sg} + c_{kp}, (d) c_p, (e) c_p + c_{hand} , (f) c_p + c_a. Best zoom in for the details.

4.1. Setup

Dataset. We have constructed our own dataset for training and evaluation. Our dataset consists of 36k human images with tight and short clothes, i.e., underwear, under normal light conditions. We use 33.6k images for training and the rest 2.4k images for test. All training images are resized to 576×768. Besides, We also include two public datasets for evaluation, namely, the VITON dataset [9] and the DeepFashion dataset [21]. On the DeepFashion dataset, we select the images with the largest side larger than 1200, so that we have 0.8k high-resolution test images.

Implementation Details. We train the BodyGAN on our own dataset, with four NVIDIA Tesla-V100 GPUs. The weights for the loss terms are set to λ_{REC}=1.0, λ_{L1}=5.0, λ_{SSIM}=1.0, λ_{VGG}=1.0, and λ_{GAN}=2.0. We adopt the Adam optimizer [17] with the learning rate of 0.0001 and the momentum of 0.5.

4.2. Comparison with SOTAs

We utilize SSIM [40], FID [10], and LPIPS [46] as the metrics to evaluate quality of synthesized images. During the evaluation, the background areas will be removed, so that our evaluation metrics can concentrate on the quality of the generated human bodies.

We compare the BodyGAN with several state-of-the-art methods, including SPADE [29], Pix2PixHD [39], and CoCosNet [45]. The inputs of these methods are different, and hence we also implement two variants, i.e., SPADE* and Pix2PixHD*, which use the same input as our BodyGAN. As reported in Table 1, our BodyGAN outperforms these methods consistently on all three datasets. These results suggest that our BodyGAN achieves the better disentangle of pose and appearance, and hence it obtains the higher performance, compared with other methods.

Qualitative results of different methods are shown in Fig. 4. The first row shows the results of different methods in reconstructing the source image from the condition maps, and we can see that our BodyGAN better restores the textures and other details of the 2D source image, compared with other methods. The second row shows synthesized images with 3D body parameters obtained through the HMR method [12]. Compared with other methods, our result demonstrates the delicate body textures (e.g., around the necks and knees), and is more visual pleasing. We also include PF-AFN [7] for visual comparison, which is a well-designed virtual try-on method. As shown in Fig. 4 (g), PF-AFN suffers from a few artifacts like color distortions and loss of details, while our method avoids these artifacts successfully.

4.3. User Study

We conduct a user study on one hundred randomly chosen test images from our dataset. We concatenate each source image with its five reconstructed counterparts generated by SPADE, SPADE*, Pix2PixHD, Pix2PixHD*, and BodyGAN for comparison. The concatenated images are presented to ten users in randomly shuffled orders. We ask the users to score the reconstructed images based on photo-realistic quality and naturalness. Our method receives the highest average score of 72.3% among all methods, while the scores of SPADE, SPADE*, Pix2PixHD and Pix2PixHD* are 3.1%, 1.6%, 2.3% and 20.7% respectively.

4.4. Ablation Study

We conduct the ablation study on the effects of conditional maps. Quantitative results of our BodyGAN with different conditional maps are shown in Table 2. These results validate that all our condition maps play important roles in enhancing the quality of the generated images. Qualitative results for this ablation study are shown in Fig. 5 as well. Both the quantitative and qualitative results indicate that our

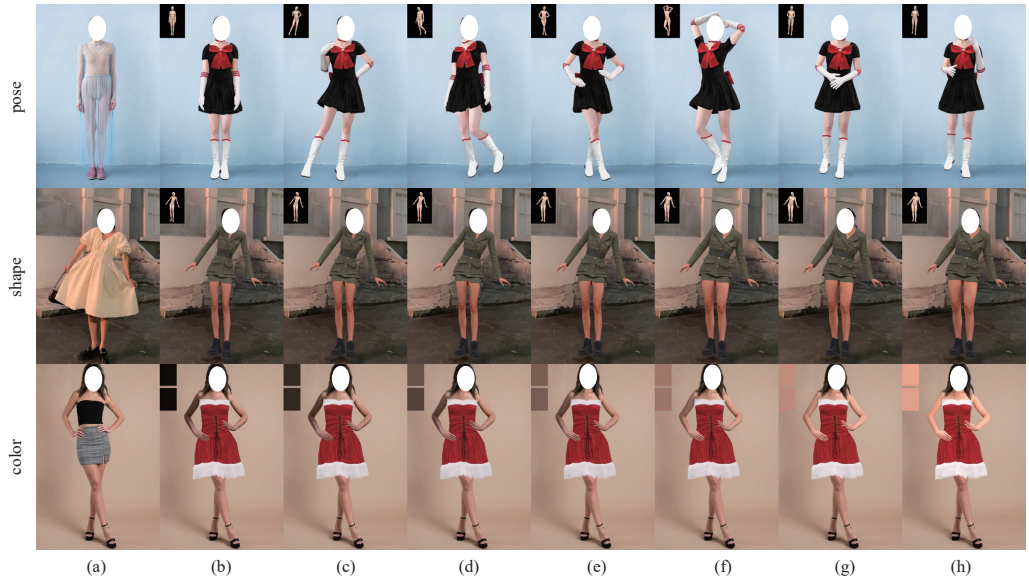


Figure 6. Visual examples of pose, shape, and skin color conditioning synthesis. (a) Source images. (b) to (h) are synthesized results with the conditions shown in the top-left corners.

	Our Test Set			DeepFashion			VITON		
Methods/Results	SSIM ↑	FID↓	LPIPS↓	SSIM ↑	FID↓	LPIPS↓	SSIM ↑	FID↓	LPIPS↓
SPADE [29]	0.6540	16.4102	0.0728	0.6186	36.7924	0.0213	0.6369	65.5421	0.0334
SPADE* [29]	0.6723	15.8825	0.0781	0.5601	37.2646	0.0234	0.5948	65.5927	0.0331
Pix2PixHD [39]	0.7782	4.4394	0.0351	0.6098	29.6315	0.0215	0.6215	53.1249	0.0359
Pix2PixHD* [39]	0.7884	4.4417	0.0346	0.6205	29.7461	0.0202	0.6383	42.0546	0.0346
CoCosNet [45]	0.3804	193.207	0.0991	0.4714	139.514	0.0401	0.5278	110.958	0.0458
Ours	**0.8307**	**3.6760**	**0.0297**	**0.8470**	**6.1654**	**0.0070**	**0.8249**	**4.4529**	**0.0146**

Table 1. Quantitative comparison of different methods. '*' denotes methods with the same input as our method.

BodyGAN with the chosen combination of conditional images outperforms those with a single condition.

5. Applications

Benefiting from the dual-branch architecture which decouples appearance and pose, our model can explicitly control body poses, shapes and skin colors of the generated body and can be applied to various human image synthesis related applications.

Virtual Try-On. The BodyGAN can be combined with the cloth generation method to accomplish virtual try-on. Fig. 4 and Fig. 6 show our synthesized human body images for try-on. As clothes are not the main concern of this paper, we render the clothes in the above figures with the computer graphics pipeline, or extract them from the source images with semantic segmentation. Fig. 6 shows that our BodyGAN can be adapted to various conditions, and generate images of the same person with different poses, shapes, and colors.

VR/AR/Metaverse. In the near future, we may design an avatar to live, socialize and work as an independent individual in the metaverse. For this application, we can design a 3D human body model (Daz or SMPL [22], etc.), and utilize the BodyGAN to generate an exclusive and realistic character with the desired appearance and pose efficiently.

6. Discussion

Potential Negative Social Impact. The proposed method can generate realistic human images, which might be misused for generating fake photos and videos.

Limitations. There are a few limitations of our BodyGAN, which we propose to tackle in the future: (i) A few source images might have non-uniform illumination distributions (e.g., with neon lighting), in this case, the skin color of the original image is distorted, and consequently the skin color of the synthesized image is distorted as well. (ii) It is hard for the subnetworks to encode the target person with heavy occlusion, as the pose might not be restored well.

Figure 7. Visual examples of our synthesized images on the DeepFashion dataset [21]. The first column are the original images, while the other columns are the synthesized images generated by the BodyGAN with various rendered clothes.

	Our Test Set			DeepFashion			VITON		
Condition Maps	SSIM ↑	FID↓	LPIPS↓	SSIM ↑	FID↓	LPIPS↓	SSIM ↑	FID↓	LPIPS↓
c_{sg}	0.7466	6.0055	0.0473	0.6141	10.8691	0.0199	0.6754	8.7129	0.0273
c_{sg} + c_{kp}	0.7602	5.7316	0.0477	0.6724	9.7372	0.01686	0.7022	7.9809	0.0263
c_p	0.7694	5.2623	0.0408	0.6580	10.1058	0.0175	0.7068	8.4752	0.0267
c_p + c_{hand}	0.7807	5.5167	0.0445	0.7665	7.4670	0.0131	0.7916	6.0040	0.0199
c_p + c_a (Ours)	**0.8307**	**3.6760**	**0.0297**	**0.8470**	**6.1654**	**0.0070**	**0.8249**	**4.4529**	**0.0146**

Table 2. Ablation study of the proposed method with different condition maps.

7. Conclusion

In this paper, we propose a general-purpose human image synthesis framework called BodyGAN, which provides explicit control over the body pose, shape and skin colors. Our BodyGAN utilizes a pose encoding branch and an appearance encoding branch, which generate pose and appearance condition maps as the bridge between user-desired synthesized effects and the actual inputs of the network. More importantly, these condition maps disentangle pose and appearance explicitly, which not only provides us with the benefit of training without paired images, but also improves the quality of our synthesized images. Our experiments on three datasets and the user study show that the BodyGAN outperforms other state-of-the-art methods significantly. Combined with physics-based clothes simulation or other well-developed simulation methods, the proposed BodyGAN can be applied to various human related applications, e.g., virtual clothes try-on, animation and metaverse.

8. Acknowledgements

This work was supported in part by Shenzhen Fundamental Research Program (Project No. RCYX20200714114642083, No. JCYJ20190807154211365), and National Natural Science Foundation of China under Grant No. 61902088.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019. 2

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. *Proceedings of the IEEE International Conference on Computer Vision*, pages 5932–5941, 2019. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018. 4

[5] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2018. 1

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 5

[7] 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. 6

[8] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 4

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 6

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 2

[12] Angjoo Kanazawa, Michael J. Black, D. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 5, 6

[13] Moritz Kappel, Vladislav Golyanik, Mohamed A. Elgharib, Jann-Ole Henningson, Hans-Peter Seidel, Susana Castillo, Christian Theobalt, and Marcus A. Magnor. High-fidelity neural human motion transfer from monocular video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1541–1550, 2021. 3

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2018. 2

[15] Tero Karras, S. Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4396–4405, 2019. 2

[16] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, M. Nießner, P. Pérez, Christian Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics*, 37:1 – 14, 2018. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ArXiv*, abs/1412.6980, 2015. 6

[18] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. Linestofacephoto. *Proceedings of the ACM International Conference on Multimedia*, Oct 2019. 1

[19] Guilin Liu, F. Reda, Kevin J. Shih, T. Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *ArXiv*, abs/1804.07723, 2018. 4

[20] Lingjie Liu, Weipeng Xu, Marc Habermann, M. Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and C. Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP, 2020. 2

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6, 8

[22] M. Loper, Naureen Mahmood, J. Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trasactions on Graphics*, 34:248:1–248:16, 2015. 4, 7

[23] Dejan Malesevic, Christoph Mayer, Shuhang Gu, and Radu Timofte. Photo-realistic and robust inpainting of faces using refinement gans. In *Inpainting and Denoising Challenges*, 2019. 1

[24] Moustafa Meshry, Dan B. Goldman, S. Khamis, Hugues Hoppe, R. Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6871–6880, 2019. 2

[25] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul 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*, June 2020. 1

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014. 2

[27] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *ArXiv*, abs/1802.05637, 2018. 2

[28] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning, 2019. 1

[29] Taesung Park, Ming-Yu Liu, T. Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normaliza-

tion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2332–2341, 2019. 2, 5, 6, 7

[30] Tiziano Portenier, Qiyang Hu, Attila Szab, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop. *ACM Transactions on Graphics*, 37(4):113, Aug 2018. 1

[31] Jian Ren, Menglei Chai, Oliver J. Woodford, Kyle Olszewski, and S. Tulyakov. Flow guided transformable bottleneck networks for motion retargeting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10790–10800, 2021. 3

[32] Tim Salimans, I. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016. 2

[33] Soubhik Sanyal, Alex Vorobiov, Timo Bolkart, Matthew Loper, Betty Mohler, Larry Davis, Javier Romero, and Michael J. Black. Learning realistic human reposing using cyclic self-supervision with 3d shape, pose, and appearance consistency, 2021. 1

[34] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and C. Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *ArXiv*, abs/2102.11263, 2021. 2

[35] S. Song, W. Zhang, J. Liu, Z. Guo, and T. Mei. Unpaired person image generation with semantic parsing transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2020. 1

[36] Ayush Tewari, Mohamed A. Elgharib, Gaurav Bharaj, Florian Bernard, H. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6141–6150, 2020. 2

[37] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision*, pages 589–604, 2018. 1

[38] T. Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, J. Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *ArXiv*, abs/1910.12713, 2019. 3

[39] 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. 2, 5, 6, 7

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5, 6

[41] XuFeng, LiuYebin, StollCarsten, TompkinJames, BharajGaurav, DaiQionghai, SeidelHans-Peter, KautzJan, and TheobaltChristian. Video-based characters. *ACM Transactions on Graphics*, 2011. 2

[42] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1

[43] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models, 2017. 1

[44] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, 2019. 2

[45] P. Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5142–5152, 2020. 6, 7

[46] Richard Zhang, Phillip Isola, Alexei A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[47] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13239–13249, 2021. 1