Sketch2Human: Deep Human Generation with Disentangled Geometry and Appearance Constraints

Linzi Qu, Jiaxiang Shang, Hui Ye, Xiaoguang Han, and Hongbo Fu

***Abstract*—Geometry- and appearance-controlled full-body human image generation is an interesting but challenging task. Existing solutions are either unconditional or dependent on coarse conditions (e.g., pose, text), thus lacking explicit geometry and appearance control of body and garment. Sketching offers such editing ability and has been adopted in various sketch-based face generation and editing solutions. However, directly adapting sketch-based face generation to full-body generation often fails to produce high-fidelity and diverse results due to the high complexity and diversity in the pose, body shape, and garment shape and texture. Recent geometrically controllable diffusion-based methods mainly rely on prompts to generate appearance. It is hard to balance the realism and the faithfulness of their results to the sketch when the input is coarse. This work presents Sketch2Human, the first system for controllable full-body human image generation guided by a semantic sketch (for geometry control) and a reference image (for appearance control). Our solution is based on the latent space of StyleGAN-Human with inverted geometry and appearance latent codes as input. Specifically, we present a sketch encoder trained with a large synthetic dataset sampled from StyleGAN-Human's latent space and directly supervised by sketches rather than real images. Considering the entangled information of partial geometry and texture in StyleGAN-Human and the absence of disentangled datasets, we design a novel training scheme that creates geometry-preserved and appearance-transferred training data to tune a generator to achieve disentangled geometry and appearance control. Although our method is trained with synthetic data, it can also handle hand-drawn sketches. Qualitative and quantitative evaluations demonstrate the superior performance of our method to state-of-the-art methods. We will release the code upon the acceptance of the paper.**

***Index Terms*—Full-body image generation, style-based generator, style mixing, sketch-based generation**

I. Introduction

REALISTIC full-body human image synthesis benefits various applications like fashion design [1], virtual try-on [2]–[5], 2D avatar creation [6], [7], and animations [8], [9]. For such applications, high-fidelity generation and interactive control are both essential to generate specific images of interest. Although existing human image generation methods have produced impressive results, they are either unconditional [7], [10] or based on coarse conditions [3], [11], [12]. The unconditional methods produce high-fidelity and diverse images but lose controllability. The coarsely conditioned methods achieve high-level control by coarse representations (*e.g.*, pose [3], [12], text [9], [13]–[15], reference image [2], [4], [16]–[19]). However, these methods fail to explicitly and flexibly control detailed geometry (e.g., body contour, garment shape) and appearance (e.g., skin color, garment texture) simultaneously. In terms of design, both novices and professionals with specific designs in mind often prefer a more subjective and specific control of results.

Sketches are often used to explicitly depict desired geometry due to their simplicity, ease of modification, and ability to represent details. Sketch-based image generation techniques have been well explored in the human face domain [20]–[22]. However, sketch-based full-body human image generation is underexplored and more challenging than face generation since human bodies involve a larger variety of poses, garment shapes, and textures. These difficulties undoubtedly require a large amount of data for training. Meanwhile, it is almost impossible to collect a real disentangled dataset including images with the same geometry but varying appearance or the same appearance with varying geometry for geometry and appearance control. Therefore, directly applying previous methods for human face generation to generate human bodies fails to produce high-fidelity results with diverse appearances (especially for subtle regions, *e.g.*, face, shoes, glasses, and garment patterns) and preserve the reasonable human body structure. Benefiting from the prior knowledge in large text-to-image diffusion models [23], [24], ControlNet [25] and T2I-Adapter [26] leverage sketches to guide the multi-class object generation. However, with the increase of sketch abstraction, it is difficult to balance the photo-realism of the results and their faithfulness to the input sketches (Figure 7) and these methods cannot support consistent appearance control via appearance examples.

To address these issues, we propose *Sketch2Human*, a novel deep generative framework for synthesizing a realistic human image from a semantic sketch (for flexible and detailed geometry control) and a reference appearance image (for appearance control), as shown in Figure 1. Due to the lack of a disentangled full-body human dataset, we start from an unconditional generation model StyleGAN-Human [10], which roughly disentangles the geometry and appearance in the latent space, as shown in Figure 2. To produce photorealistic human images from sketches with different levels of abstraction, we embed the sketch in the latent space and then use StyleGAN-Human to generate results from such embeddings. To further

Corresponding author: Hongbo Fu
L. Qu, H. Ye, and H. Fu are with the School of Creative Media, City University of Hong Kong. E-mail: linziqu2-c@my.cityu.edu.hk, huiye4@cityu.edu.hk, hongbofu@cityu.edu.hk

J. Shang is with the Department of Computer Science & Engineering, HKUST. E-mail: jshang@cse.ust.hk

X. Han is with Shenzhen Research Institute of Big Data, Chinese University of Hong Kong, Shenzhen. E-mail: hanxiaoguang@cuhk.edu.cn

Fig. 1. Our Sketch2Human generates high-quality full-body images with respect to an input semantic sketch for geometry control and a reference image for appearance control. (a), (b), (c), and (d) correspond to four different sketch inputs.

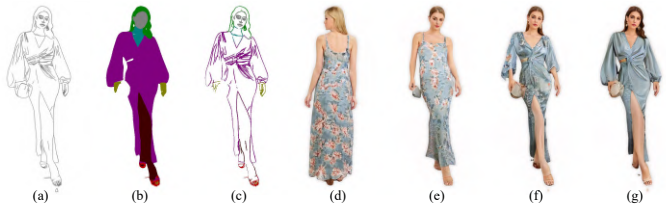

Fig. 2. The examples of inputs and style-mixing results. (a) a sketch input. (b) the corresponding parsing map. (c) the semantic sketch produced from (a) and (b). (d) an input appearance input. (e)-(g) results from mixing at layers 6, 8, and 10, respectively.

promote the disentanglement of the StyleGAN-Human without a real dataset, we propose to employ the style mixing to automatically generate appearance-transferred and geometry-preserved training data (see Figure 2 (e) and (g), respectively) that completely keep one constraint and associate it with the other constraint.

Specifically, we design a two-stage generation framework consisting of two main modules: *Sketch Image Inversion* and *Body Generator Tuning*. In Stage 1, we invert the input semantic sketch to a latent code via a sketch encoder. This encoder is trained with a large number of sampled images from StyleGAN-Human. To achieve accurate geometry inversion and avoid the influence of color and texture, we calculate sketch loss directly between the input sketches and the sketches extracted from the sampled images. Meanwhile, due to the sparsity of sketches, the sketch loss is not enough to help the encoder clearly distinguish each semantic part, especially for sketched humans with tight clothes. Therefore, we introduce a semantic loss. In Stage 2, we fine-tune the pre-trained StyleGAN-Human with the appearance-transferred and geometry-preserved results, respectively. We leverage the style loss to learn the fabric feature from the appearance-transferred results and design a content loss using the geometry-preserved results to avoid geometric changes caused by the appearance-transferred results. We still leverage the above semantic loss to enhance geometry preservation. The large amount of synthesized data generated from the original StyleGAN-Human enables the full disentanglement between geometry and appearance.

Extensive experiments show that our method achieves flexible and disentangled control of geometry and appearance. Quantitative and qualitative comparisons prove that our Sketch2Human outperforms the related techniques for full-body image generation. We also demonstrate the robustness of our method against sketches of different styles by professionals and amateurs.

II. Related Work

In this section, we review the existing techniques closely related to our method, including full-body human image synthesis and sketch-based generation and editing.

A. Full-body Human Image Synthesis

In recent years, 2D human generation has been widely investigated due to the rapid development of deep generative models. The existing solutions can be roughly divided into unconditional and conditional methods. Unconditional methods aim to generate high-fidelity and diverse images from random noises. For example, from a data-centric perspective, StyleGAN-Human [10] collects an SHHQ dataset to train a StyleGAN for the entire human body. InsetGAN [7] combines multiple pre-trained GANs, each focusing on different parts (e.g., faces, shoes), which can be seamlessly merged into a full-body person image. UnitedHuman [27] uses multi-source datasets for different body regions to jointly learn a human generative model. Although the above methods generate high-resolution and realistic images, they lack explicit control of geometry or appearance.

In contrast, conditional methods focus on controllable generation via various conditions (e.g., pose, reference images, text, etc.). *Pose-conditioned* methods [3], [28], [29] mainly depend on a canonical coordinate system of a 3D human

body (with UV parameterization) to directly establish the correspondence between pixels at each pose and then leverage a generator supervised with multi-view human datasets to refine coarse warped results. *Virtual try-on* methods [2], [4], [16]–[18] conditioned on reference images, aim to transfer garments in a reference person image to an input source person. They disentangle clothes from human identity at a segmentation generation stage and then warp clothes via a clothes deformation module to a target person. Compared with the pose-conditioned methods, virtual try-on methods further control the garment shapes and appearance via the reference images. However, they are inflexible and inaccurate for simultaneous control of geometry and texture. To facilitate intuitive control for layman users, text-conditioned methods [15], [30], [31] utilize an input text to constrain the shapes or textures of clothes. Specifically, Text2Human [30] first outputs a human parsing map from a given human pose and then synthesizes results guided by a text look-up code from a hierarchical texture-aware codebook. Similar to our work, FashionTex [15] and StyleHumanCLIP [31] leverage a pre-trained StyleGAN-Human model to enable human editing through the manipulation of a latent vector in the $W+$ space, guided by input text. Nonetheless, these approaches are limited in their inability to precisely control both geometry and appearance simultaneously.

To achieve a more explicit and flexible geometry control, we adopt a semantic sketch as a geometry representation, which provides a concrete description and can be freely drawn and modified by novices and professionals. Different from the above virtual try-on methods, which use reference images to influence the geometry and appearance of generated images, our method uses a reference image to achieve appearance control, independent of geometry control by an input semantic sketch.

B. Sketch-based Image Generation and Editing

Since it is easy for sketches to depict global geometry (shape contours) and local geometry (e.g., wrinkles), sketch-based image generation and editing have been well explored. These methods have mainly focused on the human face domain. For example, DeepFaceDrawing [32] and DeepFaceEditing [22] leverage a local-to-global strategy: they first model separate features for each key face component and then recombine them together. Wu et al. [33] attempt to apply a local-to-global strategy to sketch-based human body generation directly. However, their approach still cannot achieve realistic results due to a larger variety of poses, shapes, and garments. In addition, their approach lacks any appearance control. Benefiting from the large text-to-image (T2I) models ([23], [24]), [25], [26] learn different additional paths with geometry control inputs (e.g., sketch, pose, depth) to extract guidance features and then add them to the pre-trained T2I models. [13], [19] directly concatenate the spatial controls with the original noises and finetune the pre-trained T2I models in a self-supervised manner. However, [13], [25] lack detailed control of appearance with text and [19], [26] fail to transfer the appearance of reference images faithfully.

Exemplar-based methods [34]–[36] are designed for controllable (geometry, appearance) image translation tasks and have been tested on multiple datasets (e.g., faces, indoor images). They use image retrieval to find appearance exemplars for training, thus alleviating the lack of a disentangled dataset, including pairs with the same appearance but varying geometry. They mainly learn a dense correspondence map between an input sketch and a reference image. The learned correspondence might be inaccurate when large pose and shape differences exist between the sketch and reference image. Hence, such methods are unsuitable for generating human bodies with high complexity and diversity in pose, body shape, and fashion elements.

Recently, based on the layerwise representative of StyleGAN, embedding-based methods [37], [38] learn encoders to invert inputs to the corresponding latent space or latent features of StyleGAN. StyleGAN-Human is designed specifically for the full-body human domain. Regrettably, due to incomplete disentanglement, it lacks control over clothing patterns with geometry unchanged. Since we would like to leverage the prior of StyleGAN, we also embed sketches to the latent space. To achieve accurate geometric embedding of complex human images, we train the encoder directly supervised by sketches and semantics rather than RGB images. Additionally, we synthesize appearance-transferred and geometry-preserved data to make StyleGAN-Human achieve full disentanglement.

III. Method

For an input semantic sketch image PS_g and an appearance image I_a, our Sketch2Human aims to synthesize a high-fidelity result I_{syn} aligned with PS_g, while transferring the color and texture of face, skin, and clothing in I_a, as illustrated in Figure 3 (Left). The semantic labels in our current implementation include nine categories: hat, hair, hand, glasses, garment, torso-skin, face, foot, and shoe. Users might input semantic sketches stroke by stroke, with each stroke associated with a specific semantic label (see Supplementary Materials Section IV-A).

As illustrated in Figure 3 (Right), our system consists of two main modules, namely, Sketch Image Inversion (Section III-A) and Body Generator Tuning (Section III-B). To produce photo-realistic full-body human images from sketches with different styles, the first module focuses on inverting an input semantic sketch PS_g to a geometry latent code $\hat{w}_g$, located in the $W+$ space. Compared with the W space, the $W+$ space with 18 style vectors can represent more accurate details of the input. Due to the large variability of real human-body images, StyleGAN-Human fails to restore complex fabric patterns accurately when it roughly maintains geometry input (Figure 2 (f)). Note that a naïve approach by fine-tuning on the appearance image cannot obtain satisfactory results due to the large variance in the pose and shape (Section IV-E2). Therefore, we propose to use a novel training strategy to fine-tune the generator in the second module.

A. Sketch Image Inversion

At this stage, based on the pre-trained StyleGAN-Human $G(w;\theta)$, we train a sketch encoder responsible for embedding

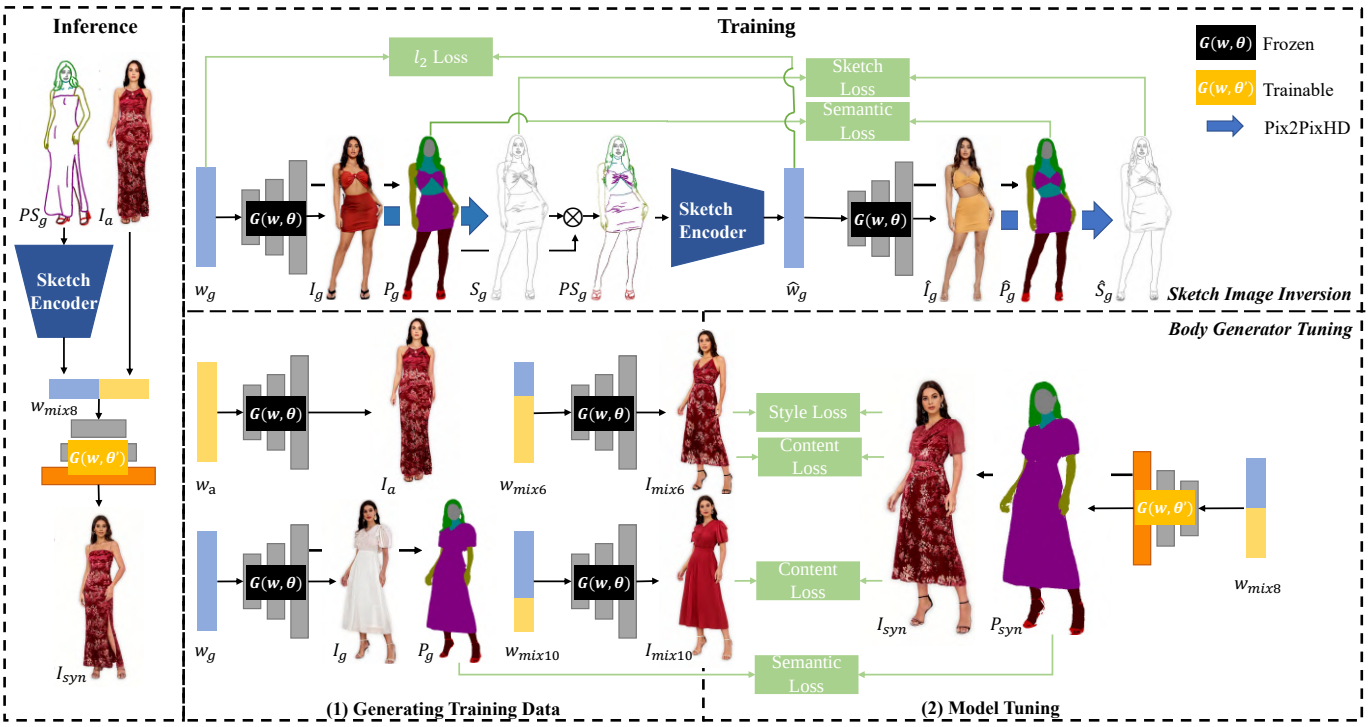

Fig. 3. An illustration about the training (right) and inference (left) pipelines of our method for full-body human image generation conditioned on a semantic sketch PS_g and a reference image I_a. The training pipeline consists of two main modules: Sketch Image Inversion (right-top) and Body Generator Tuning (right-bottom). In the Sketch Image Inversion module, we first sample a latent code w_g to generate the training triplet (semantic sketch PS_g, parsing map P_g, sketch S_g). Then, we use these data to train a sketch encoder. In the Body Generator Tuning module, given an appearance code w_a, we also sample a latent code w_g to prepare the training appearance-transferred I_{mix6} and geometry-preserved I_{mix10} samples via style mixing at different layers. Then, we use them to fine-tune the generator $G(w;\theta^{'})$. During inference, the sketch encoder first embeds PS_g into a latent code and mixes it with the appearance code derived from I_a to form w_{mix8}. Given w_{mix8}, $G(w;\theta^{'})$ produces the final result I_{syn}.

the geometry information of PS_g as comprehensively as possible. Although an input sketch could be converted to multiple latent codes with the same geometry and different appearance, the geometric parts of those multi-latent codes are the same. Thus, the sketch encoder only focuses on accurate geometry one-to-one mapping (sketch to the geometric part). Specifically, it embeds an input semantic sketch PS_g to a latent code $\hat{w}_g \in 18 \times 512$, as shown in Figure 3. During training, PS_g comes from a combination of an automatically generated sketch S_g and a parsing map P_g. During inference, users interactively provide semantic strokes PS_g, as shown in the accompanying video. Since the feature pyramid structure of Encoder4Editing (e4e) encoder [39] makes it possible to encode the input details, we adopt its structure for our encoder. StyleGAN-Human cannot fully capture the distribution of complex real full-body images because such images might be out-of-distribution for it. If trained with those real images, the encoder would introduce unreasonable structures into $\hat{w}_g$. Then such artifacts would be displayed after the style-mixing (Figure 5). To avoid this issue, we train the encoder with paired data $\{(w_g, I_g)\}$ sampled from the latent space of StyleGAN-Human.

1) Training: To provide the supervision directly on the sketches, we retrain Pix2PixHD [40] for the image-to-sketch generation task with the paired data $\{(I, S)\}$. The images $\{I\}$ are from the SHHQ dataset [10], and the corresponding sketches $\{S\}$ are extracted by the Sobel filter and the sketch simplification method [41]. The network structure and training process are unchanged. Then, Pix2PixHD generates sketches $\{S\}$ from input images $\{I\}$, as illustrated by the blue arrow in Figure 3. Inspired by [42], we add a semantic branch StyleGAN-Human to facilitate semantic label extraction. The detailed architecture is shown in Supplementary Materials Section I. Benefiting from the underlying semantic information in StyleGAN-Human, it is easy to learn such a semantic branch with sampled pairs $\{(w, I)\}$. Here, we provide the semantic label P from I using an off-the-shelf method [43]. After training, the generator $G(w;\theta)$ directly produces (I, P) for any input latent code w.

For the specific training procedure of the encoder, we randomly sample a latent code w_g and feed it into $G(w;\theta)$ to generate a synthetic full-body image I_g and a corresponding parsing map P_g. Meanwhile, Pix2PixHD produces the corresponding sketch image S_g from the input image I_g. As shown in Figure 3, we transfer the semantics from the parsing map P_g to the sketch image S_g via pixel-wise multiplying them together to get a semantic sketch PS_g, which is then projected to $\hat{w}_g$ via our sketch encoder. Finally, inputting the whole $\hat{w}_g$ to $G(w;\theta)$, the generator outputs a synthesized image $\hat{I}_g$ and a parsing map $\hat{P}_g$. We still use Pix2PixHD to produce a sketch image $\hat{S}_g$.

2) Objective Function: Based on the frozen generator $G(w;\theta)$, the goal of our sketch encoder is to reconstruct

the geometry information in S_g with a low error. To achieve this, we formulate the objective function from three aspects: (1) distribution consistency, (2) geometric accuracy, and (3) semantic consistency.

Distribution Consistency. To encourage the geometry codes follow the distribution of the $W+$ space, we first adopt the latent adversarial loss $\mathcal{L}_{adv_w}$ proposed in e4e [39]. We also use the l_2 loss $\mathcal{L}_{l_2}$, which is calculated between the inverted code $\hat{w}_g$ and the ground-truth code w_g. This simple way makes the encoder easy to project the inputs to the correct distribution, thus alleviating unreasonable embeddings.

Geometry Accuracy. Sketching inputs only contain geometric information. Our sketch encoder is not too concerned with the code related to the appearance for the high-resolution ($64^2 - 1024^2$) layers. Compared with calculating losses on RGB images and thus indirectly constraining geometry, it is more efficient to directly supervise on the sketches. Benefiting from the image-to-sketch model Pix2PixHD, we can get $\hat{S}_g$ from the generated image $\hat{I}_g$. Liu et al. [44] found that LPIPS [45] is effective for sketches due to its sensitivity for edges. Hence, we introduce the sketch loss $\mathcal{L}_{lpips}$, which aligns the features from VGG [46], given the respective inputs of a sampled sketch S_g and the corresponding generated sketch $\hat{S}_g$.

Semantic Consistency. Compared with RGB images, sketches are sparse and lack color information. Such an issue makes the encoder trained with sketches themselves easier to confuse various semantic parts (Section IV-E1), especially when the body and clothes share similar contours (e.g., due to wearing tight clothes). Hence, we incorporate a semantic loss to force the encoder to learn the semantic correspondence between the input sketch and the latent code. Specifically, the semantic loss consists of the pixel-wise cross-entropy loss $\mathcal{L}_{ce}$ and the dice loss $\mathcal{L}_{dice}$, which are generally used in the semantic segmentation task [47]. We calculate those two terms between the parsing maps P_g and $\hat{P}_g$ derived from the input sketch and the reconstructed result, respectively.

By combining the above loss terms, we define the final objective function as follows:

$$\begin{aligned}\mathcal{L}_{encoder} =& \lambda_1 \mathcal{L}_{adv_w}(\hat{w}_g) + \lambda_2 \mathcal{L}_{l_2}(w_g, \hat{w}_g) + \lambda_3 \mathcal{L}_{lpips}(S_g, \hat{S}_g) \\ &+ \lambda_4 \mathcal{L}_{ce}(P_g, \hat{P}_g) + \lambda_5 \mathcal{L}_{dice}(P_g, \hat{P}_g).\end{aligned} \tag{1}$$

In our experiments, we set $\lambda_1 = 0.1$, $\lambda_2 = 10$, $\lambda_3 = 2$, $\lambda_4 = 1$, and $\lambda_5 = 1$.

B. Body Generator Tuning

Our body generator follows a standard structure of StyleGAN-Human. The model has two conditions (i.e., the inputs): a geometry code w_g to provide geometry constraints and an appearance code w_a to provide appearance constraints. StyleGAN is composed of multiple style blocks, and each block is responsible for generating a specific level of detail in the image. These style blocks are associated with latent codes that control the different attributes. In style-mixing, different blocks can be assigned distinct codes. This allows for combining styles from multiple codes, resulting in diverse generated images. StyleGAN-Human observes that the high-layer styles (9-18) control the clothing color, middle-layer styles (5-8) control the clothing type and human face identity (geometry and appearance), and low-layer styles (1-4) control the pose. Hence, we follow the mixing rules between the high layers and the rest layers. The model $G(w;\theta')$ combines the geometry and appearance information by mixing these two codes at layer 8, denoted as w_{mix8}, to synthesize the result I_{syn}. As shown in Figure 2 (f), applying the mixed code w_{mix8} to the original generator $G(w;\theta)$ produces an image that slightly changes the geometry from S_g and only exhibits the color but not texture from I_a. To faithfully align the geometry and restore the texture, for each appearance image, we first generate a corresponding training dataset (Section III-B1) and then fine-tune the original generator $G(w;\theta)$ to a new body generator $G(w;\theta')$, which achieves a more complete disentanglement (Section III-B2). The key idea is to leverage the style mixing of an unconditional GAN to automatically synthesize the training data.

1) Generating Training Data: Intuitively, fine-tuning with a single appearance image I_a using a style loss [45] can help the generator learn the appearance. However, we found several issues with such a naïve approach (Section IV-E2). First, the model trained with the style loss tends to learn the mean textures and sometimes blends the textures from different semantic parts. Second, the appearance of clothes is pose-dependent, but transferring the style from this single appearance image I_a just copies patterns under the original pose. Meanwhile, it is challenging to build a dataset with different geometry for the same appearance.

We observe that by copying the geometry code for layers (1-6) and the appearance code for the rest layers, the results restore the reference appearance well with the pose indicated in the input geometry code but does not faithfully respect the detailed geometry information (e.g., garment change in Figure 2 (e)). We call such results the *appearance-transferred* results. On the other hand, when inputting the geometry code for layers (1-10), the results preserve the global and local geometry well but only gain the color without any fabric pattern from the appearance code (Figure 2 (g)). We thus refer to such results as the *geometry-preserved* results. Although the two kinds of style-mixing strategies we find above fully maintain only geometry or appearance information, they can help the generator achieve a comprehensive disentanglement of geometry and appearance.

Hence, we first sample a large number of latent codes $\{w_g\}$ from the latent space of the generator $G(w;\theta)$ to represent different geometry. Then, for each appearance input I_a, we generate the synthesized appearance-transferred $\{I_{mix6}\}$ and geometry-preserved $\{I_{mix10}\}$ pairs by feeding the pre-trained generator $G(w;\theta)$ with the mixed codes between the appearance code w_a and geometry codes $\{w_g\}$ at layers 6 and 10 separately (Figure 3). The representative training examples are shown in Figure 4.

2) Model Tuning: The goal of our body generator is to restore the whole appearance information (color and texture) on the premise of preserving the geometry information defined in the latent code $\hat{w}_g$ obtained in Section III-A. We freeze

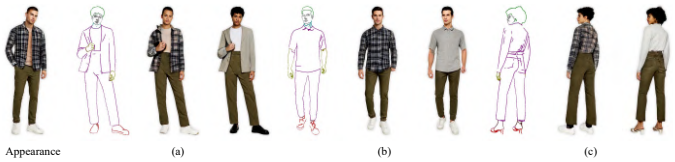

Fig. 4. Three examples of prepared data. Each example in (a)-(c) shows an input semantic sketch and its corresponding appearance-transferred and geometry-preserved results.

the affine transformation layers since we need to utilize the original latent space. Additionally, Alaluf et al. [48] find that altering the *toRGB* layers harms the editing capabilities of GAN. Therefore, we only tune the *non-toRGB* convolution layers.

For appearance learning, as illustrated in Figure 3, we first incorporate the style loss $\mathcal{L}_{style}$ [45] between each semantic region i of the generated image I_{syn} and target image I_{mix6} via computing the Gram matrix for the features extracted by VGG [46]. Those regions are extracted from the corresponding parsing maps P_{syn} and P_{mix6}. Note that the generated images $\{I_{syn}\}$ and $\{I_{mix6}\}$ are not entirely aligned in terms of geometry, especially for the clothing shapes. Hence, we calculate a content loss $\mathcal{L}_{content_6}$ [45] which involves spatial features within each semantic union region via a union mask M. M_i refers to the overlap regions between semantic label i of P_{mix6} and P_{syn}. The extraction of the M is illustrated in Supplemental Materials Section II.

However, only using the above loss harms the preservation of the geometry information since respecting the image I_{mix6} more or less changes the geometry defined by the input sketch. It is important for our body generator $G(w;\theta^{'})$ to promote the ability to restore the geometric requirements. Hence, we propose geometric constraints from three aspects. First, since the different layers of StyleGAN-Human control different visual attributes, we update the (9-18) convolution layers related to the appearance code, leaving the (1-8) convolution layers related to the geometry unchanged. Such a constraint can not only preserve the geometry prior but also reduce the training parameters, thus saving training time.

Second, benefiting from the semantic branch mentioned in Section III-A, we get the paired parsing maps with the synthesized results simultaneously. When inputting the whole w_g, the generator provides the corresponding parsing map P_g. Similarly, we can get P_{syn} with the input of w_{mix8}. Then, we only measure the dice loss $\mathcal{L}_{dice}$ [47] between them, since we find that the additional *ce* loss here is harmful to the visual quality.

Third, semantic supervision only focuses on maintaining the global geometry but cannot preserve local features (e.g., wrinkles), thus decreasing the generation quality. Therefore, we calculate the high-level content loss $\mathcal{L}_{content_{10}}$ to encourage the output image I_{syn} to be perceptually consistent with the geometry-preserved image I_{mix10} in terms of the image content and spatial structure.

Method	CD ↓	mIoU ↑	FID ↓	IS ↑
pSp	4.66	0.72	41.38	2.69
e4e	5.06	0.76	22.64	2.73
CoCosNetV2	**1.07**	**0.83**	39.09	2.71
T2I-Adapter	3.22	0.74	56.73	**3.74**
Ours	4.95	0.80	**22.04**	3.05

TABLE I
QUANTITATIVE COMPARISON BETWEEN OUR SKETCH2HUMAN AND FOUR RELATED METHODS ON THE DEEPFASHION DATASET [52].

The final loss objective is defined as follows:

$$\begin{aligned}\mathcal{L}_{decoder} =& \lambda_6 \sum_{i=1}^{n} \mathcal{L}_{style}(P_{mix6,i} \cdot I_{mix6}, P_{syn,i} \cdot I_{syn}) \\ &+ \lambda_7 \sum_{i=1}^{n} \mathcal{L}_{content_6}(M_i \cdot I_{mix6}, M_i \cdot I_{syn}) \\ &+ \lambda_8 \mathcal{L}_{content_{10}}(I_{mix10}, I_{syn}) + \lambda_9 \mathcal{L}_{dice}(P_g, P_{syn}),\end{aligned} \tag{2}$$

where $\lambda_6 = 1000$, $\lambda_7 = 0.6$, $\lambda_8 = 1$, and $\lambda_9 = 30$. n is the number of semantic labels.

IV. EXPERIMENT

In this section, we have done extensive experiments from five aspects, namely, baseline comparison (Section IV-A), face body montage with InsetGAN (Section IV-B), real appearnce image (Section IV-C), user study (Section IV-D) and ablation study (Section IV-E). For the quantitative comparison, we leverage the symmetric *Chamfer Distance* (CD) [49] and *Mean Intersection over Union* (mIoU) to evaluate the geometry alignment with the sketch input. The visual quality is measured by the $Fréchet\ Inception\ Distance$ (FID) [50] and Inception Score (IS) [51]. For the user study, each participant was asked to consider four aspects, including the accuracy of geometry preservation (GP), appearance transfer (AT), visual quality (VQ), and user preference (UP), and then choose the best one for each aspect. The details about system implementation and evaluation metrics can be found in Supplementary Materials Section II.

A. Baseline Comparison

1) Qualitative Comparison: To the best of our knowledge, no prior work generates full-body human images conditioned on geometry and appearance images. Hence, we first repurpose the related image generation methods with the above two inputs, including pSp [37], e4e [39], CoCosNetV2 [34], and T2I-Adapter [26] for our controllable full-body generation task. The evaluation details of the above methods and text prompts for T2I-Adapter are in Supplementary Materials Section III-A.

Geometry Quality. Figures 5 and 6 show the results from fine and coarse sketches separately. pSp can basically capture the global geometry of sketch images varied in abstraction. Still, this approach tends to add or lose semantic parts, resulting in salient artifacts (e.g., glasses in Figure 5 (a) and hat in Figure 5 (d)). Additionally, the local geometry is rough since several strokes are ignored (see the pSp results in Figure 6 (c) and (f)). In addition, the geometry code from pSp sometimes affects subsequent color expression

Fig. 5. Qualitative comparisons between our method and four related sketch-based methods. Our method shows the best geometry and appearance transfer results. The sketch images are extracted from the DeepFashion dataset. The appearance images sampled from the StyleGAN-Human include pure color and texture images. Pure color images denote garments containing one or more colors without fabric patterns (a)-(b), while texture images include both (c)-(f).

(Figure 5 (b) and Figure 6 (e)). This is mainly because the predicted embeddings might be out-of-distribution. The e4e can effectively infer latent codes that belong to a latent space distribution because of the latent discriminator. So, the visual quality of e4e's results is relatively high. However, their results might fail to capture the accurate postures (see the e4e results in Figure 5 (a)) and local details (see Figure 5 (b) and (d)) possibly due to the co-learning of the geometry and appearance. CoCosNetV2 completely aligns the geometry of results with the sketches. Thus, it fails to synthesize realistic full-body images from coarse sketches with inaccurate human proportions and shapes (see the CoCosNetV2 results in Figure 6). T2I-Adapter generates satisfactory results even with rough input sketches since it leverages an adjustable weight to control the geometry consistency. Nevertheless, with the almost full geometric alignment, they produce unrealistic faces (see the T2I-Adapter results in Figure 6 (a) and (d)) and fail to restore the real body proportion (see Figure 6 (c) and (e)), which is roughly depicted in the input sketches. Benefiting from the large sampled data and the explicit supervision of geometry and semantics in our network, our results better preserve the global geometry (poses, clothing types) and synthesize each semantic element more faithfully.

Appearance Quality. As for the appearance transfer, pSp, e4e, and CoCosNetV2 can only transfer the primary colors of the appearance inputs. The reason for pSp and e4e is that the high layers of StyleGAN-Human can only control the colors. CoCosNetV2 trained with the DeepFashion dataset fails to build accurate dense correspondence when the sketch input does not depict the texture boundaries. T2I-Adapter regards the appearance input as a style, so it cannot accurately restore appearance details, and their results tend to be non-photorealistic (see the T2I-Adapter results in Figure 6 (a) and (d)). As embedded in the latent space of StyleGAN-Human, our method can generate high-fidelity results from different degrees of sketches. Meanwhile, thanks to altering the generator with the roughly appearance-transferred data, our method can best transfer color and texture simultaneously from the reference images.

2) Comparison with Human Image Generation Methods: To prove the flexibility and detailed control of our inputs, we also compare our method with full-body image generation methods with different inputs, including PWS [3], NTED [5], Text2Human [30], and ControlNet [25]. Although the modalities of their inputs are different, we derive the inputs required by different methods from the same RGB images. The details for input extraction and text prompts for ControlNet are in Supplementary Materials Section III-B.

As shown in Figure 8, our method achieves the highest fidelity. PWS can only change the human pose in a given appearance image, while NTED aims to transfer the garments in the appearance image to the human in the geometry image. Besides a pose and specific cloth, our method can also change the hairstyle, garment shape, etc. For the virtual try-on effect similar to NTED, it is vital to retain the face completely (i.e., both geometry and appearance) of the geometry input image, while our method is designed to retain only the geometry of the geometry image (Figure 8 Column *Ours*) since we assume the appearance is from the reference image. Our method could potentially achieve this application via drawing a similar clothing type with the reference image and directly swapping the face of the geometry image with our result (e.g., by using InsetGAN) (Figure 9 (a)). Text2Human and ControlNet take parsing maps and canny maps as input, respectively. They achieve satisfactory geometry preservation. However, the appearance of these two methods is mainly controlled by text.

Fig. 6. Qualitative comparisons between our method and three related methods (without e4e since it takes as input RGB images). Our method shows high-fidelity results with the best geometry and appearance consistency. The sketch images are collected from users.

Fig. 7. Qualitative comparisons with two diffusion-based methods with the decreasing sketch conditioning weight. The text prompt for the two compared methods is "a woman wears a dress with salmon color".

Text2human generates the results with a completely irrelevant appearance since it supports only five types of appearance text (e.g., stripe, plaid). Based on a large text-to-image model, ControlNet produces more specific results from sufficient text prompts. However, compared with a given appearance image, a text prompt is still too coarse to describe detailed colors and textures.

3) Quantitative Comparison: Table I shows the quantitative comparison results. Since the results of CoCosNetV2 are strictly aligned with the input sketches (Figures 5 and 6), it achieves the highest CD and mIoU. Here we set the weight of T2I-Adapter to 1, so it produces the comparable CD. However, their method might ignore several strokes located in the semantic boundary (e.g., Figure 5 (d), (f) and Figure 6 (e), (f)), resulting in a low mIoU. Since T2I-Adapter is based on a large text-to-image model, its results tend to exhibit more diversity according to reference images, achieving the highest IS. However, its results fail to preserve the color and texture in the target references. Additionally, CoCosNetV2 and T2I-Adapter lack realism, as reflected in their high FID scores. The realism tends to be worse when the input sketches get rougher (Figure 6). For a fair comparison between the embedding-based methods, we use the same generator, i.e., StyleGAN-Human, for testing. For FID, our method has achieved a similar result to e4e, proving that our encoder trained with the sampled data does not affect the fidelity of the generator. The CD of pSp is lower than us, but this is inconsistent with our observations of more serious artifacts with their results, as shown in Figure 5. Meanwhile, the highest FID of pSp also reflects the lowest quality of synthesized images. Our method significantly outperforms the other four methods in terms of FID with comparable IS and mIoU. This indicates that our results are not only the most similar to the sketch inputs but also of the highest image quality. This is consistent with our findings based on the qualitative comparisons.

B. Face Body Montage with InsetGAN

Since our method combines the geometry and appearance information from two inputs, the face identity (involving both geometry and appearance) of the reference image might be lost. However, several tasks (e.g., virtual try-on, human image editing) require preserving human face identity. Benefiting from InsetGAN [7], which introduces a multi-GAN optimization method, our pipeline can be easily combined with the state-of-art face model [53] to achieve face identity preservation. To keep the identity of the reference image or geometry image (used to extract a sketch input), we first invert the corresponding face to the latent space of the pretrained FFHQ model [53] and then iteratively optimize the face latent codes and body latent codes by the FFHQ [53] and our generator, respectively. Figure 9 shows the results of combining geometry (Column *Ours Body + Geometry Face*) or appearance (Column *Ours Body + Appearance Face*) faces with bodies

Fig. 8. Qualitative comparisons of our method with four human image generation methods. Our method demonstrates the most flexible and detailed control in both geometry and appearance.

generated by our full-body generator. After optimizing both latent codes, we successfully obtain composition results that maintain coherence and preserve the identity.

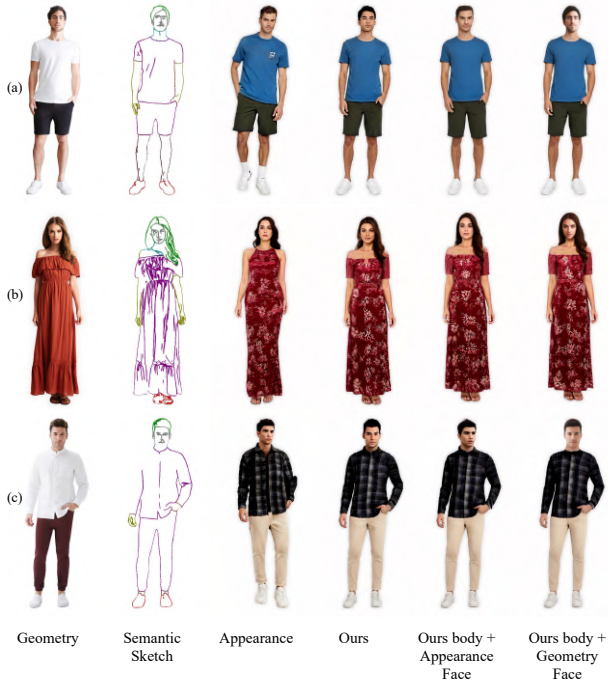

Fig. 9. Three face body montage results using InsetGAN. Given target geometry faces/appearance faces and bodies generated by our generator with the input of semantic sketches (extracted from geometry images) and appearance images, InsetGAN jointly optimizes the corresponding latent codes to achieve coherent results.

C. Real Appearance Image

For a real appearance image, we first need to encode it into the latent space of StyleGAN-Human as w_a using an off-the-shell encoder [39]. As Figure 10 shows, with the accurately inverted latent codes, our method successfully transfers the appearance of the real images.

Fig. 10. The results with the inputs of a semantic sketch and a real appearance image. To achieve them, we first invert the real appearance image to the latent space of StyleGAN-Human via an off-the-shell encoder [39].

D. User Study

We first prepared a set of input appearance images (50 in total) with diverse styles randomly picked from StyleGAN-Human, and then applied the four compared methods (Section IV-A1) to each input appearance image with a semantic sketch image (25 randomly selected from our test set and 25 from freehand sketches (Supplementary Materials Section II)). The evaluation was done through an online questionnaire. There were, in total, 50 participants. We showed each participant

Method	GP(%)	AT(%)	VQ(%)	UP(%)
pSp	5.2	7.3	6.6	6.3
e4e	11.0	13.9	15.7	15.6
CoCosNetV2	31.6	5.9	6.9	7.6
T2I-Adapter	12.2	4.5	4.5	4.5
Ours	**40.0**	**68.4**	**66.3**	**66.0**

TABLE II
SUMMARY OF THE VOTING RESULTS FROM THE USER STUDY WITH THE INPUT SKETCHES EXTRACTED FROM THE DEEPFASHION DATASET.

Method	GP(%)	AT(%)	VQ(%)	UP(%)
pSp	6.6	9.4	10.8	9.4
CoCosNetV2	28.8	2.8	3.1	3.1
T2I-Adapter	10.4	4.5	4.5	4.2
Ours	**54.2**	**83.3**	**81.6**	**83.3**

TABLE III
SUMMARY OF THE VOTING RESULTS FROM THE USER STUDY WITH THE INPUT SKETCHES COLLECTED FROM USERS (SUPPLEMENTARY MATERIALS SECTION II) VIA OUR INTERFACE.

all the compared results in random order with the sketch and appearance inputs, set by set. Noted that we exclude e4e since its input is an RGB image, but there is no such images for those user-drawn sketches. Tables II and III show the statistics of the voting results from the different style sketches. No matter what level of abstraction the input sketches were, our method was most chosen for every aspect. Therefore, human preference proves that with diverse inputs, our method maintains geometry and texture well and generates high-fidelity results.

E. Ablation Study

We report some results of the ablation study of our method in terms of the effectiveness of the Sketch Image Inversion module (Section III-A) and the Body Generator Tuning module (Section III-B). We use semantic sketches as input for the ablation study.

1) Effectiveness of Sketch Image Inversion: Since the StyleGAN-Human cannot fully capture the distribution of complex real full-body images, the encoder supervised with those real images results in inaccurate geometry and artifacts (Figures 5 and 6). We adopt the sampled data to train our Sketch Image Inversion module with a series of loss functions. We evaluate the effectiveness of each loss by testing the performance of our model trained in various ways, including trained only with the full supervised loss ($w/\mathcal{L}_{l_2}$), trained with the self-supervised loss ($w/\mathcal{L}_{lpips}$), trained with both of them ($w/\mathcal{L}_{l_2} + \mathcal{L}_{lpips}$). All the experiments are implemented with $\mathcal{L}_{adv_w}$.

As shown in Figure 11, only the results supervised with all the proposed losses (*ours*) achieve consistent geometry (e.g., poses, clothing types) with the sketch inputs. While the model trained with $\mathcal{L}_{l_2}$ can synthesize high-quality results (Column $w/\mathcal{L}_{l_2}$), these results do not faithfully respect the input sketches. For example, cases ((a) and (e)) change the head pose, and cases ((b) and (c)) produce undesirable clothing contours. It is mainly because full supervision can stably predict the geometric code conforming to the distribution, thus ensuring the generation quality, but small deviations in the latent space would be amplified and reflected in the geometric shape. The encoder supervised with $\mathcal{L}_{lpips}$ provides more accurate poses and shapes (Column $w/\mathcal{L}_{lpips}$) but introduces artifacts on the clothing ((a) and (b)) and influences the color representative ((c)), possibly due to the strong geometry constraint to ignore the original distribution. Hence, applying both of them can solve the above issues to a certain extent (Column $w/\mathcal{L}_{l_2} + \mathcal{L}_{lpips}$). However, due to the sparsity of sketches, some line segments can easily be ignored at the training stage. That leads to the unstable prediction of semantic elements ((a)), especially for those (e.g., glasses, hats) with a low probability of occurrence in the training data and inaccurate clothing boundaries, especially when clothing and body contour tend to be consistent ((c)). To solve this problem, we explicitly add the semantic supervision $\mathcal{L}_{ce}$ and $\mathcal{L}_{dice}$ so that the encoder can focus on the accuracy of semantic objects separately.

Method	CD ↓	mIoU ↑
$w/\mathcal{L}_{l_2}$	7.06	0.68
$w/\mathcal{L}_{lpips}$	**3.95**	0.75
$w/\mathcal{L}_{l_2} + \mathcal{L}_{lpips}$	4.98	0.77
Ours	4.95	**0.80**

TABLE IV
QUANTITATIVE RESULTS OF THE ABLATION STUDY FOR THE SKETCH IMAGE INVERSION MODULE ON THE DEEPFASHION DATASET.

According to the quantitative metrics provided in Section IV-A3, we show the quantitative results of the ablation study for the Sketch Image Inversion module in Table IV. The baseline model with $\mathcal{L}_{lpips}$ gets the highest CD. This is because the artifacts inside the generated images can heavily influence the CD metric. Our full model achieves the best performance on the mIoU metric. Meanwhile, removing any components degrades the model's performance integrally.

2) Effectiveness of Body Generator Tuning: Since the high-resolution layers of the existing StyleGAN-Human fail to control the specific textures, we propose the Body Generator Tuning module to alter the weights. To validate the efficiency of this module, we compare our method with its variants from the perspective of loss terms, including supervised with a single appearance input (w/I_a), supervised with multiple appearance images after mixing at layer 6 (w/I_{mix6}), and supervised with multiple appearance images after mixing at layer 6 and the semantic loss ($w/I_{mix6} + \mathcal{L}_{dice}$). We only fine-tune the high-resolution layers for the above experiments. We also implement the full-weight updates (Full-layer).

As shown in Figure 12, it is obvious that supervised with a single appearance input, the results exhibit mean textures but lack the global and local structure distribution (Column w/I_a). Those textures do not vary with the poses of the input sketches ((c) and (e)). Additionally, for cases with multiple textures of different clothing, the results tend to mix textures ((b) and (e)). The possible reason is that the optimization with a single appearance image easily falls into a local optimum. Therefore, we propose to fine-tune the generator with the style-mixing results I_{mix6}, thus improving the quality of textures (Column w/I_{mix6}). However, the body and clothing shapes change a lot ((b) and (f)) since I_{mix6} is not completely aligned with the geometric requirements. By comparing Columns w/I_a and $w/I_{mix6} + \mathcal{L}_{dice}$, we can see that this issue is greatly solved after incorporating semantic

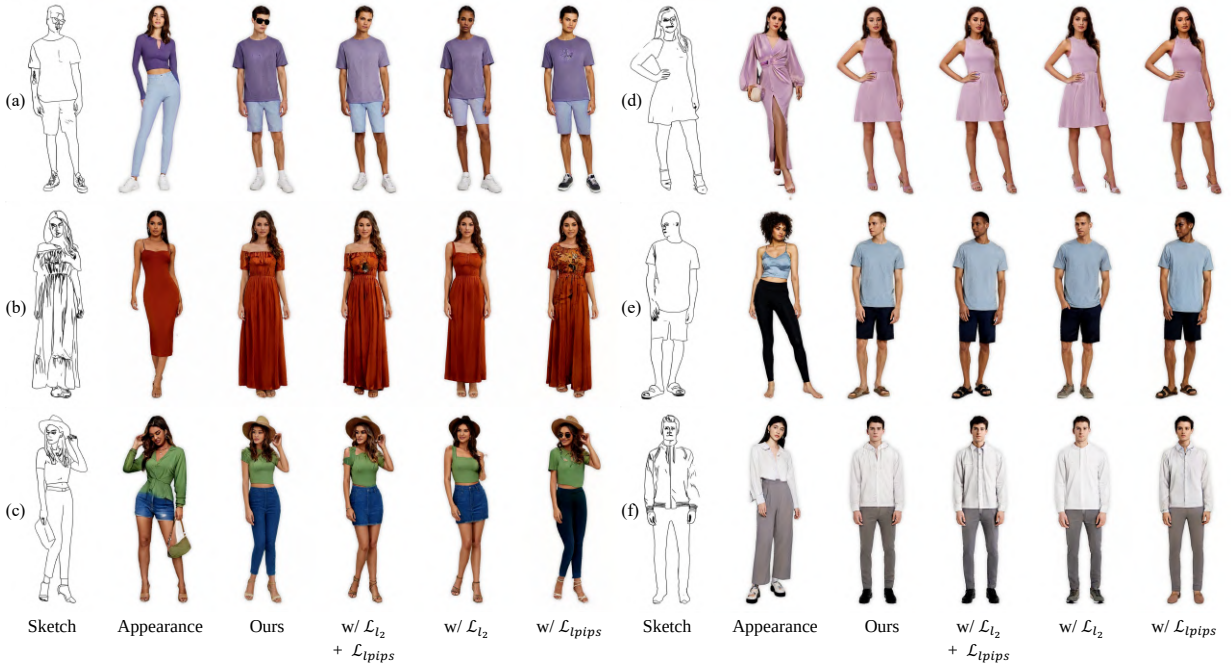

Fig. 11. The ablation study for the Sketch Image Inversion module.

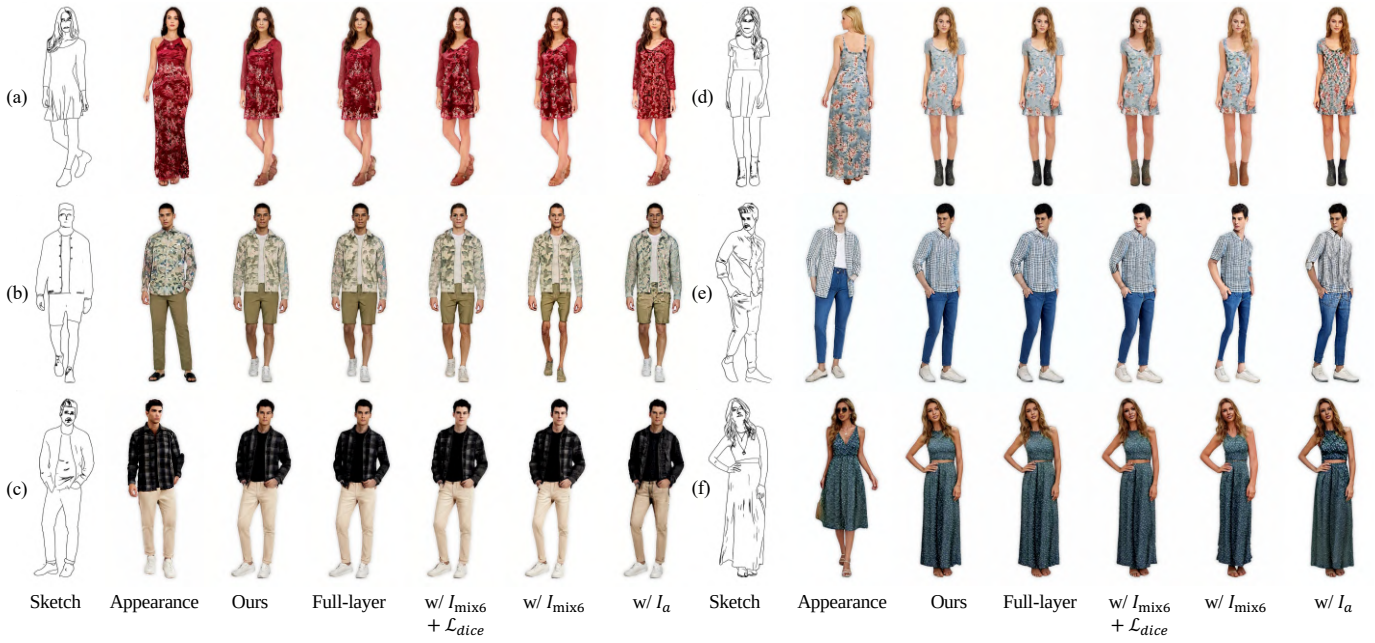

Fig. 12. The ablation study for the Body Generator Tuning module.

supervision. However, such semantic supervision leads to the emergence of non-photorealistic faces (see (b) and (f) in Column $w/I_{mix6} + \mathcal{L}_{dice}$). In other words, this supervision affects the fidelity of the generator. As shown in Column *Ours*, by adding the content loss with I_{mix10}, the model avoids such artifacts. Although there is a slight difference in the outer contour between full-layer updates and high-resolution-layer updates (see (a) and (e) in Column *Full-layer*), only fine-tuning the high-resolution layers can decrease the number of training parameters, thus slightly reducing the time required for fine-tuning.

V. Conclusion and Limitations

We presented Sketch2Human, the first system for controllable human full-body image generation given sketch and appearance constraints. The first stage of our system inverts the sketch input to the latent space. To achieve high-precision geometric encoding, we proposed to train our sketch encoder with infinitely sampled images and directly calculate the loss on the sketch level. Its second stage aims to enhance

the texture expressiveness for the existing StyleGAN-Human under different postures and shapes. Due to the lack of a proper dataset, we proposed two style-mixing strategies to synthesize the data that well preserve the texture and geometric information, respectively. Although they are changed on the other side, they can be used as good guidance together. Hence, we utilized these results to fine-tune the generator. Extensive experiments have demonstrated that our Sketch2Human generates results preserving both geometry and texture/color consistency with the two inputs and outperforming the four compared methods. The flexible control of inputs and attractive results from the applications (Supplementary Materials Section IV) show the practicality and significance of our method.

Our method can still be improved in various ways. First, since we choose to embed the sketch input into the latent space, our method prefers to produce reasonable results but sometimes ignores the user's intent (Figure 13 (b)). This problem could be solved by embedding the spatial guidance for the generator. Meanwhile, in future work, we intend to investigate a controllable tradeoff mechanism that effectively balances fidelity and insensitivity towards sketches provided by both novice and professional users. Second, for real appearance images, our method is influenced by the appearance codes from the image inversion method [39]. For those with complex textures, the corresponding textures cannot be transferred due to the inaccurate appearance codes and the limited generative power of StyleGAN-Human (Figure 13 (a)). Additionally, our method relies on the style mixing results of StyleGAN-Human. Such results fail to keep the hairstyles stable and miss the appearance of the shoes and glasses. These issues might be alleviated or addressed by more powerful full-body StyleGAN.

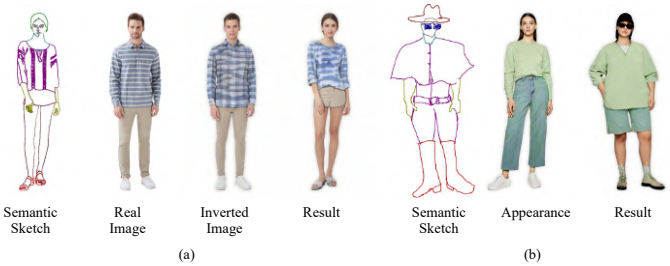

Fig. 13. Two less successful cases. (a) covers an example of transferring the appearance of a real image with e4e inversion. (b) shows an unsatisfactory result, which respects the training image distribution but not the creative design.

REFERENCES

[1] H. Dong, X. Liang, Y. Zhang, X. Zhang, X. Shen, Z. Xie, B. Wu, and J. Yin, "Fashion editing with adversarial parsing learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8120–8128.

[2] X. Han, X. Hu, W. Huang, and M. R. Scott, "Clothflow: A flow-based model for clothed person generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 471–10 480.

[3] B. Albahar, J. Lu, J. Yang, Z. Shu, E. Shechtman, and J.-B. Huang, "Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–11, 2021.

[4] X. Dong, F. Zhao, Z. Xie, X. Zhang, D. K. Du, M. Zheng, X. Long, X. Liang, and J. Yang, "Dressing in the wild by watching dance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3480–3489.

[5] Y. Ren, X. Fan, G. Li, S. Liu, and T. H. Li, "Neural texture extraction and distribution for controllable person image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 535–13 544.

[6] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5084–5093.

[7] A. Frühstück, K. K. Singh, E. Shechtman, N. J. Mitra, P. Wonka, and J. Lu, "Insetgan for full-body image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7723–7732.

[8] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5933–5942.

[9] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu, "Avatarclip: Zero-shot text-driven generation and animation of 3d avatars," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–19, 2022.

[10] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, C. Qian, C. C. Loy, W. Wu, and Z. Liu, "Stylegan-human: A data-centric odyssey of human generation," in *Proceedings of European Conference on Computer Vision*. Springer, 2022, pp. 1–19.

[11] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3408–3416.

[12] J. Tang, Y. Yuan, T. Shao, Y. Liu, M. Wang, and K. Zhou, "Structure-aware person image generation with pose decomposition and semantic correlation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2656–2664.

[13] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing," *arXiv preprint arXiv:2304.02051*, 2023.

[14] Y. Jiang, S. Yang, T. L. Koh, W. Wu, C. C. Loy, and Z. Liu, "Text2performer: Text-driven human video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 747–22 757.

[15] A. Lin, N. Zhao, S. Ning, Y. Qiu, B. Wang, and X. Han, "Fashiontex: Controllable virtual try-on with text and texture," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–9.

[16] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7543–7552.

[17] S. Choi, S. Park, M. Lee, and J. Choo, "Viton-hd: High-resolution virtual try-on via misalignment-aware normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 131–14 140.

[18] S. He, Y.-Z. Song, and T. Xiang, "Style-based global appearance flow for virtual try-on," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3470–3479.

[19] K. Kim, S. Park, J. Lee, and J. Choo, "Reference-based image composition with sketch via structure-aware diffusion model," *arXiv preprint arXiv:2304.09748*, 2023.

[20] W. Chen and J. Hays, "Sketchygan: Towards diverse and realistic sketch to image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9416–9425.

[21] Y. Li, X. Chen, B. Yang, Z. Chen, Z. Cheng, and Z.-J. Zha, "Deepfacepencil: Creating face images from freehand sketches," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 991–999.

[22] S.-Y. Chen, F.-L. Liu, Y.-K. Lai, P. L. Rosin, C. Li, H. Fu, and L. Gao, "DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 90:1–90:15, 2021.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[25] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[26] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[27] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, W. Wu, and Z. Liu, "Unitedhuman: Harnessing multi-source data for high-resolution human generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7301–7311.

[28] K. Sarkar, L. Liu, V. Golyanik, and C. Theobalt, "Humangan: A generative model of human images," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 258–267.

[29] K. Sarkar, V. Golyanik, L. Liu, and C. Theobalt, "Style and pose control for image synthesis of humans from a single monocular view," *arXiv preprint arXiv:2102.11263*, 2021.

[30] Y. Jiang, S. Yang, H. Qju, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.

[31] T. Yoshikawa, Y. Endo, and Y. Kanamori, "Stylehumanclip: Text-guided garment manipulation for stylegan-human," *arXiv preprint arXiv:2305.16759*, 2023.

[32] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deepfacedrawing: Deep generation of face images from sketches," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 72–1, 2020.

[33] X. Wu, C. Wang, H. Fu, A. Shamir, S.-H. Zhang, and S.-M. Hu, "Deepportraitdrawing: Generating human body images from freehand sketches," *arXiv preprint arXiv:2205.02070*, 2022.

[34] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, and F. Wen, "Cocosnet v2: Full-resolution correspondence learning for image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 465–11 475.

[35] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, and C. Zhang, "Marginal contrastive correspondence for guided image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 663–10 672.

[36] S. Liu, J. Ye, S. Ren, and X. Wang, "Dynast: Dynamic sparse transformer for exemplar-guided image generation," in *Proceedings of European Conference on Computer Vision*. Springer, 2022, pp. 72–90.

[37] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2287–2296.

[38] W. Su, H. Ye, S.-Y. Chen, L. Gao, and H. Fu, "Drawinginstyles: Portrait image generation and editing with spatially conditioned stylegan," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[39] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

[40] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[41] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: fully convolutional networks for rough sketch cleanup," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[42] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8300–8311.

[43] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7450–7459.

[44] F.-L. Liu, S.-Y. Chen, Y. Lai, C. Li, Y.-R. Jiang, H. Fu, and L. Gao, "Deepfacevideoediting: Sketch-based deep editing of face videos," *ACM Transactions on Graphics(TOG)*, vol. 41, no. 4, p. 167, 2022.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[47] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[48] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, "Hyperstyle: Stylegan inversion with hypernetworks for real image editing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 511–18 521.

[49] S.-Y. Wang, D. Bau, and J.-Y. Zhu, "Rewriting geometric rules of a gan," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–16, 2022.

[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[51] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[52] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[53] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

Linzi Qu received her B.S. degree in Electronic Engineering from XiDian University in 2018. She is currently a PhD candidate in the School of Creative Media, City University of Hong Kong. Her research interests lie in computer graphics and computer vision.

Jiaxiang Shang received a Ph.D. degree from the Hong Kong University of Science and Technology. Before joining HKUST, he had his Bachelor degree in the Department of Computer Science and Engineering at Sun Yat-sen University. His primary research topic is about face reconstruction.

Hui Ye received a BA degree from the University of Science and Technology of China in 2016 and a Ph.D. degree from the City University of Hong Kong in 2022. She is now a RGC Postdoctoral Fellow at the School of Creative Media, City University of Hong Kong. Her research interests lie in the intersection between Human-Computer Interaction and Computer Graphics.

Xiaoguang Han is now an Assistant Professor and President Young Scholar of the Chinese University of Hong Kong (Shenzhen) and the Future Intelligence Network Research Institute. He received his PhD degree from the University of Hong Kong in 2017. His research interests include computer vision and computer graphics. He has published nearly 50 papers in well-known international journals and conferences, including top conferences and journals SIGGRAPH (Asia), IEEE TVCG, CVPR, ICCV, ECCV, NeurIPS, ACM TOG, etc. He is currently a guest editor of Frontiers of Virtual Reality and also an associate editor of the Journal of Computers & Graphics.

Hongbo Fu received a BS degree in information sciences from Peking University, China, in 2002 and a PhD degree in computer science from the Hong Kong University of Science and Technology in 2007. He is a Full Professor at the School of Creative Media, City University of Hong Kong. His primary research interests fall in computer graphics and human-computer interaction. He has served as an Associate Editor of The Visual Computer, Computers & Graphics, and Computer Graphics Forum.