Supplemental Materials
Sketch2Human: Deep Human Generation with Disentangled Geometry and Appearance Constraints

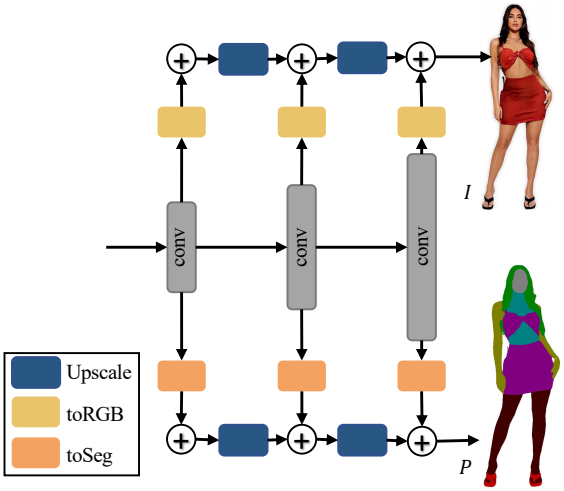

Fig. 1. The architecture of StyleGAN-Human with a newly added semantic branch. Here, we omit the mapping and affine transformation layers for brevity.

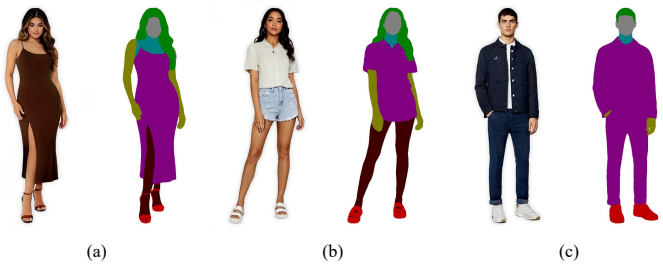

Fig. 2. Three representative examples of synthesized results and parsing maps. (a)-(c) are generated from StyleGAN-Human with the semantic branch. The synthesized result and the parsing map in each example are generated from the same latent codes.

I. Architecture of StyleGAN-Human with Semantic Branch

To enhance the geometry control via semantics, we modify the architecture of StyleGAN-Human to make it synthesize an image I and a parsing map P simultaneously. The architecture is similar to [1]. Besides the mapping and affine transformation layers, StyleGAN-Human consists of feature space convolutions and toRGB blocks. Feature space convolutions first learn layer-wise semantic information, and then toRGB blocks utilize convolutions to transfer this information to the RGB space. Considering that such semantic information is universal, we only add symmetric toSeg blocks, as illustrated in Figure 1. The output channel size of toSeg blocks is the same as the number of semantic labels. Different from [1], we find that freezing the original StyleGAN-Human's weights and only training toSeg blocks can achieve satisfactory results (Figure 2), without training all the weights together. We decided to use nine semantic labels since we can segment the human body with them while ensuring the accuracy of existing segmentation methods [2].

II. Implementation Details

Training. Our model is trained on one NVIDIA Tesla V100 GPU with the data randomly sampled from StyleGAN-Human and the Adam optimizer. For the Sketch Image Inversion module, we freeze the generator and train the sketch encoder for 80K iterations with a learning rate of 10^{-4}, and a batch size of 4. The resolution of the input sketch is 256×512. Then, for the Body Generator Tuning module, we update the generator for 1K iterations with a learning rate of 10^{-3} and a batch size of 1 for each sampled appearance code. For the losses $\mathcal{L}_{style}$ and $\mathcal{L}_{content_6}$, we choose the feature of $[relu1\_1, relu2\_1, relu3\_1, relu4\_1, relu5\_1]$. For the loss $\mathcal{L}_{content_{10}}$, we only use the feature of $[relu5\_1]$ to avoid the color features of I_{mix10} learned in the high-resolution layers to interfere with the final results. It takes 8 minutes to fine-tune the original generator at 512×1024 resolution on a single 3090Ti GPU.

Union Mask for Model Tuning. Model tuning mainly relies on appearance-transferred results. However, since the geometry of such results is not completely consistent with the input geometry and the content loss is calculated on the spatial features, directly training with them loses the expected geometry (see Figure 12 in the main paper). With a union mask M across different semantic labels, we only calculate the content loss at the position with similar semantics. For each semantic label i, a value of 1 in M_i signifies the semantic alignment between P_{mix6} and P_{syn} at that specific location, while a value of 0 indicates a difference in semantic representation between the two. The semantic branch of StyleGAN-Human is capable of producing the corresponding semantics (Figure 3 (b) and (d)). Since the parsing maps are represented by one-hot embedding, the union mask can be obtained by the pixel-wise and semantics-wise multiplication of P_{mix6} and P_{syn} (Figure 3 (e)).

Testing Data. Considering it is challenging for users to draw the entire textures, we expect sketch images to control only the global shape (e.g., clothing contours) and local details (e.g., wrinkles) and use reference images to provide such

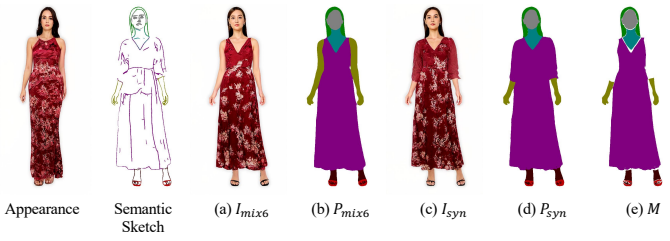

Fig. 3. An example of a union mask (e).

complex textures. Hence, we manually chose 2,000 sketches from DeepFashion [3] as our test set and excluded those with many lines within garments. We also collected freehand sketches based on the interface introduced in Section IV-A. We invited 5 users to draw full-body human sketches, and each of them drew 5 sketches. Through the self-assessment of their drawing skills, two of these users considered themselves professional, and the rest were novices. As for reference images, we randomly selected 100 styles from the latent space of StyleGAN-Human, including 50 pure color images and 50 texture images (e.g., floral, stripe, plaid). Pure color images denote garments containing one or more colors without fabric patterns (Figure 10 (a)-(d)), while texture images include both (Figure 10 (e)-(h)).

Metrics for Quantitative Comparison. We evaluate our method from two aspects. First, we expect the sketch encoder to embed all the geometry details from the sketch input successfully. Also, we hope the whole framework generates high-fidelity results. Similar to [4], we calculate the symmetric *Chamfer Distance* (CD) between the contour image of a synthesized image and the sketch input S_g to assess the accuracy of geometry control. We use Pix2pixHD to compute contour images. Note that we prepare a set of pure color appearance images to avoid the influence of garment textured lines on this CD metric for quantitative comparison. Additionally, we adopt *Mean Intersection over Union* (mIoU) to measure the semantic alignment between the generated images and the semantic input P_g. To measure the quality of synthesized image results, we incorporate $Fréchet\ Inception\ Distance$ (FID) [5] and Inception Score (IS) [6].

Metrics for User Study. Since there is no ground truth for combining sketch images with appearance images, it is difficult to judge the qualities using existing metrics fairly. We thus conducted a user study to evaluate the results from the perspective of human viewers. The evaluation criteria for the user study include four aspects. The first is geometry preservation (GP), which refers to the geometric alignment (pose, body shape, clothing shape) between a synthesized image and the input sketch. The second aspect is the accuracy of appearance transfer (AT). AT means whether the generated image and the input appearance image have the same or similar appearance. The third is visual quality (VQ). On the premise that GP and AT are relatively accurate, VQ measures which method shows the best visual quality. The last one is the user preference (UP). On the premise that GP and AT are relatively accurate, UP shows users' subjective preferences.

III. Evaluation Details

A. Details for Qualitative Comparison

For the compared methods, we retrain pSp [7] and CoCosNetV2 [8] with the SHHQ dataset [9] and DeepFashion dataset [3] respectively using their official implementations. CoCosNetV2 shows success for the edge-to-face and pose-to-body translation tasks. For a fair comparison, we repurpose the input of this method with a full-body sketch rather than a skeleton to achieve edge-to-body translation. As for e4e [10], we directly adopt the weights provided with StyleGAN-Human. We use the composable style adapter and the sketch adapter from T2I-Adapter [11] with pretrained models. For the sketches extracted from real images, the condition weights are set to 1 for the two adapters. For user sketches, the condition weights are set to 1 and 0.8 for the style adapter and sketch adapter. The input of pSp, CoCosNetV2, and T2I-Adapter is a sketch image similar to our sketch input but with no associated semantics, and e4e takes an RGB image as input. Additionally, we provide the same text prompt, "a photorealistic human", for T2I-Adapter.

B. Details for Human Image Generation Methods

Given a geometry RGB image and an appearance RGB image, we illustrate how to generate the specific inputs for the full-body image generation methods. For geometry input, we leverage DensePose [12], OpenPose [13], Graphonomy [2] to respectively extract Image-space UV coordinate maps, skeletons, parsing maps for PWS [14], NTED [15], and Text2Human [16]. ControlNet [17] can automatically produce sketches from the RGB input. The appearance input of PWS and NTED is the appearance image. Since Text2Human and ControlNet do not support appearance images, they need extra text input. According to the textures pre-defined by Text2Human (i.e., pure color, stripe/spline, plaid/lattice, floral, denim), we choose the one closest to the appearance image out of the five textures. We describe each appearance image with text for ControlNet. All the above methods utilize the officially published weights.

The text prompts corresponding to Figure 8 in the main paper are (a) A man wears a blue shirt and black pants. (b) A woman wears a pink dress. (c) A woman wears a gray shirt and red pants. (d) A woman wears a light blue dress with a red floral pattern. (e) A man wears a green floral shirt and army green shorts. (f) A woman wears a black dress with a white points pattern. The text prompts corresponding to Figure 12 are (a) A woman wears a purple dress. (b) A woman wears a dress with a leopard print texture. (c) A woman wears a white floral shirt and black pants. (d) A woman wears a blue floral dress. (e) A woman wears a black and white plaid coat and army green pants. (f) A woman wears a dress in salmon color.

IV. Applications

Due to separate control of geometry and appearance, our system can be applied to various applications. In this section, we present three typical applications.

A. Sketching Interface for Full-body Human Generation

To facilitate the human character design, we design a sketch-based full-body generation user interface (Figure 4). Considering that the overall human structure is challenging to draw, especially for novices, we provide geometry guidance in various poses (Figure 4 (a)). Meanwhile, a list of reference appearance images (pure color and texture) (Figure 4 (b)) is placed next to (a). The control panel (Figure 4 (c)) mainly consists of brush colors corresponding to nine semantic parts. When the "Generate" button is pressed, our system generates a synthesized result based on the current sketch and the selected reference appearance image. We also offer an eraser tool for editing the currently drawn sketch. To facilitate the drawing process, our interface operates on Microsoft Surface Pro 4 with a Surface pen to mimic a paper-and-pencil setting. Additionally, our system can produce realistic faces even with simple sketches, and thus it is friendly even for users with little drawing experience. Several results from users can be found in Figure 13.

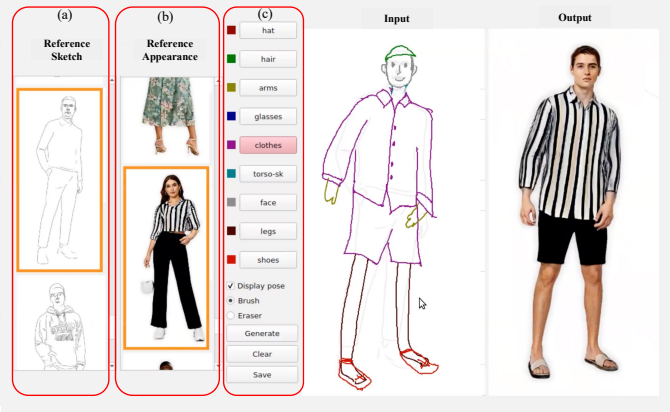

Fig. 4. A screenshot of our sketching interface for full-body human generation. The semantic sketch is displayed on the left canvas. The corresponding synthesized result given a selected reference appearance image (the orange box in (b)) is displayed on the right. Users might select one of the provided reference sketches (the orange box in (a)) as sketching guidance. The middle buttons (c) allow users to draw and edit according to different semantic components.

B. Full-body Geometry and Appearance Editing

The full-body geometry and appearance editing task aims to change the appearance or geometry (by modifying a corresponding sketch) while preserving its geometry or texture. Benefiting from the full disentanglement between the two aspects, our module enables users to edit from one aspect. For geometry editing, with the appearance code w_a unchanged and given a series of edited sketch images $\{S_{g,1}, S_{g,2}, \cdots\}$, our method can extract the corresponding geometry code $\{w_{g,1}, w_{g,2}, \cdots\}$ to synthesize edited results via our body generator. Figure 5 shows the progressive editing sequence by a user. Our sketch encoder can provide reasonable latent codes for crude or incomplete sketches. This is illustrated through an example involving a coarse-to-fine sketch editing process in Figure 6. Similarly, modifying the appearance code w_a can achieve appearance editing, as shown in Figure 13. The results

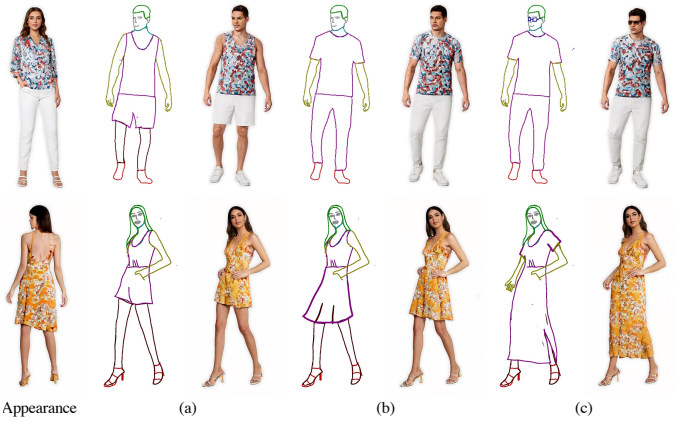

Fig. 5. Application of full-body sketch geometry editing. (a), (b) and (c) respectively consist of the edited semantic sketches and the corresponding results with the same appearance image.

demonstrate that our proposed method is controllable and can generate high-quality and diverse human images.

C. Full-body Geometry Interpolation

Benefiting from the smooth latent space of StyleGAN-Human, our method can produce continuous geometric transformation results from two predefined sketches. With the input of $I_{g,1}$ and $I_{g,2}$, our sketch encoder generates the corresponding geometry codes $w_{g,1}$ and $w_{g,2}$. After mixing each of these two codes with the texture code w_a, we can interpolate linearly between the mixed codes $w_{mix8,1}$ and $w_{mix8,2}$.

$$w_{inter} = \alpha \times w_{mix8,1} + (1 - \alpha) \times w_{mix8,2}. \quad (1)$$

When α varies between $(0, 1)$, our body generator produces the corresponding results. See Figure 7 for two interpolation examples.

V. Additional Qualitative Results

A. Robustness of Our System

Benefiting from the loss $\mathcal{L}_{adv_w}$, our sketch encoder maps the input sketch to a latent code within the actual distribution of W in StyleGAN-Human. Thus, with a poor sketch input (Figure 13) or even an empty image (with all pixels in white) (Figure 9 (a)), our method can still generate a realistic standing human image. However, it is important to note that for sketches depicting rare poses (Figure 8 (a) and (c)) and body proportions (Figure 8 (b)) in the training data for StyleGAN-Human, ensuring precise geometric alignment between the generated results and the input sketches becomes challenging. Due to the lack of sufficient training data, the generative power of StyleGAN-Human is limited. Thus, it is difficult to find efficient codes for those poses and proportions in the latent space.

Given the laborious nature of creating semantic sketches, our method generates reasonable and globally geometrically-aligned human images even when strokes are assigned incorrect semantics. However, the introduction of erroneous semantics disrupts the encoding procedure, resulting in subtle impacts on the pose (Figure 9 (c)), clothing texture (Figure 9 (d)) and gender (Figure 9 (d)) of the generated images.

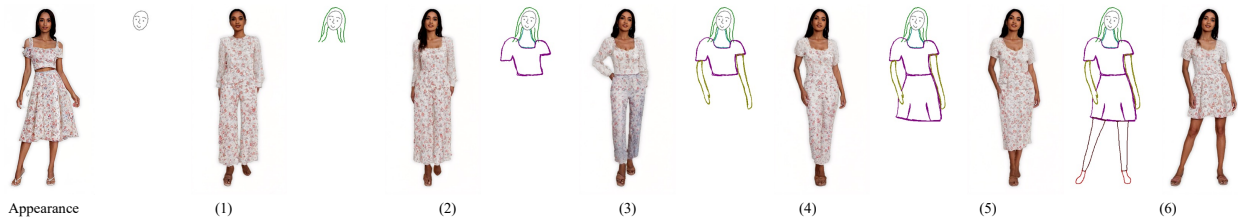

Fig. 6. The coarse-to-fine sketch editing process. For each step in this process, we show the freehand sketch (Left) and the corresponding result (Right).

Fig. 7. Two examples of full-body geometry interpolation.

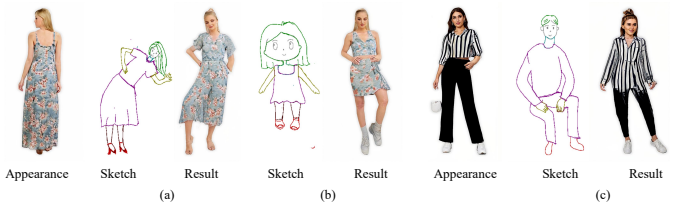

Fig. 8. Three less successful cases. The sketch inputs are not common in the training data in terms of body pose (a) and (c) and proportion (b).

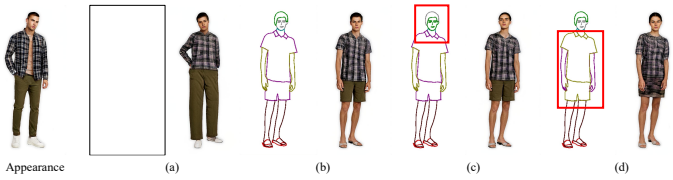

Fig. 9. Four results with a similar appearance input. Each example in (a)-(d) presents an input semantic sketch alongside its corresponding outcome. (a) illustrates the absence of a sketch input. (b) showcases the utilization of a correct semantic sketch input. (c) demonstrates the utilization of a wrong semantic sketch input, wherein the face and hair semantics have been exchanged. (d) displays the utilization of a wrong semantic sketch input, wherein the garment and hand semantics have been exchanged.

B. More Results for Baseline Comparison

Here we provide more comparison results with pSp [7], e4e [10], CoCosNetV2 [8], and T2I-Adapter [11], given the inputs of fine sketches (Figures 10) and coarse sketches (Figures 11). Additionally, Figure 12 presents the results of our approach compared with PWS [14], NTED [15], Text2Human [16], and ControlNet [17]. These results demonstrate that our method can flexibly and accurately control geometry and appearance.

C. More Results for Freehand Sketches

Figure 13 shows several examples of sketching from users. No matter if it is given a coarsely or finely drawn sketch, our method can provide a high-quality full-body human image, even with limited input information (e.g., for the face in Figure 13 (g)). We also provide multiple results for each sketch input with different appearance images. That proves that our method well disentangles geometry and appearance.

VI. Ethical Issues

Since we directly use synthetic images generated by StyleGAN-Human as training data, our method inherits biases from the original data distribution. Hence, it tends to generate lean and medium-height humans. Additionally, every male sketch with a hat results in a cap figure. Our method might be used to edit an actual portrait image. The appearance control via the reference image without geometry changes may disturb the gender and race, thus resulting in the original portrait's ethics. Using our work to spread false information or damage someone's reputation is strongly disapproved. Therefore, it is crucial to exercise caution when employing this technology.

Fig. 10. More visual comparison results.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[14] 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.

[15] 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.

[16] 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.

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

Fig. 11. More visual comparison results. The input sketches are collected from the users with our interface.

Fig. 12. More visual comparison results from the full-body image generation methods. Although these methods have different inputs, they are all produced from the same geometry and appearance images.

Fig. 13. More results from test users via our interface.