Supplementary Material for SemanticHuman-HD: High-Resolution Semantic Disentangled 3D Human Generation

In this **Supplementary Material**, we first introduce the implementation details of our proposed SemanticHuman-HD. Furthermore, we present additional results related to SemanticHuman-HD. For video demonstrations, including animations of generated humans and 3D-aware human image synthesis, please see the **Supplementary Video**.

1 Implementation Details

To assist others in reproducing our work or specific components, we provide implementation details of SemanticHuman-HD, including network architecture, dataset processing, training strategies and other relevant information.

Local Generator. For each local generator in SemanticHuman-HD, we utilize the same generator architecture as employed in AG3D [4]. However, to reduce the computational cost, we halve the number of channels in the network. Remarkably, this adjustment does not compromise the performance of model, as each local generator focuses solely on generating its corresponding semantic part. During the training, different local generators share the same semantic latent code as a condition, ensuring consistency across different semantic parts. In inference, we can feed different semantic latent codes to different local generators, enabling interesting applications such as semantic-aware interpolation and out-of-distribution image synthesis.

Discriminators. Throughout the training of stage 1 and stage 2, we employ multiple discriminators, including image discriminators, normal discriminators, semantic discriminators and face discriminators, to supervise the training. These discriminators share a common network architecture, only differing in channels and resolutions of input. Particularly, we opt not to use paired input of images and semantic masks. Our experiments reveal that employing two independent discriminators—one for images and another for semantic masks—enhances training stability. To facilitate conditional image synthesis, the discriminators also take human pose P and semantic label L_s as input. Additionally, we amplify the R1 penalty for the semantic discriminator by a factor of 10 compared to other discriminators, as we observed that introducing the semantic discriminator made the model more prone to collapse.

3D-Aware Super-Resolution Module. Our proposed super-resolution module leverages a feature super-resolution component to obtain high-resolution tri-plane representations. This module shares the same network design as the local generators and takes as input the low-resolution tri-plane representations generated in stage 1. Additionally, it conditions on the same semantic latent code used in stage 1. In the depth-guided sampling, we upsample the depth maps using a 2D FIR filter with coefficients [1, 3, 3, 1], which align with the super-resolution of tri-plane representations. Similarly, in the semantic-guided sampling, we first upsample the semantic masks and subsequently mask out those semantic parts whose values fall below the threshold δ. The chosen value for δ is 0.0005, ensuring that valid semantic parts are not inadvertently masked out. Notably, even with this small δ, the semantic-guided sampling can still exclude other semantic parts except one specific part for most pixels, providing evidence of the semantic disentanglement achieved by our method.

GAN Inversion. In GAN inversion, our goal is to derive a latent code that can be mapped to generated results similar to the given target results. This process involves optimizing the latent code by minimizing the differences between the target results and the generated results. As in SemanticStyleGAN [7], we employ LPIPS [9] and L1 loss between target and generated results to supervise the optimization of latent code. The target results encompass human images, semantic masks and face images. Notably, the face images are cropped from human images based on their SMPL [2] parameters.

Semantic Label. The semantic label L_s serves as an indicator for gender and whether the image contains specific types of garments. It is represented as a 21-bit vector, where each value can be either 0 or 1. Apart from gender, the remaining 20 bits correspond to the following garment types: top, outer, skirt, dress, pants, leggings, headwear, eyeglass, neckwear, belt, footwear, bag, ring, wrist wearing, socks, gloves, necklace, rompers, earrings and tie.

Dataset Processing. The DeepFashion dataset provides images and semantic masks. The original masks contain 24 categories, and we simplified them into 6 broader categories: "Body" covers face, hair, and skin; "Tops" covers tops, dresses, and rompers; "Outer" covers outer; "Bottoms" covers skirts, pants, and leggings; "Shoes" covers socks and shoes; "Accessories" covers everything else.

2 Additional Results

In this section, we present additional results. Notably, some of these results are not attainable by existing methods [1, 3–5, 8], such as semantic-aware interpolation and 3D garment interpolation. While view control and pose control are common applications, we specially apply them to the 3D garments disentangled from 3D humans. These novel capabilities demonstrate the versatility and effectiveness of our proposed SemanticHuman-HD.

3D Human Interpolation. Given that each semantic latent code corresponds to a generated 3D human, interpolating between two latent codes allows for smooth transitions between two distinct 3D humans. Fig. 1 illustrates the results of such interpolations, accompanied by corresponding semantic masks and normal maps.

Semantic-Aware Interpolation. Our method enables semantic-aware interpolation of 3D humans. This means we can interpolate between specific semantic parts of two generated 3D humans. For instance, it allows us to smoothly alter a specific semantic part while keeping other parts unchanged. Please refer to Fig. 2 for visual examples. The semantic-aware interpolation is achieved by interpolating between two semantic latent codes, with the interpolation confined to the codes corresponding to specific semantic part.

3D Garment Interpolation. The independence of our generation process extends to 3D garment generation and interpolation. By setting the densities of other semantic parts (excluding the specific garment) to zero, we achieve 3D garment generation. Fig. 3 showcases the results of interpolating between two 3D garments. Additionally, we provide normal maps during the interpolation process to visualize the geometric transitions.

View Control. To demonstrate the 3D-consistent generation capability of our method, we render 3D humans and 3D garments from different viewpoints. Specifically, we disentangle the 3D garments from the generated 3D humans by setting the densities of the “body” to 0. The resulting renderings are shown in Fig. 4. Notably, our proposed 3D-aware super-resolution module enables high-resolution image synthesis without compromising 3D consistency. In other words, the images rendered from different viewpoints maintain coherence. In contrast, works [4,8] using a 2D super-resolution module do not achieve full 3D-consistency. For a clear demonstration of this point, please refer to the videos provided by AG3D [4].

Pose Control. Our method leverages the deformer to enable pose control for both generated 3D humans and 3D garments. In practical terms, this means that given a sequence of human poses, our method can animate the generated results—making them run, walk, and perform other actions. This flexibility makes our approach suitable for various downstream applications, such as virtual reality and video games. Similar to the view control, the images synthesized in different poses remain consistent, thanks to our proposed 3D-aware super-resolution module. The results are shown in Fig. 5. For a comprehensive view of the animations, please refer to the **Supplementary Video**. The pose sequences used in the animation are sourced from [6].

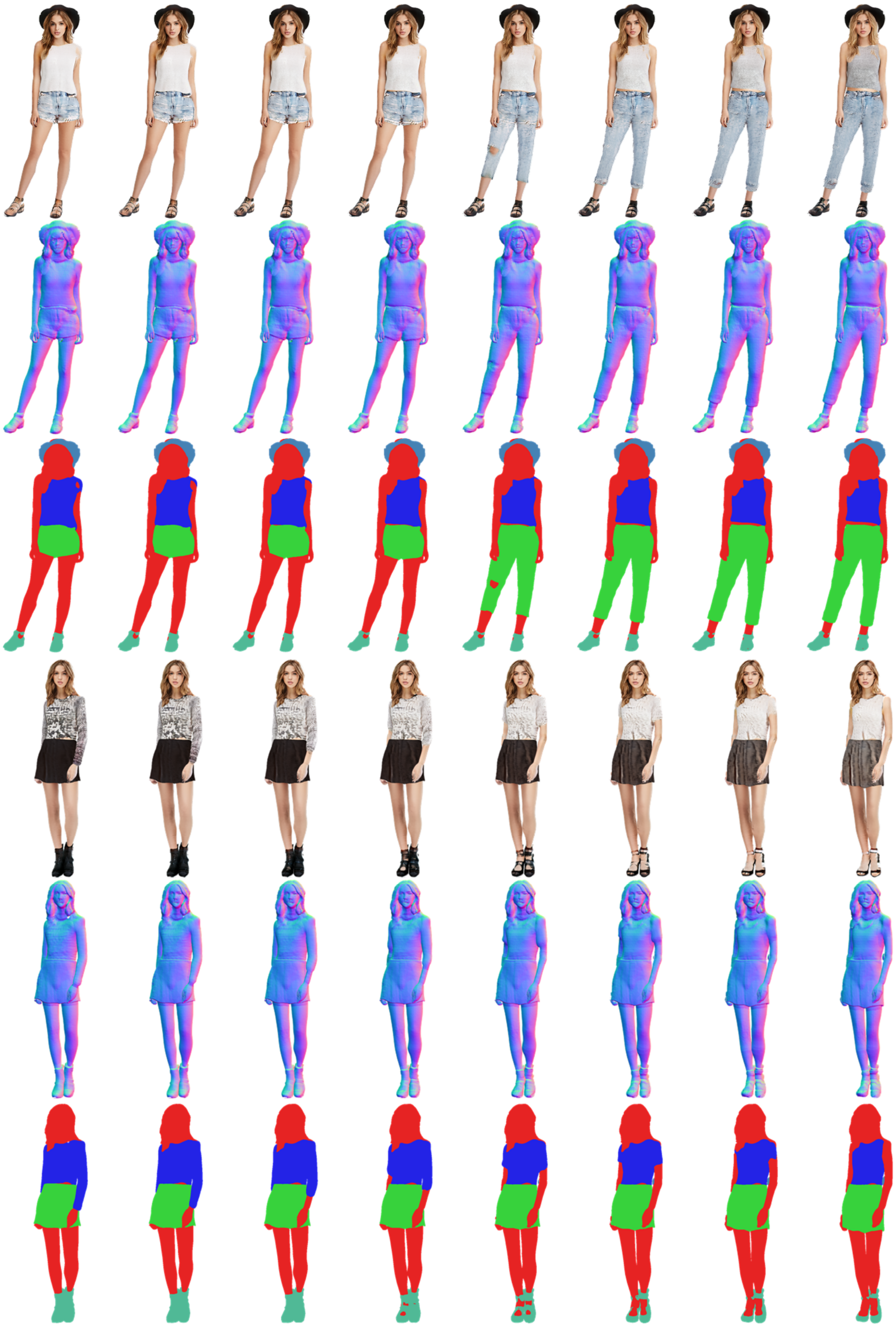

Fig. 1: 3D human interpolation. Each interpolation result includes an image, a normal map and a semantic mask.

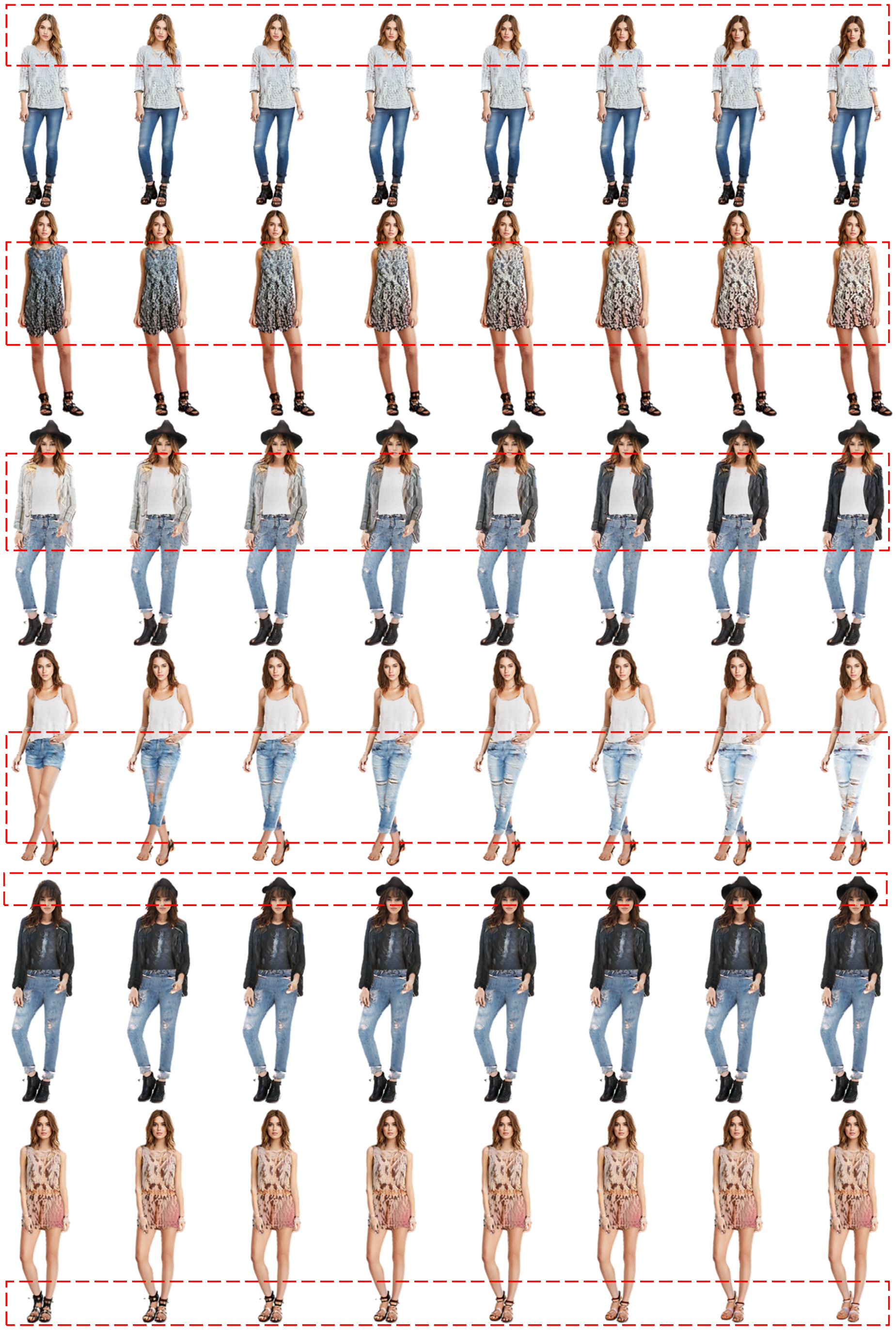

Fig. 2: Semantic-aware interpolation. Red dashed rectangles on the images indicate chosen semantic parts during the interpolation.

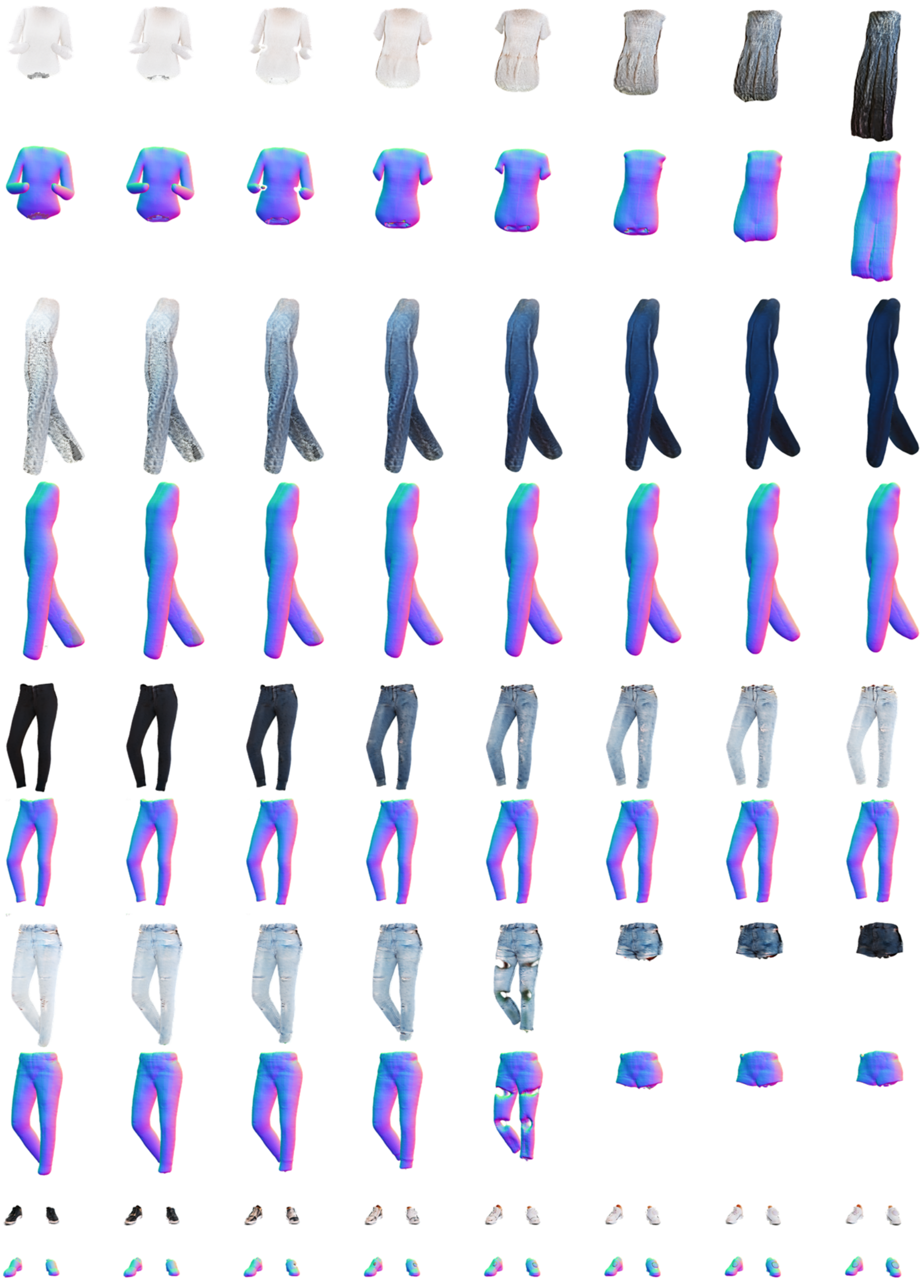

Fig. 3: 3D garment interpolation, including images and normal maps. For a closer view, please zoom in to see the details.

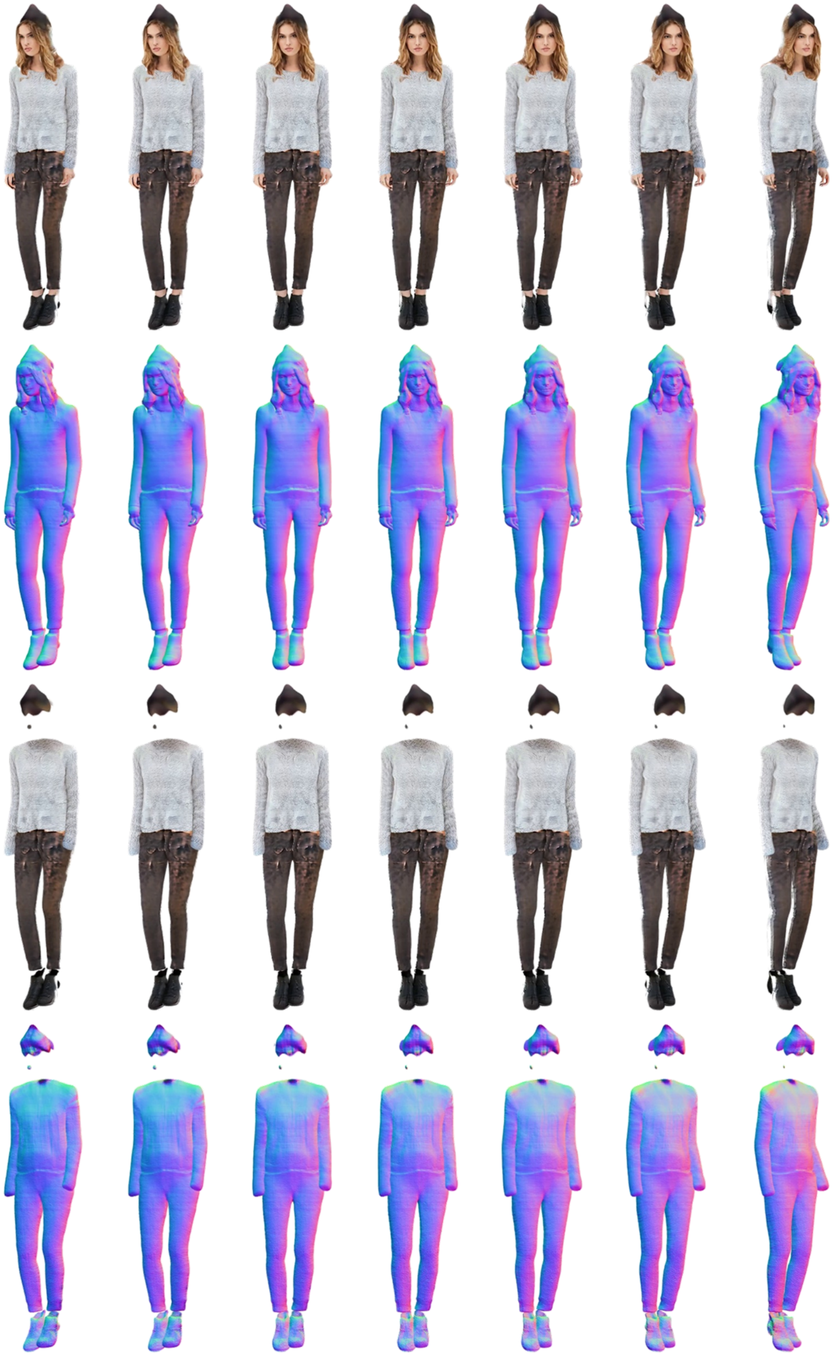

Fig. 4: View control. The images in the last two rows depict 3D garments disentangled from 3D humans shown in the first two rows.

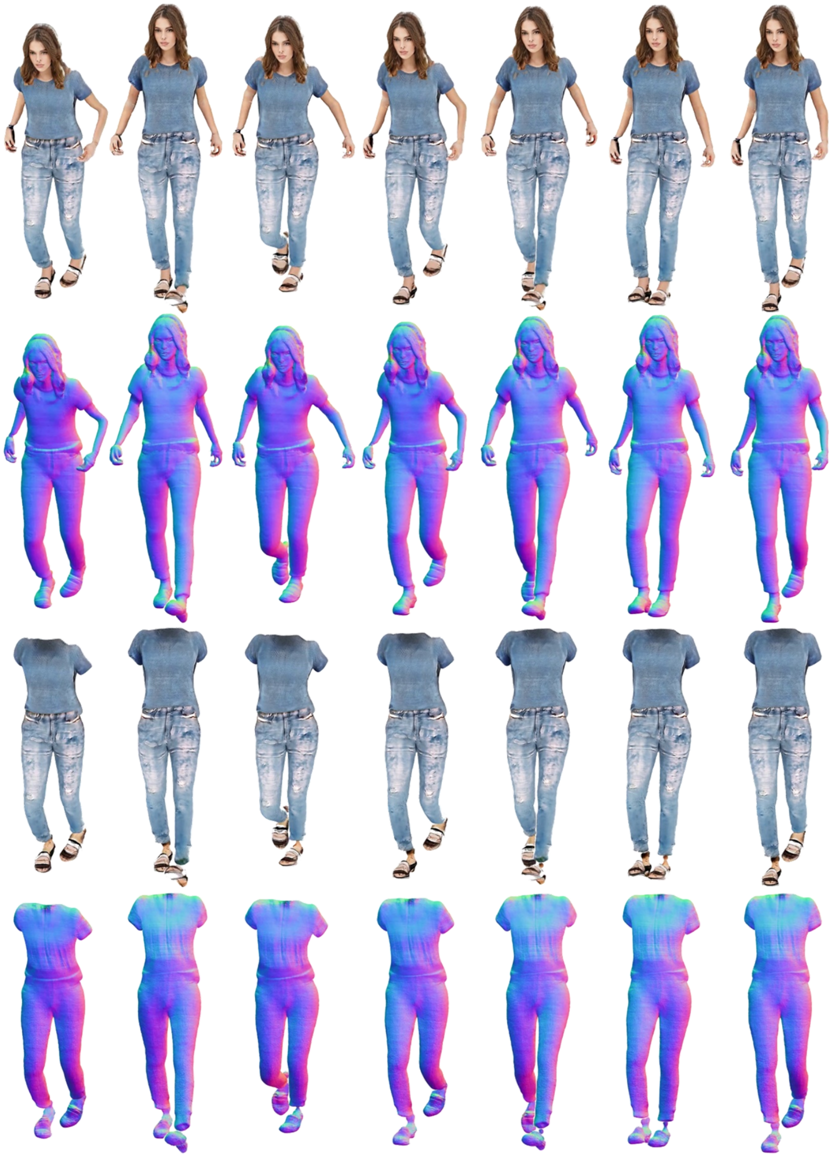

Fig. 5: Pose control. The images in the last two rows depict 3D garments disentangled from 3D humans shown in the first two rows.

References

1. Abdal, R., Yifan, W., Shi, Z., Xu, Y., Po, R., Kuang, Z., Chen, Q., Yeung, D.Y., Wetzstein, G.: Gaussian shell maps for efficient 3d human generation. arXiv preprint arXiv:2311.17857 (2023) 2
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016) 2
3. Chen, X., Huang, J., Bin, Y., Yu, L., Liao, Y.: Veri3d: Generative vertex-based radiance fields for 3d controllable human image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8986–8997 (2023) 2
4. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. arXiv preprint arXiv:2305.02312 (2023) 1, 2, 3
5. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: Eva3d: Compositional 3d human generation from 2d image collections. arXiv preprint arXiv:2210.04888 (2022) 2
6. Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2019), https://amass.is.tue.mpg.de 3
7. Shi, Y., Yang, X., Wan, Y., Shen, X.: Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11254–11264 (2022) 2
8. Yang, F., Chen, T., He, X., Cai, Z., Yang, L., Wu, S., Lin, G.: Attrihuman-3d: Editable 3d human avatar generation with attribute decomposition and indexing. arXiv preprint arXiv:2312.02209 (2023) 2, 3
9. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 2