Supplementary materials for AttriHuman-3D: Editable 3D Human Avatar Generation with Attribute Decomposition and Indexing

Fan Yang[1] Tianyi Chen[2] Xiaosheng He[1] Zhongang Cai[1,3]
Lei Yang[3] Si Wu[2] Guosheng Lin[1]
[1]S-Lab, Nanyang Technological University
[2]School of Computer Science and Engineering, South China University of Technology
[3]SenseTime Research
fan007@e.ntu.edu.sg, csttychen@mail.scut.edu.cn, gslin@ntu.edu.sg

1. More Implementation Details

Implementation Details. Our generator is built on the top of StyleGAN2 [1]. To achieve better generation quality and efficiency, we adopt an improved version of generator structure following IDE3d [4]. Instead of using different latent codes for different components in the generator, we adopt the same latent code for all the components and keep the other structure the same as IDE3d. The learning rates of generator and discriminator are 0.0025 and 0.002, respectively. The gamma value of RGB image and semantic masks are set to 10 and 100 to prevent the semantic masks from dominating the backward gradient thus degrading the quality of generated images. Our model is trained in two stages by gradually increasing the neural rendering resolution from 64 to 128.

Volumen Rendering. Similar as StyleSDF [3], we transform the signed distance value d into the density value σ as $\sigma(x) = \frac{1}{\sigma} \cdot Sigmoid(\frac{-d(x)}{\alpha})$. For each ray $\mathrm{r}(t) = \mathrm{o} + t\mathrm{v}$, we query the sampled points and integrate all the values with volume rendering equations:

$$I(R) = \sum_{i=1}^{N} \left(\prod_{j=1}^{i-1} e^{-\sigma_j \cdot \delta_j} \right) \cdot (1 - e^{-\sigma_i \cdot \delta_i}) \cdot f_i, \qquad (1)$$

where $\delta = \|x_i - x_{i-1}\|$ and N is the number of sampled points along each ray.

Discriminator. To model the joint distribution of real RGB images and semantic masks, We adopt a dual-branch discriminator D_{dual}, which takes both RGB images and segmentation masks as input. We feed the SMPL parameters [2] and gender label into the discriminators as conditions to further stabilize the training process. Moreover, to enhance the generation quality of the human face, we further involve a face discriminator D_{face}. In detail, we crop the head regions from the generated high-resolution fake images and feed it into another shallow face discriminator. We show the detail structure of our discriminator in Figure 1.

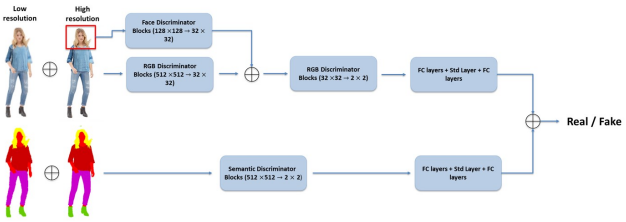

Figure 1. Structure of our discriminator.

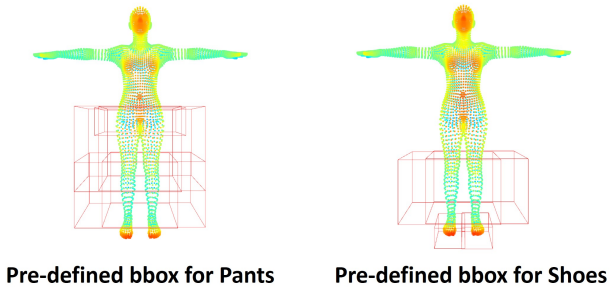

Figure 2. Examples of pre-defined attribute bounding boxes.

Attribute-specific Sampling Strategy. We show examples of our pre-defined attribute-specific bounding boxes in Figure 2. We define attribute-specific sampling bounding box for different attribute based on the pre-defined body joints on the SMPL templates [2]. Our attribute-specific sampling strategy limits the influence of each attribute explicitly inside the bounding box, which avoids the entanglement between attributes with little over-lap areas and further improves the computational efficiency by sampling fewer points for each attribute only inside the bounding boxes.

Edit Pipeline. To edit certain attribute, we could change

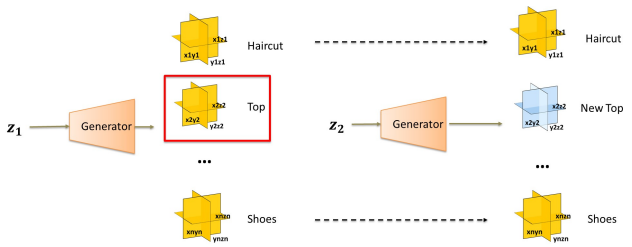

Figure 3. Pipeline in the editing stage (Edit Top clothes).

the generated feature planes of the selected attribute with another ones generated from new sampled latent codes. The pipeline of editing is shown in Figure 3. Benefit from our 4D space-attribute representation of the generated human avatars, we are capable to modify the generated features of certain attribute while keeping others fixed. Moreover, we also support deleting or adding more attributes into the generation process, allowing us to achieve the Try-on effect.

2. Limitation and Future work.

Our model allows fine-grained image editing and generates high-quality view- consistent 3D human avatars. However, there are still some limitations for further improvement. 1) Although our model supports the precise editing on certain semantic region, we could not control the specific style of the edited results. Involving stronger control signals such as text could be a promising future direction. 2) The estimation of SMPL parameters could be inaccurate, leading to noisy label for the discriminator to learn in the training process, which may degrade the generation quality. Improve the accuracy of SMPL estimation may lead to better results.

3. Additional Experimental Results

More Qualitative Results. We show more qualitative results in this section. In detail, we show the 3D-consistent image, semantic mask and geometry rendering results of our method in Figure 4, 5. We show the 3D-consistent editing results of our method, including editing certain attributes and try on new attributes in Figure 6, 7. In Figure 8, we show more latent space interpolation results for Pants and Dress. In Figure 9, we show more results for animation.

References

[1] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1

[3] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 1

[4] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 1

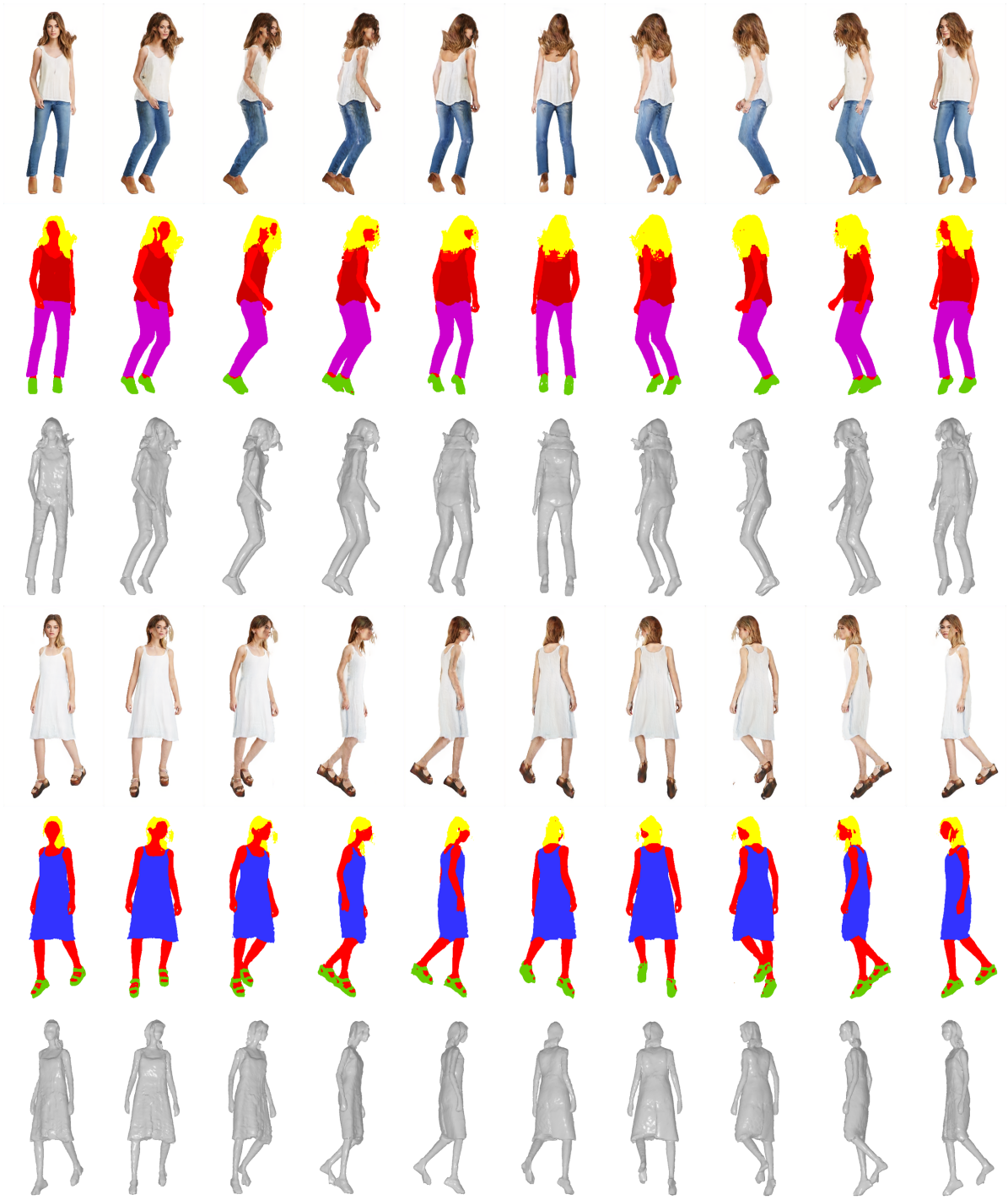

Figure 4. The 3D-consistent image, semantic mask and geometry rendering results of our method.

Figure 5. The 3D-consistent image, semantic mask and geometry rendering results of our method.

Figure 6. The 3D-consistent editing results of our method.

Figure 7. The 3D-consistent editing results of our method.

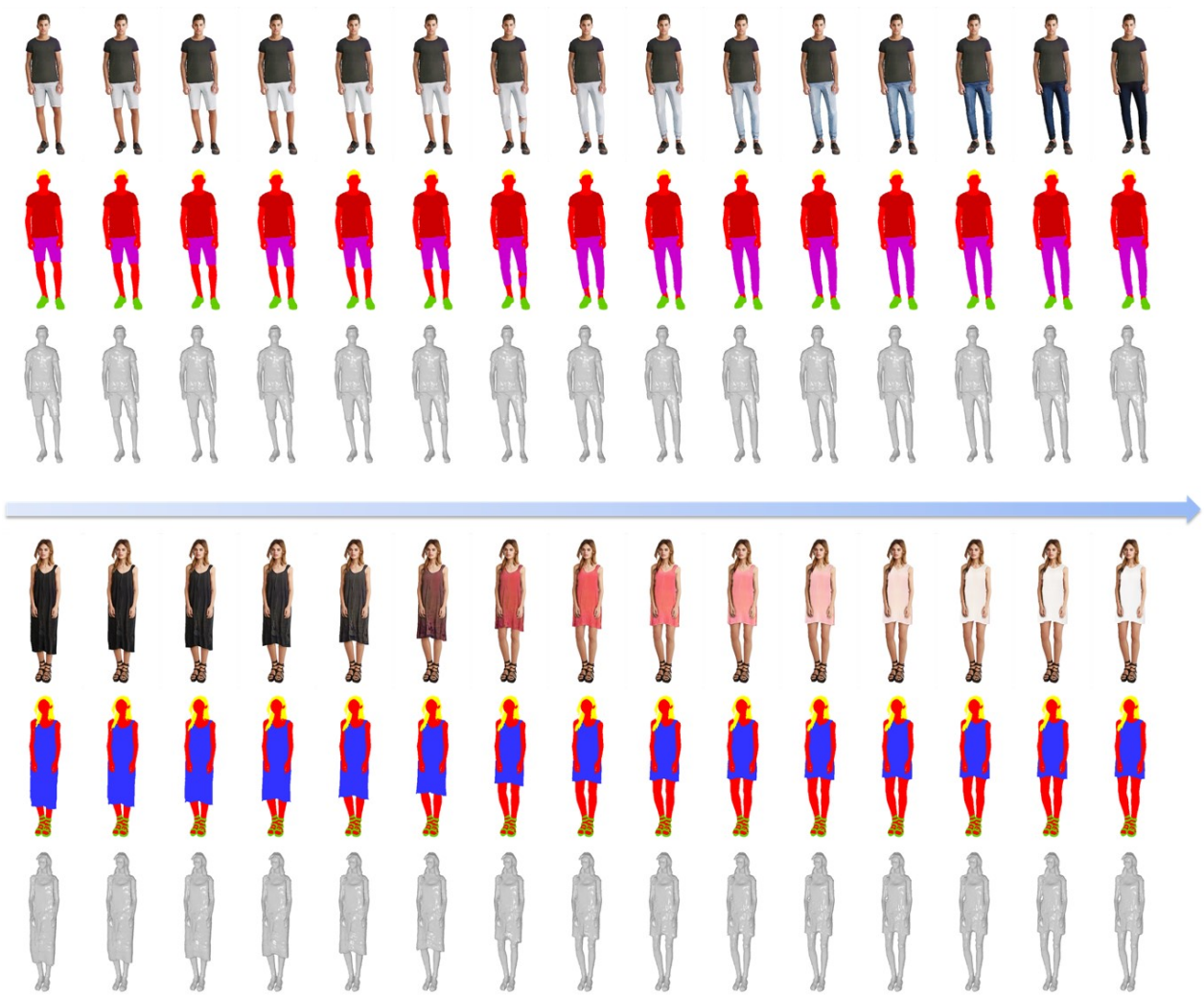

Figure 8. More latent space interpolation results. (Pants and Dress)

Figure 9. More animation results.