7. Supplemental Material

7.1. Implementation Details

We train our HumanGen using 4 NVIDIA A40 GPUs. We first train the geometry branch, which takes about 12 hours to converge. The trained implicit function $\mathbb{F}_{sdf}$ has hidden neurons of (273, 128, 32, 1). The loss weights are $\lambda_{mask} = 1$, $\lambda_{3D\_SDF} = 1$ and $\lambda_{eik} = 0.1$. For the texture branch, we first train the base model for around 24 hours and then continue to train each model for another roughly 18 hours. For generating the tri-plane feature, we generate planes of shape 256×256 with channel size of 32. We further add another two stylegan [31] synthesis blocks with upscale equals to 1 so as to extend the layer number of latent w^+ to 18, which is compatible with the layer number of w^+ from Stylegan-human [13] generator. The loss weights are $\lambda_{2D\_front} = 1$, $\lambda_{3D\_RGB} = 8$, $\lambda_b = 1\text{e}{-2}$, $\lambda_{2D\_back} = 8$ and the r1 regularization term of adversarial training (Eqn. 5) has weight $\lambda = 10$.

7.2. Additional Evaluation

We further conduct a texture fitting evaluation. Given the already-trained geometry branch, we only use anchor image and photometric loss to fit the frontal texture of given geometry. Let M denotes the mapping network from Stylegan-human. As shown in Fig.13, without the mapping network from Stylegan-human (**w/o** M), we train a newly-initialized mapping network to map the same noise z to some latent and synthesize the tri-plane feature. However, it generates results that are all blurred and have closing color with each other because the newly-initialized mapping network fails to recover the original latent space of Stylegan-human. Without the geometry embedding feature given to the decoder (**w/o geometry embedding**), though the fitting texture can follow the color consistency with anchor image, it tends to have no details. With both M and geometry embedding (**full**), the fitting results can maintain better consistency with anchor image while recovering some geometry details as well. The corresponding quantitative results are provided in Tab.4. However, the texture fitting results are still less-detailed. Therefore, we further add the blending scheme and GAN training to improve the effects. We provide interpolation results in Fig.15 and more generation results of our complete model in Fig.14. Note that our approach can generate more photo-realistic results than previous methods with detailed geometry and free-viewing ability.

Table 4. Quantitative evaluation of texture fitting.

Method	PSNR↑	SSIM↑	LPIPS ↓
w/o M	15.86	0.7835	0.2596
w/o geometry embedding	19.89	0.8236	0.2075
full	**21.13**	**0.8434**	**0.1803**

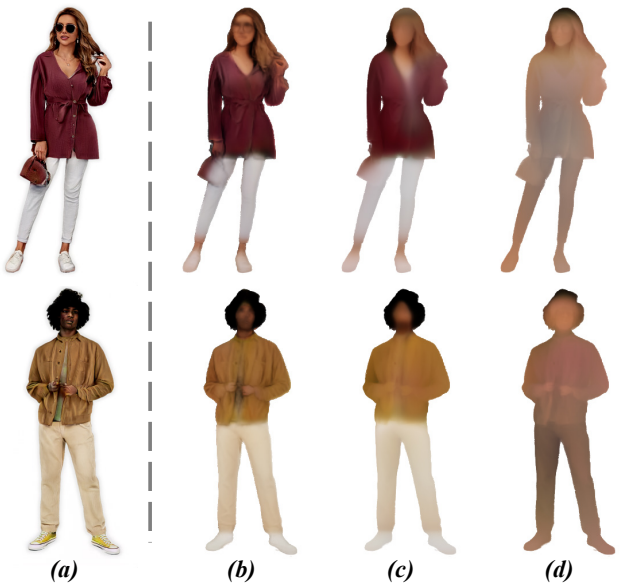

Figure 13. Qualitative evaluation of texture fitting. (a) anchor image; (b) full; (c) w/o geometry embedding; (d) w/o M.

Figure 14. More generation results using our HumanGen. Note that our approach enables photo-realistic human generation with detailed geometry and free-view rendering.

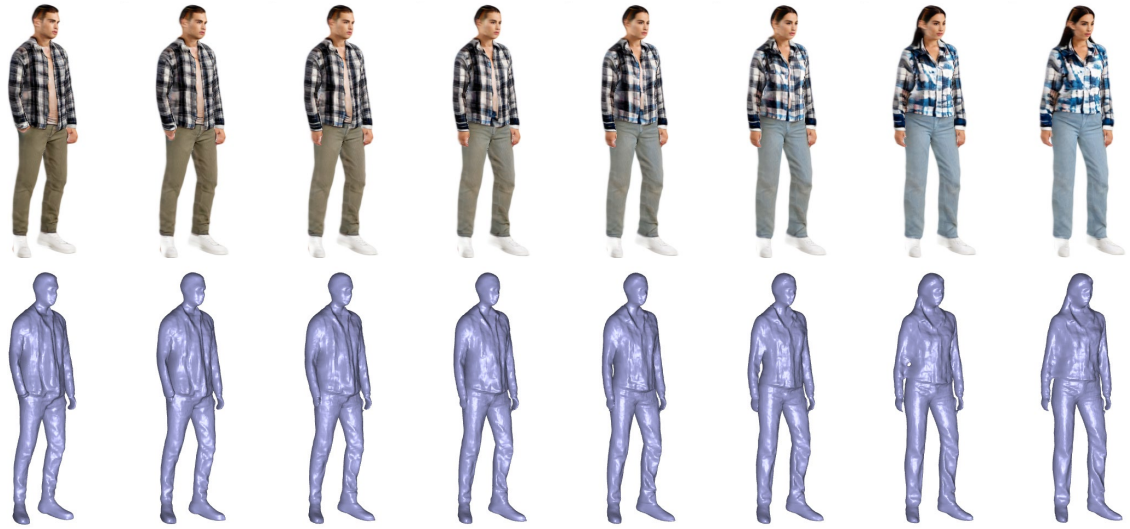

Figure 15. Interpolation results of our HumanGen.