Supplementary Material for VeRi3D: Generative Vertex-based Radiance Fields for 3D Controllable Human Image Synthesis

Xinya Chen[1], Jiaxin Huang[1], Yanrui Bin[2], Lu Yu[1], Yiyi Liao[1*]

[1]Zhejiang University [2]Huazhong University of Science and Technology

{hust.xinyachen, jaceyh919, binyanrui}@gmail.com, {yul, yiyi.liao}@zju.edu.cn

Abstract

*In this **supplementary document**, we first clarify implementation details in Section 1. Next, we provide more results for part control and additional experimental results in Section 2. Finally, we discuss our limitations in Section 3. Our **supplementary video** further demonstrates the controllability of our method over camera pose, human pose, shape, appearance, and parts.*

1. Implementation Details

In this section, we describe the implementation details of our proposed VeRi3D. We also provide more details of the datasets.

1.1. VeRi3D

Network Architecture: The generator composes of three parts: a mapping network, a convolutional backbone, and an MLP decoder. The mapping network and the convolutional backbone follow the official implementation[1] of StyleGAN2 [3]. The convolutional backbone consists of 7 style blocks, conditioning on a 256-dimensional Gaussian noise input and producing a 64-channel 256 $\times$ 256 feature image. Each block consists of modulation, convolution, and normalization. The MLP decoder consists of 2 hidden layers of 256 units. The input to the MLP includes the 64-channel averaged feature vector $\bar{\mathbf{f}}$ and the positional encoded local coordinate information $\gamma(\mathbf{x}^l)$ with $L = 10$.

1.2. Datasets

Surreal: Following ENARF [4], we crop the first frame of all videos to 180 $\times$ 180 with the center at the pelvis joint and then resize it to 128 $\times$ 128. 68033 images are obtained in total. The background mask provided by the dataset is used to replace the background with black color.

AIST++: Following ENARF [4], we crop the images to 600 $\times$ 600 with the center at the pelvis joint, and then resize it to 256 $\times$ 256. 3000 frames are sampled for each subject, resulting in 90K frames in total. We use an off-the-shelf segmentation model [1] to paint the background black.

DeepFashion: We use the data provided by [2]. The dataset filters out images with partial observations and inaccurate SMPL estimations, resulting in 8K images for training. We render the images at the resolution of 512 $\times$ 256 pixels.

*Corresponding author.

[1]https://github.com/NVlabs/stylegan3

2. Additional Experimental Results

2.1. Part Control

Part Control using Different PCA Components: We further show the results of different PCA components in Fig. 1. Here, we change one component coefficient for each part respectively and keep the remaining component coefficients fixed. We find that different components capture different semantic meanings. For example, the components of the head capture hair length and color, whereas the components of the upper/lower body control cloth length, tightness, and color. Interestingly, we observe that one component of the head controls its orientation. We note that this is due to the distribution mismatch between the GT poses and the noisy SMPL poses, where the generator needs to model the orientation to compensate for the inaccurate pose distribution.

Figure 1: **Additional Qualitative Results** for different PCA Components. We show three components for each part and three samples for each component.

2.2. Additional Qualitative Results

DeepFashion: Fig. 2 shows a qualitative comparison on the DeepFashion dataset. We provide the corresponding real image as a reference for each input pose. Our method has more fidelity and higher pose accuracy.

ZJU-MoCap: Fig. 3 shows a qualitative comparison with HumanNerf on ZJU-MoCap. Our method achieves competitive performance to the state-of-the-art human reconstruction method HumanNeRF.

Figure 2: **Qualitative Comparison on DeepFashion.** We provide the corresponding real image as a reference for each input pose.

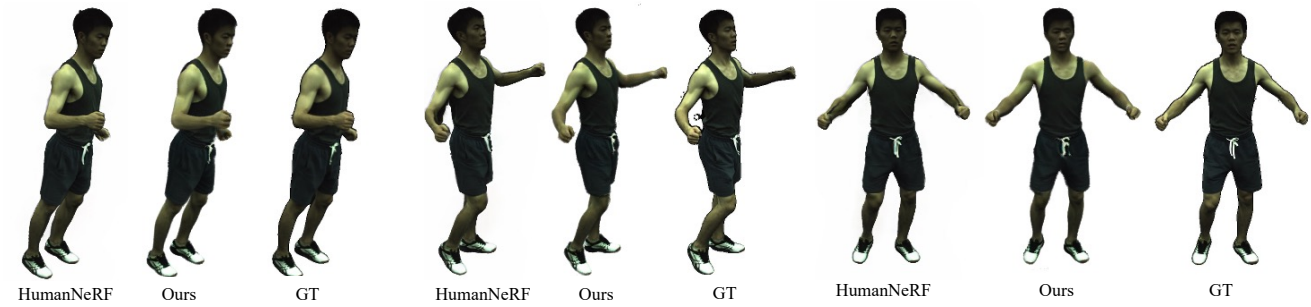

Figure 3: **Qualitative Comparison with HumanNerf on ZJU-MoCap.**

3. Limitations

Our method may be negatively affected by inaccurate data. In the DeepFashion dataset, the distribution of the estimated SMPL poses mismatches the image poses. Many bent legs are estimated as straight ones. To match the image pose distribution during training, the model needs to generate the bent leg on top of the straight SMPL leg as a sort of "clothes". Therefore changing the appearance may alter the pose. This could be addressed by jointly refining the pose distribution in future work. In AIST++ dataset, the background is not removed accurately, especially in the hand regions. Therefore our model may also generate stuff around the hand to match the image distribution.

References

[1] PaddlePaddle Contributors. Paddleseg, end-to-end image segmentation kit based on paddlepaddle. https://github.com/PaddlePaddle/PaddleSeg, 2019. 1

[2] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *International Conference on Learning Representations*, 2023. 1

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1

[4] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 1