# Supplementary Materials for

# "En3D: An Enhanced Generative Model for Sculpting 3D Humans from 2D Synthetic Data"

[1]Yifang Men, [1]Biwen Lei, [1]Yuan Yao, [1]Miaomiao Cui, [2]Zhouhui Lian, [1]Xuansong Xie

[1]Institute for Intelligent Computing, Alibaba Group

[2]Peking University, China

In this document we provide the following supplementary contents:

- Video demonstration of generative 3D human model.
- Results of 3D human generation.
- Comparison with state-of-the-art methods.
- Details of network architectures.
- The semantical UV partitioning.
- Exemplar of synthetic 2D images.
- Limitations and future work.

1. Video demonstration of generative 3D human model

Without any 3D or 2D datasets, we constructed a generative model capable of producing visually realistic, geometrically accurate and content-wise diverse 3D humans. The generated avatars can be seamlessly animated and easily edited. We also demonstrated the scalability of our approach for content-style free adaptation, i.e., portrait or Disney cartoon character synthesis. The results are provided in the supplemental video (file '3929 video.mp4').

2. Results of 3D human generation

By combining the enhanced 3D generative model with two optimization modules, our method achieves the synthesis of visually realistic and geometrically accurate high-fidelity 3D human avatars. It enables the production of diverse 3D humans, covering a wide range of age groups, genders, races, appearances, and clothing styles. Our results of synthesized 3D humans rendered in various viewpoints are shown in Figure 1, 2, 3.

Figure 1: Results of synthesized 3D human in various viewpoints.

Figure 2: Results of synthesized 3D human in various viewpoints.

Figure 3: Results of synthesized 3D human in various viewpoints

3. Comparison with state-of-the-art methods

We provide more comparison results with state-of-the-art methods in Figure 4.

(a) EVA3D [3]

Figure 4: Qualitative comparison with state-of-the-art methods.

(b) AG3D [2]

Figure 4: Qualitative comparison with state-of-the-art methods.

(c) EG3D [1]

Figure 4: Qualitative comparison with state-of-the-art methods.

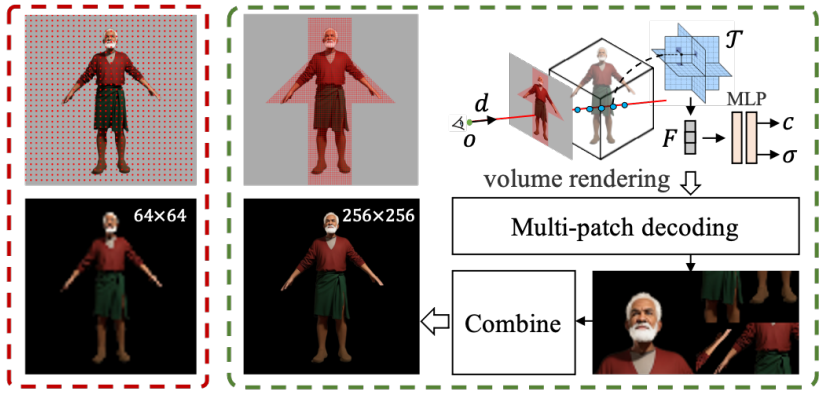
(d) Ours

Figure 4: Qualitative comparison with state-of-the-art methods.

4. Details of network architectures

For the 3D generative module, the architectures of our triplane generator, neural renderer and discriminators follow the implementations in EG3D [1] and the super-resolution network is also used with dual-discrimination. We adopt patch-composed rendering by adaptive ROI sampling and full-image decoding from rendered multiple patches, as illustrated in Figure 5. This strategy enables feature images with a resolution of $256^2$ rendered with the same computational efficiency as a resolution of $64^2$.

For the geometric sculpting module, we parametrize DMTET as a MLP to predict the SDF value and the position offset for each vertex, following by a tetrahedra layer [5] to extract the triangular mesh. Details of the network structure is shown in Table 1. For the explicit texturing module, a differentiable rasterizer from [4] is employed to optimize the UV texture map and the computation of UV coordinates are described in Section 5.




(a) Original solution (b) ROI patch-based sampling and composition

Figure 5. Visualization of patch-composed neural rendering.

Table 1: Details of geometric sculptor.

| Operation | | Output Size |
|---|---|---|
| 3D points | | $N_p \times 3$ |
| Encoder | HashGrid | $N_p \times 32$ |
| MLP | Linear + ReLU | $N_p \times 32$ |
| | Linear + ReLU | $N_p \times 32$ |
| | Linear | $(s, \delta v),\ N_p \times 4$ |
| Tetrahedra layer | | $\mathcal{M}_{tri}, N_v \times 3$ |

5. The semantical UV partitioning

Based on the canonical properties of synthesized bodies, we semantically split the vertices of the triangular mesh into 5 body components and assign the position of each component in UV map with start and end points ($start_w$, $end_w$, $start_h$, $end_h$ ). We rotate each component to be vertical to the horizontal plane and compute the corresponding UV coordinates via cylinder unwarping. The split of the UV map is presented in Table 2 and the visualization of the final UV texture is shown in Figure 6.

Table 2: Details of UV splitting.

| Component | $start_w$ | $start_h$ | $end_w$ | $end_h$ |
|---|---|---|---|---|
| Trunk | 1/6 | 0 | 5/6 | 1 |
| Left arm | 0 | 0 | 1/6 | 1/2 |
| Right arm | 5/6 | 0 | 1 | 1/2 |
| Left leg | 0 | 1/2 | 1/6 | 1 |
| Right leg | 5/6 | 1/2 | 1 | 1 |

Figure 6: Visualization of explicit UV texture

6. Exemplar of synthetic 2D images

The synthetic 2D images initially consists of 7 views covering from horizontal 0° to 180°, which are flipped for full 360° images. The exemplar of our synthetic 2D images in 7 views are shown in Figure 7.

(a) 0° view images

(b) 30° view images

(c) 60° view images

Figure 7: Exemplar of our synthetic images covering 7 viewpoints.

(d) 90° view images

(e) 120° view images

(f) 150° view images

(g) 180° view images

Figure 7: Exemplar of our synthetic images covering 7 viewpoints.

## 7. Limitations and future work

Due to training on synthetic human images without accessory, our model cannot produce exaggerated accessories (e.g., hats) with realistic geometry and appearance. The failure case is shown in Figure 8.

Our model learns a shared 3D human representation with canonical pose utilizing synthetic structured human images. Without pose deformation modeling, our method only supports inputting standard human images in A pose for image guided synthesis. This constraint could be potentially released by introducing a human re-pose function in the preprocess step. Alternatively, exploring the modeling of articulated humans using 2D synthetic data could also address this issue and is encouraged for further work.

The proposed framework is not tailor to 3D humans and holds potential to produce more diverse and realistic results for other 3D objects (e.g., cars, chairs). The exploration of generating common objects is a promising avenue worth pursuing. Additionally, our generative model establishes a manifold, which presents opportunities for attribute editing applications.

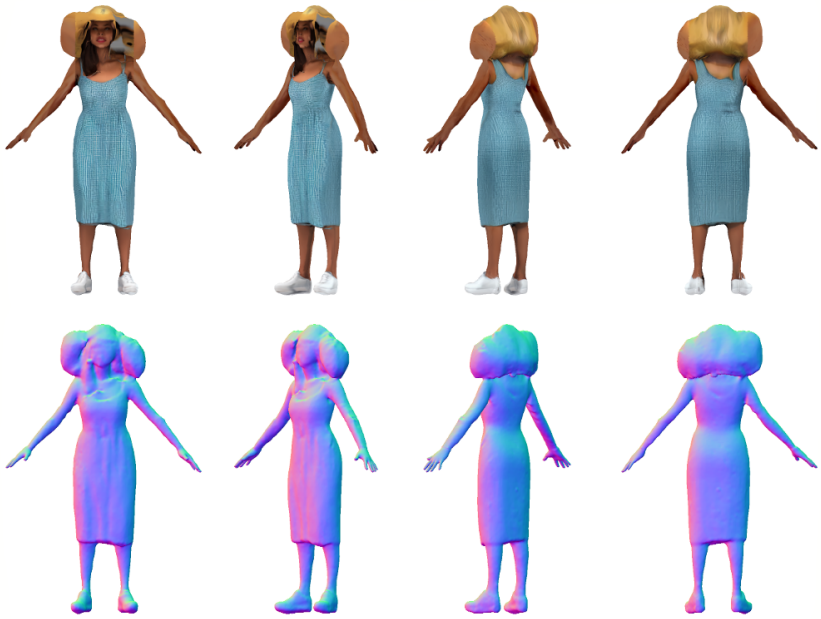


Figure 8: The failure case of 3D human wearing exaggerated accessories (e.g., hats).

## Reference


[1] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16123–16133, 2022.

[2] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Ot- mar Hilliges, and Andreas Geiger. Ag3d: Learning to gen- erate 3d avatars from 2d image collections. arXiv preprint arXiv:2305.02312, 2023.

[3] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and 602 Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. arXiv preprint arXiv:2210.04888, 2022.

[4] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. ACM Transac- tions on Graphics (TOG), 39(6):1–14, 2020.

[5] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid represen- tation for high-resolution 3D shape synthesis. In Advances in Neural Information Processing Systems (NeurIPS), 2021.