Dress-up: Generating Animatable Clothed 3D Humans via Latent Modeling of 3D Gaussian Texture Maps

Kim Youwang[1] Lee Hyoseok[2] Gerard Pons-Moll[3,4,5] Tae-Hyun Oh[2]
[1]Dept. of Electrical Engineering, POSTECH [2]School of Computing, KAIST
[3]University of Tübingen [4]Tübingen AI Center, Germany [5]Max Planck Institute for Informatics, Germany

Figure 1. ***Dress-up*** generates unlimited clothed 3D humans in diverse identities and clothing by sampling from a learned latent space of 3D Gaussian UV texture maps. The generated 3D clothed humans can be animated by skinned parametric 3D human poses for applications.

Abstract

We present Dress-up, a generative framework for creating diverse, animatable 3D human avatars with novel identities and clothing. Unlike prior methods that mainly produce in-domain results with limited variation, Dress-up synthesizes high-fidelity 3D humans with diverse identities and clothing, achieved via efficient latent generative modeling and leveraging multi-view 3D captures spanning a wide range of identities, poses, and outfits. Specifically, we design a latent space modeling of clothed 3D human Gaussian texture maps with a latent diffusion model for realistic clothing and appearance generation. The framework ensures multi-view geometric and texture consistency, while remaining robust to novel poses for animation. Dress-up can generate realistic, fully animatable avatars within ∼5 seconds, supporting real-time deployment. This can enable a wide range of creative and immersive applications, from virtual production to interactive VR/AR experiences.

1. Introduction

Generating realistic 3D virtual human avatars is the key technology facilitating immersive applications in industries such as movies, games, online commercials and VR/AR. While recent advances in generative models significantly enhanced the realism of the generated human images [4, 9, 10], generating 3D human avatars poses distinctive technical challenges, *e.g.*, ensuring multi-view consistency for geometry and appearance or achieving robustness to novel poses since the avatars could be animated. Moreover, the vast diversity of 3D human body shapes, textures, and clothing further com-

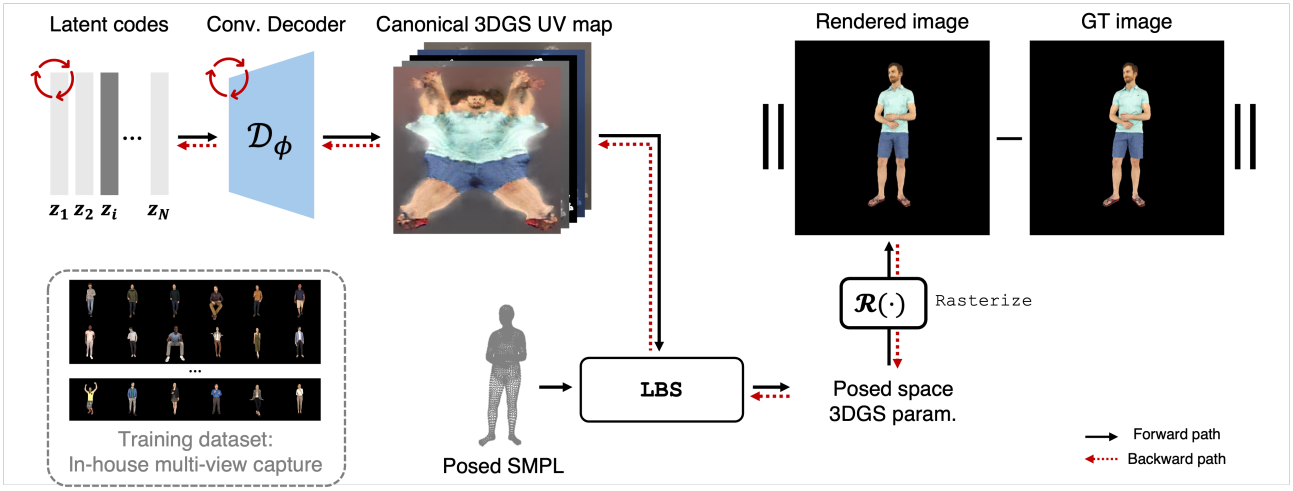

Figure 2. **Training latent space of 3D Gaussian (3DGS) UV texture maps.** We design an auto-decoder for modeling the latent space of clothed 3D human textures and geometries. Given the per-identity latent code as an input, a shared convolutional decoder outputs 3DGS UV texture maps, which are then attached to the posed SMPL mesh surface points and rasterized. By back-propagating the photometric loss between the multi-view rendered human images and the ground-truth human capture images, we jointly train the per-identity latent codes and a shared 3DGS decoder. We use in-house multi-view human capture images (∼2K identities) for training.

plicates the task. To mitigate such challenges, generative methods for 3D human avatars have been extensively studied [1, 3, 6–8, 11, 12, 15–17]. While these methods generate plausible-looking 3D humans by leveraging prior knowledge from domain-specific human image datasets [2, 13], the generated results are typically in-domain and cannot generate diverse identities or clothing.

In this work, we aim to enhance the diversity of generated clothed human avatars. We propose *Dress-up*, a system that generates an unlimited number of animatable clothed 3D humans in novel identities and new clothing. *Dress-up* is composed of two modules: 1) an auto-decoder that learns the latent space for clothed 3D humans, and 2) a latent diffusion model that learns the distribution of realistic 3D human clothing and appearances. For training the core modules of *Dress-up*, we leverage high-fidelity, 3D multi-view captures of diverse clothed humans, which cover human identities with broad age, gender, poses, and clothing. Once trained, *Dress-up* unconditionally generate novel human identities under new clothing within ~ 5 seconds, and can be animated in real-time for application scenarios (see Fig. 1).

2. *Dress-up*: Latent generative model for animatable clothed 3D humans

We propose *Dress-up*, a system that unconditionally generates clothed 3D humans, represented as 3D Gaussian UV texture maps, animatable with a parametric human body model. To facilitate the efficient generation process, we employ a latent generative modeling framework. We first train a compact latent space of 3D Gaussian UV texture maps, then

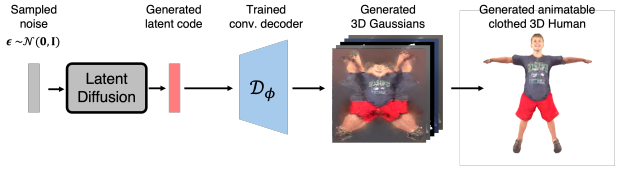

Figure 3. **3D human sampling from the learned latent diffusion.**

train a latent diffusion model to learn the data distribution of realistic 3D Gaussian UV texture maps.

Latent space modeling of 3D clothed humans. We design an auto-decoder for learning the latent space of realistic 3DGS UV texture maps (see Fig. 2).

Given a multi-view human capture image dataset composed of N identities and V views, we construct a latent codebook $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, where $\mathbf{z}_i \in \mathbb{R}^d$ refers to a dedicated latent code for the identity i. We design a convolution-based decoder $\mathcal{D}_\phi$ shared across all the identities. The decoder $\mathcal{D}_\phi$ gets a latent code $\mathbf{z}_i$ as an input and outputs canonical pose 3DGS UV texture maps $\mathbf{M}_i = [\mathbf{c}_i, \mathbf{s}_i, \mathbf{q}_i, \delta\mathbf{t}_i] \in \mathbb{R}^{(3+1+4+3)\times H\times W}$, where $\mathbf{c}_i, \mathbf{s}_i, \mathbf{q}_i$ and $\delta\mathbf{t}_i$ denotes the 3DGS color, scale, rotation and displacement (from a skinned SMPL surface), respectively. We fix the 3DGS opacity as ones, *i.e.*, fully opaque, and set isotropic 3DGS, following [14]. We then transform the canonical 3DGS into the posed space and rasterize to obtain the rendered images. Finally, we compute photometric loss and jointly optimize the codebook $\mathbf{z}$ and the shared decoder $\mathcal{D}_\phi$. The auto-decoder training loss $\mathcal{L}_{\text{AD}}$ is written as:

$$\mathcal{L}_{\text{AD}} = \sum_{i=1}^{N}\sum_{v=1}^{V} \|\mathbf{I}_{i,v} - \mathcal{R}\left(\boldsymbol{\theta}, \mathcal{D}_\phi(\mathbf{z}_i)\right)\|_1 + \|\mathbf{z}_i\|_2^2,$$

Figure 4. **Generated clothed 3D human avatars with novel identities and diverse outfits.**

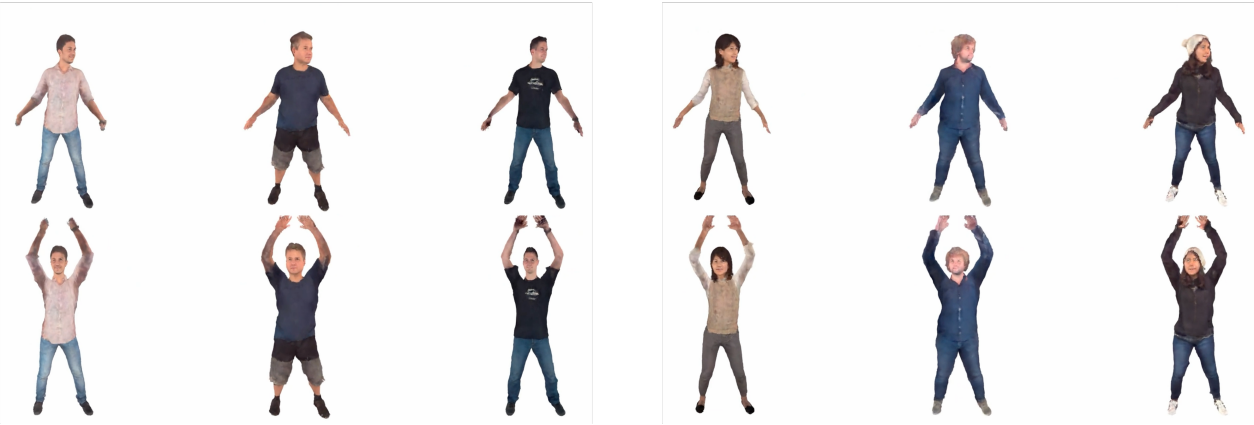

Figure 5. **Generated clothed 3D human avatars in animated poses.**

where $\mathbf{I}_{i,v}$ is the ground-truth capture image of the identity i in view v, $\mathcal{R}(\cdot)$ is the differentiable 3DGS rasterizer, $\boldsymbol{\theta}$ is the SMPL pose parameter, and $\|\mathbf{z}_i\|_2^2$ is a regularization term for encouraging the latent codes to stay near $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Latent diffusion for 3D human generation. After training the latent space and the 3DGS UV texture map decoder, we train the latent diffusion model that enables sampling from the realistic latent space of 3D humans. Specifically, given the trained latent codebook $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, we train a diffusion model with the noise prediction objective [5]:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{\mathbf{z}_i,\, \epsilon,\, t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\alpha_t}\,\mathbf{z}_i + \sqrt{1-\alpha_t}\,\epsilon,\, t\right)\right\|_2^2\right],$$

where ϵ_θ is the noise prediction network and α_t denotes the variance schedule at timestep t. After training, we sample a random noise ϵ and obtain the sampled latent code, which is then queried to the trained 3DGS decoder $\mathcal{D}_\phi$ to generate an animatable clothed 3D human avatar (see Fig. 3).

3. Results

In Figs. 4 & 5, we visualize the generated human avatars in the canonical pose and animated poses. The qualitative results demonstrate that *Dress-up* can generate highly realistic, animatable, clothed 3D human avatars with diverse ages, genders, body shapes, and outfits.

4. Conclusion

We present *Dress-up*, a generative model for animatable clothed 3D human avatar synthesis. With the latent space modeling of 3DGS texture maps for clothed 3D humans in the auto-decoding formulation, along with the usage of a high-quality real human scan dataset, *Dress-up* can generate realistic clothed 3D human avatars that can be driven with SMPL pose parameters. Since this work is in progress, we plan to extend *Dress-up* to support multi-modal generation, *e.g.*, text-conditioned generation, in our future work.

Acknowledgment. Kim Youwang, Lee Hyoseok and Tae-Hyun Oh were supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02263036, Generative AI-based Pre-visualization Technology for Media Production Coordination, Contribution Rate: 50%) and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of barrier-free experiential XR contents technology to improve accessibility to online activities for the physically disabled, Project Number: RS-2024-00396700, Contribution Rate: 50%). Gerard Pons-Moll is a Professor at the University of Tübingen endowed by the Carl Zeiss Foundation, at the Department of Computer Science and a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645, and is funded by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The project was made possible by funding from the Carl Zeiss Foundation. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] Dan Casas and Marc Comino-Trinidad. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *British Machine Vision Conference (BMVC)*, 2023. 2

[2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[3] Zijian Dong, Xu Chen, Jinlong Yang, Michael J.Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[4] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[6] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4):1–19, 2022. 2

[7] 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 (ICLR)*, 2023.

[8] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[9] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. 2023. 1

[10] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[11] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *International Conference on 3D Vision (3DV)*, 2023. 2

[12] Yufei Liu, Junwei Zhu, Junshu Tang, Shijie Zhang, Jiangning Zhang, Weijian Cao, Chengjie Wang, Yunsheng Wu, and Dongjin Huang. Texdreamer: Towards zero-shot high-fidelity 3d human texture generation. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[13] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[14] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[15] Soubhik Sanyal, Partha Ghosh, Jinlong Yang, Michael J. Black, Justus Thies, and Timo Bolkart. SCULPT: Shape-conditioned unpaired learning of pose-dependent clothed and textured human meshes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[16] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. CLIP-Actor: Text-driven recommendation and stylization for animating human meshes. In *European Conference on Computer Vision (ECCV)*, 2022.

[17] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2