Gaussian Shell Maps for Efficient 3D Human Generation

Rameen Abdal[*1] Wang Yifan[*1] Zifan Shi[*†1,2] Yinghao Xu[1] Ryan Po[1] Zhengfei Kuang[1]

Qifeng Chen[2] Dit-Yan Yeung[2] Gordon Wetzstein[1]

[1]Stanford University [2]HKUST

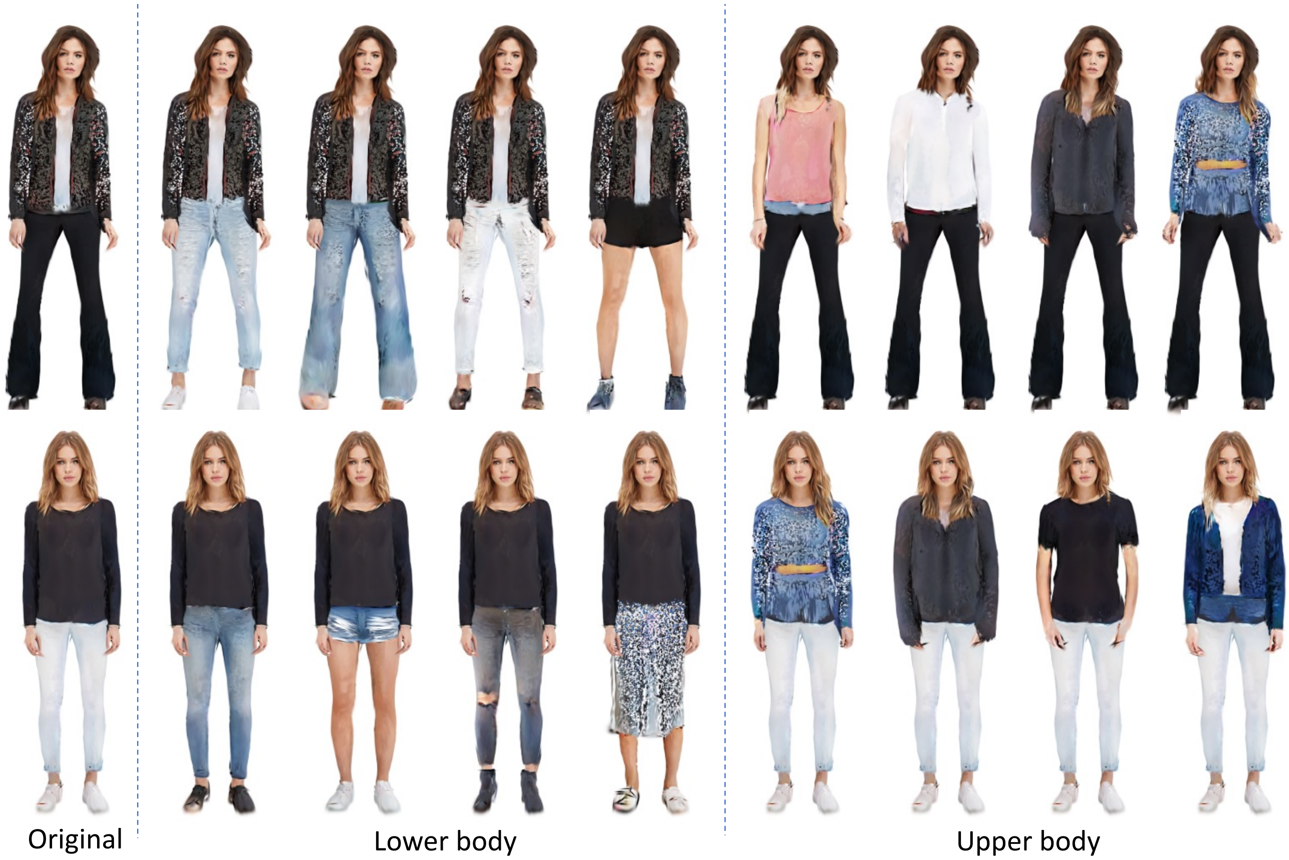

Figure S1. **Appearance Editing.** 3D Gaussians offer an explicit representation, thereby facilitating convenient post-generation editing. In this example, we demonstrate swapping the clothing of two generated identities. Please refer to Appendix C.3 for further details.

Contents

A. Further Analysis

A.1. Types of Gaussian Anchors.

GSM anchors the 3D Gaussians on shell meshes. We evaluated multiple alternative ways to anchor the Gaussians,

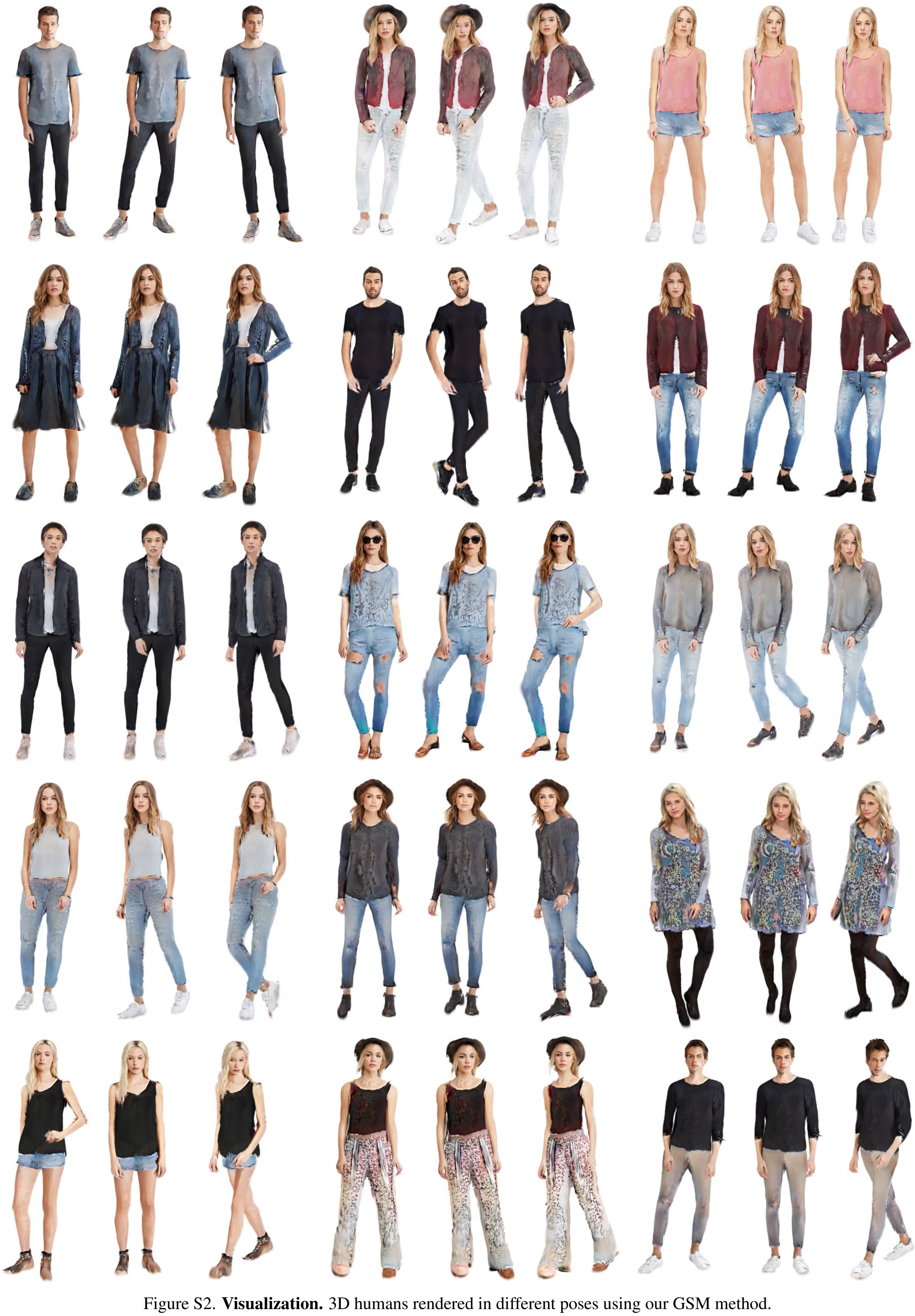

Figure S2. **Visualization.** 3D humans rendered in different poses using our GSM method.

testing the following three methods at 128^2 resolution after training with $1.6k$ Kimgs: (i) **in bounding box**: The Gaussians are uniformly sampled within the bounding box of the 3D human mesh, with Gaussian features interpolated from axis-aligned triplane features. This variant differs from our proposed GSM as it does not utilize the shell map to learn features in texture space. Instead, 3D Gaussian features are learned in world space, requiring the generator to also model the distribution of diverse human body poses. With this variant, we demonstrate the importance of using the shell map. (ii) **on a single shell with learned offset**: This variant samples only on the base mesh, the SMPL mesh, but allows deviations from the mesh template by applying a learned offset per Gaussian, predicted by the generator as part of the feature textures. This approach emulates the typical pipeline of existing 3D human GANs, where clothing and hair are captured by offsetting the template unclothed mesh. (iii) **in tets**: The Gaussians are sampled not only on the shell meshes but also in between them in tetrahedra, constructed by connecting mesh vertices. This variant is more akin to the original Shell Map proposed by Porumbescu *et al*. As shown in Table S1, the bounding box variant underperforms, as the generator struggles to handle deformation jointly with appearance. Learning offsets to model surface details different from the template mesh yields subpar quality. This suggests that varying the Gaussians' positions complicates the already non-convex optimization problem, as the positions are highly correlated with the rest of the Gaussian properties. Finally, sampling in tets shows slow convergence and does not improve the FID. Additionally, this model exhibits a slower rendering speed (speed for deformation and rasterization for a generated identity) of 20 ms/img versus 9 ms/img.

Table S1. **Anchoring types.** Ablation on the anchoring type performed on 128^2 resolution trained for $1.6k$ KImgs on SHHQ.

Anchoring	bbox	tets	learned offset	triangles (proposed)
FID ↓	63.90	24.66	29.30	**20.63**

A.2. Sampling Densities

We evaluate the effect of the number of Gaussians on the generation quality. For this study, we train on 512^2 resolution and evaluate the FID score after training with $10k$ KImgs. Since the sampling density will affect the Gaussian scale, we adjust the scaling regularization and initialization accordingly. As shown in Table S2, using 100K Gaussians yields empirically the best result in terms of FID for the SHHQ dataset. Using $50k$ Gaussian samples yields the highest FID, suggesting that Gaussians are likely too few to fully model the appearance complexity exhibited in the dataset. On the other hand, using too many Gaussians, *e.g.* $200k$, can harm the FID. We observe that this drop under a high sampling density scenario is due to the tendency of Gaussians to learn small scales while modeling high-frequency details. This adds complexity to the already challenging task of optimizing opacity and scaling. As a result, we might notice unwanted dotted patterns, especially in cloth areas. The FID score easily detects such an unnatural appearance.

Table S2. **Ablation: Number of Gaussians.** Ablation on the number of Gaussians performed on 512^2 resolution trained for 10K KImgs on SHHQ dataset.

Number	$50k$	$100k$	$200k$
FID ↓	23.83	13.30	19.96

A.3. Relation and Comparison with LSV-GAN

Concurrent work, LSV-GAN [8], also employs shell meshes and rasterization to efficiently model diverse human shapes and appearances. Our approach, however, distinguishes itself from LSV-GAN by populating the shell meshes with 3D Gaussians and employing differentiable Gaussian Splatting for rendering [3]. The spatial span of these Gaussians fills the space between shells with continuous functions. As illustrated in Figure S6 and Figure S3, LSV-GAN often exhibits artifacts at silhouette boundaries due to its discontinuous representation of shell volume. In contrast, our method yields smoother and more natural boundaries. Moreover, the utilization of 3D Gaussians allows us to define RGB and alpha values beyond the boundary shell, effectively expanding our capability to model deviations from the template mesh. This leads to more varied geometry and appearances of the human body, including loose clothing and accessories, as seen in Figure 1 of the main manuscript and Figure S2 in this document. To further substantiate this, we conducted a new user study (see Figure S5), wherein 138 participants from Amazon Mechanical Turk assessed 46 images generated by both methods under identical conditions. A significant majority (78.26%) favored our method, citing improved realism in faces (79 users), clothing (44 users), hands (32 users), and feet (27 users) as key factors. We believe the gap between FID and human evaluation shows that FID neglects geometric distortion and non-local artifacts. We would also like to point out that LSV generates separated fingers. It is a result of the naive workaround to address the difficulty of modeling more complicated concave geometry using discrete offsetted SMPL meshes. As written in their paper, they opt for a single layer for the fingers, leading to all generated results having unnaturally separated fingers that deviate from the data distribution (see Figure S4).

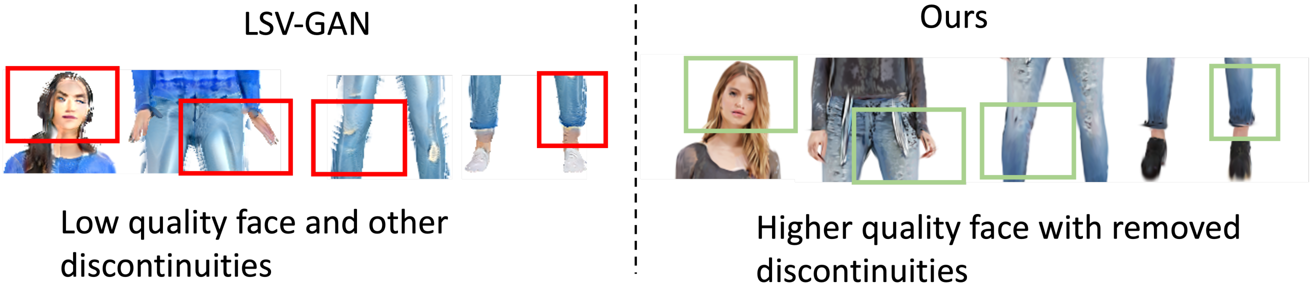
LSV-GAN
Ours
Low quality face and other discontinuities
Higher quality face with removed discontinuities

Figure S3. **LSV Comparison.** LSV-GAN vs. Ours (without curation)

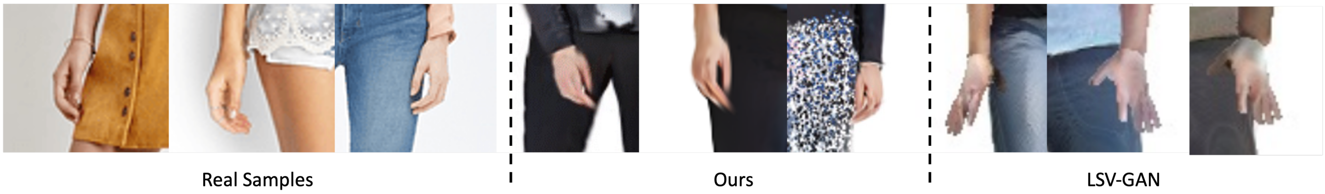

Figure S4. **LSV Comparison - Hands.** Our hands align better with the dataset. LSV-GAN produces artifacts.

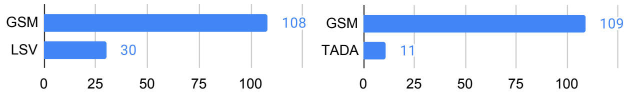

Figure S5. **User Study.** User preference comparing against LSV (left) and TADA (right)

B. Comparison to Diffusion Models

While diffusion-based zero-shot synthesis approaches are very popular [4, 5], the GANs-based methods, such as ours, have their undisputed advantages, including much higher generation speed (28ms with GSM vs several hours using TADA) and, more importantly, significantly more realistic appearance and pose accuracy (see Figure S5 and Figure S7).

C. Additional Qualitative Results

In this section, we present further qualitative results. All visual examples have been sampled using the truncation technique detailed in EG3D [1]. Additional animated results can be found in the supplementary HTML webpage.

C.1. Random Samples

To showcase the quality and diversity of our GSM method, we display randomly sampled results under identical poses in Figure S8. For both the DeepFashion [6] and SHHQ [2] datasets, our method successfully generates a variety of body shapes, accessories like hats, loose clothing, and intricate details on clothes.

C.2. Articulation and Novel View Rendering

In the accompanying HTML webpage, we present videos demonstrating articulation and novel view rendering results. The articulation sequences are provided by the AMASS [7] dataset. Notably, our method avoids the temporal flickering artifacts common in other models, as it directly renders at the target resolution. This efficiency is due to our use of rasterization instead of the more costly volumetric rendering approach.

C.3. Appearance Editing

A significant advantage of explicit representations like 3D Gaussians, especially when compared to implicit representations such as radiance fields, is their enhanced editability. In Figure S1, we illustrate this benefit through a redressing application, where we interchange the upper and lower body appearances between multiple generated instances. This editing process involves selecting Gaussians within a specific region (*e.g.*, lower or upper body) and then swapping their properties with those from another instance. This method is feasible because the Gaussian positions are anchored on the shell meshes and remain consistent across instances, with the appearance being defined solely by their properties.

References

[1] 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 *CVPR*, 2022. 4

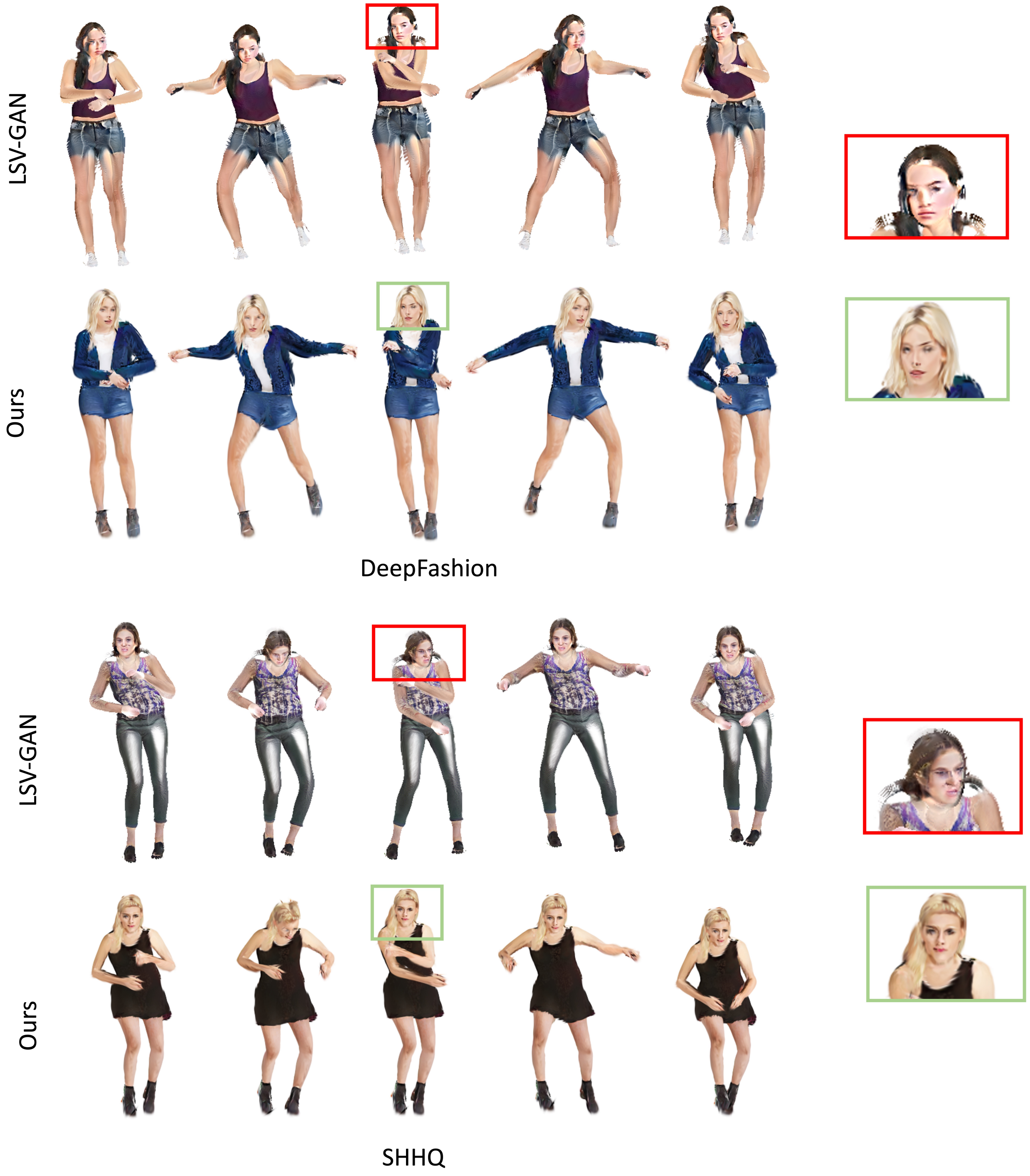

Figure S6. **LSV comparison.** LSV-GAN [8] suffers from discontinuities and facial artifacts in DeepFashion and background bleeding into textures in SHHQ. In comparison, our results show high-quality facial details and consistency.

[2] 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 *ECCV*, 2022. 4, 7

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 3

[4] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Ed-

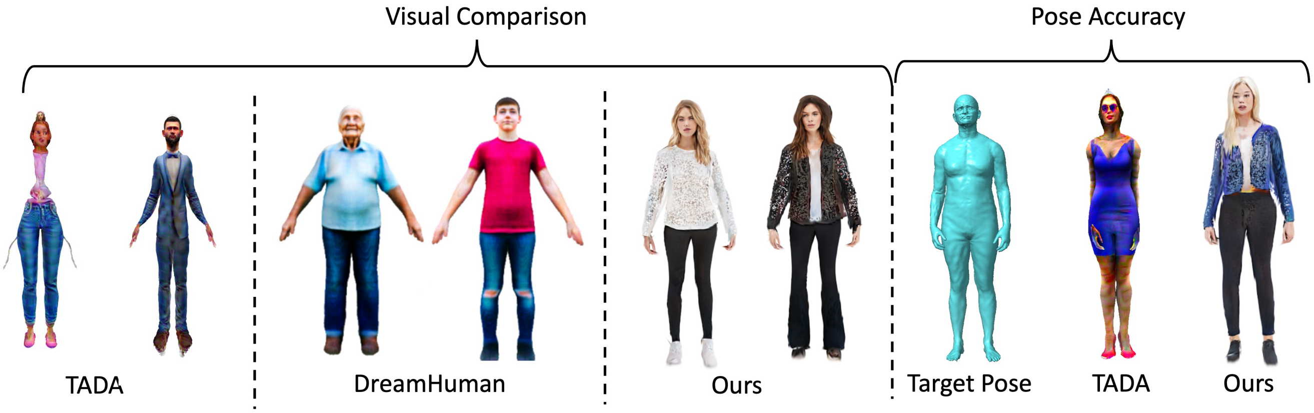

Figure S7. Results comparing TADA, DreamHuman, and ours.

uard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *ArXiv*, abs/2306.09329, 2023. 4

[5] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 4

[6] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 4, 7

[7] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 4

[8] Yinghao Xu, Wang Yifan, Alexander W. Bergman, Menglei Chai, Bolei Zhou, and Gordon Wetzstein. Efficient 3d articulated human generation with layered surface volumes. In *3DV*, 2024. 3, 5

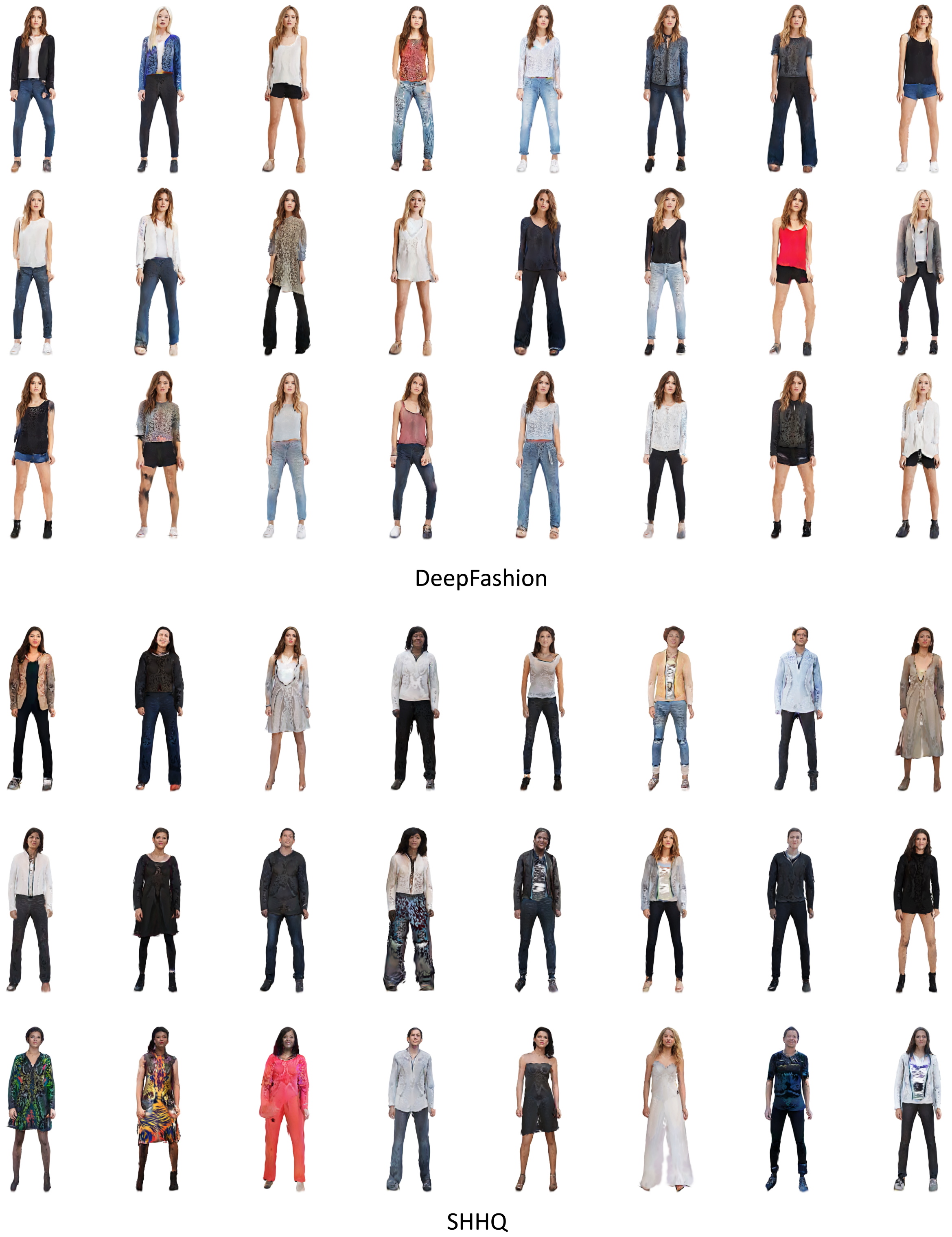

Figure S8. **Random Samples.** Randomly generated samples of 3D humans under same pose using or GSM method trained on DeepFashion [6] and SHHQ [2] datasets, without truncation.