

SinGS: Animatable Single-Image Human Gaussian Splats with Kinematic Priors

Supplementary Material

Yufan Wu^{1*} Xuanhong Chen^{1†} Wen Li^{1,3} Shunran Jia² Hualiang Wei^{1*} Kairui Feng⁴
Jialiang Chen¹ Yuhan Li¹ Ang He^{1*} Weimin Zhang¹ Bingbing Ni¹ Wenjun Zhang¹

¹Shanghai Jiao Tong University, Shanghai, China ²DreamX Inc ³Akool Research

⁴The National Key Laboratory of Autonomous Intelligent Unmanned Systems, Shanghai, China

{chen19910528,wmzhang,nibingbing,zhangwenjun}@sjtu.edu.cn, elvinfkr@tongji.edu.cn

In this supplementary material, we have provided additional experimental results, detailed analyses, video outcomes, and more, which could not be included in the main text due to space limitations. **We strongly recommend that readers watch our result video.** The main content of the supplementary material is summarized as follows:

- Section 1: Details of the 3D Pose Encoder
- Section 2: More Results of Single-Image 3D Avatar Reconstruction
- Section 3: Details of Geometry-Preserving 3D Gaussian Splatting
- Section 4: Details of Optimization losses
- Section 5: Animation Results

1. Details of the 3D Pose Encoder

Due to the limitations of 2D pose estimation methods (e.g., OpenPose, DWPose) in effectively representing the human body in three-dimensional space, our Kinematic Human Diffusion model employs SMPL as input to guide the generation of various poses within 3D space, facilitating 3D reconstruction. We begin by rendering the SMPL mesh into an image, which is then processed by a 3D Pose Encoder to convert the image into a latent code of dimensions $\mathbb{R}^{8 \times \frac{H}{8} \times \frac{W}{8}}$. This latent code is subsequently concatenated with noise of the same size along the channel dimension. In our experiments, the height H is set to 768 pixels, the width W to 512 pixels, and the number of channels C for the SMPL-corresponding images is 3. As depicted in Figure 3, we present the architecture of the 3D Pose Encoder, which is trained jointly with the entire pipeline. The input layer of the 3D Pose Encoder consists of a single convolutional layer designed to extract low-level information from the image. Subsequently, the features are downsampled twice through two cascaded residual blocks [2]. Finally, the features pass through another convolutional layer

to transform the dimensionality to $8 \times \frac{H}{8} \times \frac{W}{8}$. To enhance the performance of our network, we have adopted the SiLU activation function [1]. This choice has proven to be effective in improving the network’s capabilities.

2. More Results for Single-Image 3D Avatar Reconstruction

In this section, we present additional single-image reconstruction results for comparison. Given that the primary objective of our method is to create animatable 3D human avatars, our comparative strategy involves reposing the reconstructed 3D avatars into T-Pose or A-Pose for comparison. This approach allows us to showcase both the reconstruction and animation capabilities within a single result. It is important to note that the majority of existing methods do not provide code that supports animation after single-image reconstruction, or they are incapable of animation altogether. Consequently, the results in Figure 1 and Figure 2 demonstrate our enhanced animation functionality, which we have added based on the official open-source code. We have also observed that most methods perform poorly in terms of animation capabilities within the single-image reconstruction setting. We have included additional single-image reconstruction results in Figure 1 and Figure 2, where it is evident that our pipeline has a significant advantage in both reconstruction and animation capabilities.

3. Details of Geometry-Preserving 3D Gaussian Splatting

Semantic Laplacian Regularization. Although Laplacian regularization can effectively reduce the problem of floating Gaussian spheres caused by information inconsistency, an excessively high Laplacian can cause the Gaussian sphere to concentrate too much, leading to collapse and a complete loss of geometric shape. There is a significant structural difference in different parts of the human body; for

*Work done during an internship at Shanghai Jiao Tong University.

†Corresponding author: Xuanhong Chen.



Figure 1. In our comparison with SiTH and TeCH, the images presented are all of males, and it can be observed that our method has a significant advantage in terms of reconstruction shape and texture consistency. It’s important to note that we chose to compare with SiTH and TeCH because even after using third-party tools like Repose, their results remain within an acceptable range. Other methods face significant challenges when attempting single-image reconstruction with Repose directly.

example, the hands are intricate and complex, while the abdomen is simple and flat. If a homogeneous Laplacian is applied to both, it would result in a poor representational capability of the overall 3D model. Therefore, we propose an Anisotropic Semantic Laplacian Regularization to resolve these contradictions.

Semantic Parts. Semantic Laplacian regularization is highly dependent on the division of semantic regions. We have readjusted the official semantic region division of SMPL¹, removing overly fragmented divisions, and ultimately formed 15 semantic regions. The visualization results are shown in Figure 4.

4. Optimization Loss

After we use KHD to obtain the results of augmenting a dynamic single image into a dynamic video, we utilize this video to train our GPGS. We employ the following losses to train our model, which include the commonly used reconstruction loss and the Semantic Laplacian Regularization that we propose:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{RGB} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{LPIPS} + \lambda_4 \mathcal{L}_{SLR}, \quad (1)$$

¹https://github.com/Meshcapade/wiki/blob/main/assets/SMPL_body_segmentation/smpl/smpl_segmentation_on_template.png

where \mathcal{L}_{RGB} represents the pixel-level \mathcal{L}_1 reconstruction loss, while \mathcal{L}_{SSIM} and \mathcal{L}_{LPIPS} are both reconstruction losses aimed at improving the performance of high-frequency detail reconstruction. \mathcal{L}_{SLR} is the Semantic Laplacian Regularization that we propose, which is designed to constrain the distribution of Gaussian spheres, making them more compact and reducing the occurrence of floating Gaussian spheres.

5. Animation Results

Natively supporting animation is a key feature of our SinGS system. Unlike other single-image 3D avatar systems such as SiTH [3], TeCH [4], which require manual editing using repose tools for frame-by-frame reposing, our SinGS can achieve animation directly without the need for any third-party repose toolboxes². We only require input in the form of SMPL poses or 3D skeleton sequences to bring our avatars to life with animation. **We strongly encourage readers to watch the video in our supplementary materials to appreciate the animation capabilities of our method.**

²<https://github.com/custom-humans/editable-humans>

References

- [1] Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. [1](#)
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778. IEEE Computer Society, 2016. [1](#)
- [3] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR 2024*, pages 538–549. IEEE, 2024. [2](#)
- [4] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *3DV 2024*, pages 1531–1542. IEEE, 2024. [2](#)

SiTH(CVPR2024)



TeCH(3DV2024)



SinGS(Ours)



Figure 2. In our comparison with SiTH and TeCH, the images presented are all of female, and it can be observed that our method has a significant advantage in terms of reconstruction shape and texture consistency. It's important to note that we chose to compare with SiTH and TeCH because even after using third-party tools like Repose, their results remain within an acceptable range. Other methods face significant challenges when attempting single-image reconstruction with Repose directly.

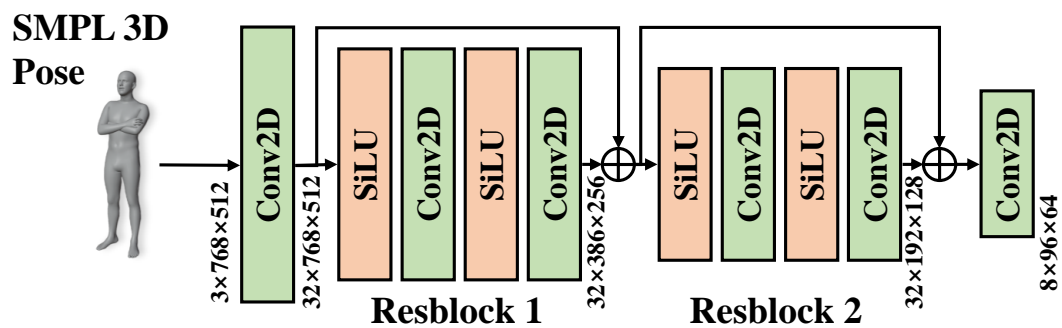


Figure 3. Detailed architecture of the proposed 3D Pose Encoder.



Figure 4. Semantic regions of our framework.