AAGaussian: Animable Avatar Generator with Portrait and Effect Prompt

Team Member 1*
Shanghai Jiao Tong University

Team Member 2*
Shanghai Jiao Tong University

Team Member 3*
Shanghai Jiao Tong University

ABSTRACT

Creating realistic, drivable 3D virtual human avatars is a valuable and challenging task. Previous works faced issues such as lack of detail, slow training, and inability to generate new postures. This paper proposes an efficient AAGaussian framework based on the emerging 3D Gaussian scene representation, designed to create virtual human avatars with intricate structures and realistic appearances. Specifically, building upon existing work, we imports SMPL-X as a body type and gender prior, and designs a facial rendering loss function to enhance the naturalness of faces. Experiments prove that our framework can produce efficient and high-quality results for various inputs, allowing users to craft lifelike and diverse virtual avatars. This framework also includes an animation generator that can add smooth animations to the generated virtual figures based on motion video input.

KEYWORDS

drivable avatar, animation generator, 3D Gaussians

1 INTRODUCTION

Creating high-quality 3D human avatars based on user specifications and needs has become a cornerstone in numerous virtual reality applications, such as virtual try-on, medical imaging, and VR gaming, increasingly drawing focus in the field of computer graphics research. A fundamental aspect of this task is the accurate reconstruction of human figures, typically regressing the 3D representation from dense multi-view data inputs. This process often utilizes Neural Radiance Fields (NeRFs), with extensive research targeting optimization issues like slow NeRF rendering and suboptimal surface effects [3, 23, 24, 36]. Some studies use monocular inputs to extract human features, but face challenges in accuracy, resolution, texture, and color. Recent research by Kerbi et al. introduced 3D Gaussian Splatting [13], rendering higher quality images faster than NeRFs without the need for highly accurate imaging.

A crucial characteristic of virtual avatars is their drivability, notably the ability to generate animations of new poses and movements. Originally, NeRFs and Gaussians were designed for static scenes, with traditional methods involving manually crafted processes to rig, skin, texture, and animate [2, 11] the derived 3D models. Some studies introduced temporal representations to NeRFs and Gaussians [39], but these models typically only replay previously observed content and are not suited for representing new movements. By integrating parametric human body models from overall body posture modeling [22] to detailed modeling of parts like face, hands, and jaw [17, 26], some work [8, 20, 42] has incorporated parametric human models as priors into NeRFs and Gaussians. These approach has enhanced the detail in human reconstruction models, creating animations using learned deformation fields, skeletal rigging, and parameter sequences.

Another facet of drivability is the ability to modify the avatar as per user requirements, like changing clothes. Recent text-to-3D models [5, 20, 29, 40] use 2D diffusion as a guiding model, incorporating SDS guidance to optimize NeRFs and Gaussians. This approach has been effective for simple objects but faces challenges with complex requirements like human generation. Firstly, it cannot utilize the user's original physical characteristics (e.g., body type, facial features); secondly, the modeling results may suffer from issues like naturalness and multi-view coherency.

In this paper, we propose a framework, **AAGaussian**, designed to create virtual human avatars with intricate structures and realistic appearances, based on the joint input of user portraits and specified effect prompts. Our approach leverages the efficient training efficiency of 3D Gaussian and its sensitivity to structural guidance. We expand upon the existing text-to-3D work of HumanGaussian [20], drawing on its pre-trained SDS and annealing negative prompt guidance. In our method, Gaussians are anchored to the SMPL-X [26] mesh processed through the SMPLify-X method. Information such as body type and gender is extracted from user portraits to guide the subsequent training process. To address the issue of facial fragmentation in the generated avatars, we extracted facial features from user portraits and designed a render loss to supervise the optimization of Gaussians for improvement and introduced this loss with varying weights. Users have the option to guide the virtual facial generation using either prompts or their own photographs. Additionally, we have developed an application that extracts pose sequences from motion videos to animate the resulting avatar representations.

In summary, our main contributions are as follows:

- We propose the **AAGaussian** framework, which creates high-quality virtual avatars based on joint inputs from images and text prompts.
- We initialize Gaussians with SMPL-X parameters fitted by SMPLify-X, extracting user body shape features.
- We introduce facial rendering loss, combined with dual-branch SDS, to extract user facial information.
- We develop an application that transfers pose sequences from motion videos, creating animations for the generated virtual avatars.

2 RELATED WORKS

2.1 3D Neural Representations

Traditional 3D scene representations such as voxel, point cloud and mesh each have their own problems, including low rendering resolution, difficulty in capturing fine details, and inability to create

*These authors contributed to the work equally.

accurate topological structures, posing challenges in 3D reconstruction tasks. NeRF [23] first introduced neural radiance fields, training an implicit function that continuously maps the coordinates and observation directions of points in space to color and volume density. This allows for modeling complex scenes at any resolution, but its optimization and inference processes are too slow to do real-time rendering with few exceptions. Subsequent work has focused on optimizing both effectiveness and efficiency, such as using surface encoding instead of volume encoding for accurate surface modeling in NeuS [36] and NeuDA [3], and introducing spatial multi-resolution hash encoding for minute-level convergence in the training process with instantNGP [24]. Recently, 3D Gaussian Splatting [13] has shown impressive results in the 3D reconstruction, surpassing previous representations with higher quality and faster convergence. In this work, we try to unlock the potential of 3D Gaussian splatting in the challenging arena of drivable 3D human avatars generation.

2.2 Text-to-3D Generation

Early efforts in text-to-3D generation predominantly utilized CLIP guidance [30] to enhance multi-view image-text alignment [10, 33] or employed 3D native pipelines for direct data capture [12]. More recent works have leveraged diffusion models, learning from rich priors in the 2D domain. For instance, DreamFusion [29] employs Score Distillation Sampling (SDS) to achieve unprecedented rendering quality in text-to-image models. Subsequent developments have introduced progressive learning strategies from coarse to fine [18], separated the learning of geometry and materials, structure and texture [4, 20], or integrated variational score distillation to refine SDS [37]. The long training time of NeRF also motivates concurrent works [5, 40] to adapt the representation of Gaussian splatting for text-to-3D generation.

Notably, some projects like SyncDreamer [21] have achieved spatially consistent text-to-multi-view image generation by learning joint multi-view probabilistic distributions and noise predictors based on diffusion. These are then converted to 3D scenes through naive 3D reconstruction methods such as NeRF.

2.3 Animatable Human Avatar

Statistical mesh templates such as SMPL [22] pioneered human body modelling, fitted to scans of numerous unclothed individuals, providing a parametric representation of human body shapes. These models, easily animated through graphic deformation tools such as Linear Blend Skinning (LBS) [16], laid the groundwork for more advanced techniques. In the realm of NeRF [23], Neural Body [28] anchors latent codes to the vertices of the SMPL model, while Neural Actor [19] and Animatable NeRF [27] each define a canonical NeRF, learning the correspondences from observation to canonical space. AvatarCLIP [8] integrates NeuS [36] with SMPL prior, guided by CLIP, and DreamHuman [15] employs a pose-conditioned NeRF based on imGHUM [1] to learn albedo and density fields using Score Distillation Sampling (SDS). HumanGaussian [20] fast-tracks human body reconstruction by anchoring Gaussian initialization to the SMPL-X [26] human template.

The complexity of facial features and expressions makes head avatars the most challenging aspect of character avatars. IMavatar [41] addresses this by learning a 3D morphable head avatar using neural implicit functions, solving the mapping from observed to canonical space through iterative root-finding. HeadNeRF [9] develops a NeRF-based parametric head model, leveraging 2D neural rendering for efficiency. INSTA [42] deforms query points to a canonical space by finding the nearest triangle on a FLAME [17] mesh, combining this approach with InstantNGP [24] for fast rendering.

3 METHODS

We present the **AAGaussian** framework, which generates avatars with superior geometry and appearance quality. The overall framework is depicted in Figure 1. To ensure the content is self-contained and the narration is clear, we first introduce some prerequisites in Section 3.1. Then, we propose methods to perceive and model user body types and facial details (Section 3.2 and 3.3). Finally, we implement an application that extracts pose sequences from motion videos and transfers them to virtual avatars to generate animations (Section 3.4).

3.1 Preliminaries

SMPL-X. SMPL-X [26] is a human model that jointly captures the body together with face and hands, which contains $N = 10,475$ vertices and $K = 54$ joints. The model is defined by the function $M(\theta, \beta, \psi)$, parameterized by pose θ, body shape β, and facial expressions ψ. The pose θ is further decomposed into jaw pose θ_f, finger pose θ_h, and body pose θ_b. The model formulation is given as:

$$\begin{aligned} M(\beta, \theta, \psi) &= W(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}), \\ T_p(\beta, \theta, \psi) &= \overline{T} + B_S(\beta) + B_E(\psi) + B_P(\theta), \end{aligned} \tag{1}$$

where $\overline{T}$ is the template shape mesh, and B_S, B_E, B_P are blend shape functions for shape, expression, and pose, respectively. $W(\cdot)$ is the linear blend skinning function [16] that transforms $T(\theta, \beta, \psi)$ into target pose θ, $J(\beta)$ and $\mathcal{W}$ are skeleton joints and blend weights defined on each vertex. Together presented with the model is SMPLify-X, an approach to fit SMPL-X to a single RGB image and 2D OpenPose joint detections.

3D Gaussian Splatting. Recently, 3D Gaussian Splatting(3DGS) [13] offers a novel technique in 3D reconstruction. Distinct from implicit methods like NeRF [23], it models scenes using anisotropic Gaussians. Each Gaussian is defined by a center position $\mu_i \in \mathbb{R}^3$, covariance $\Sigma_i \in \mathbb{R}^7$, color $c_i \in \mathbb{R}^3$, and opacity $\alpha_i \in \mathbb{R}$. These Gaussians are projected onto the camera's imaging plane, facilitating point-based rendering:

$$\begin{aligned} G(p, \mu_i, \Sigma_i) &= \exp\left(-\frac{1}{2}(p - \mu_i)^T \Sigma_i^{-1}(p - \mu_i)\right), \\ c(p) &= \sum_{i \in N} c_i \sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j), \quad \sigma_i = \alpha_i G(p, \mu_i, \Sigma_i), \end{aligned} \tag{2}$$

where p represents the pixel position. Here, μ_i, Σ_i, c_i, α_i, and σ_i denote the center, covariance, color, opacity, and density of the i-th Gaussian, respectively, and N is the set of Gaussians contributing

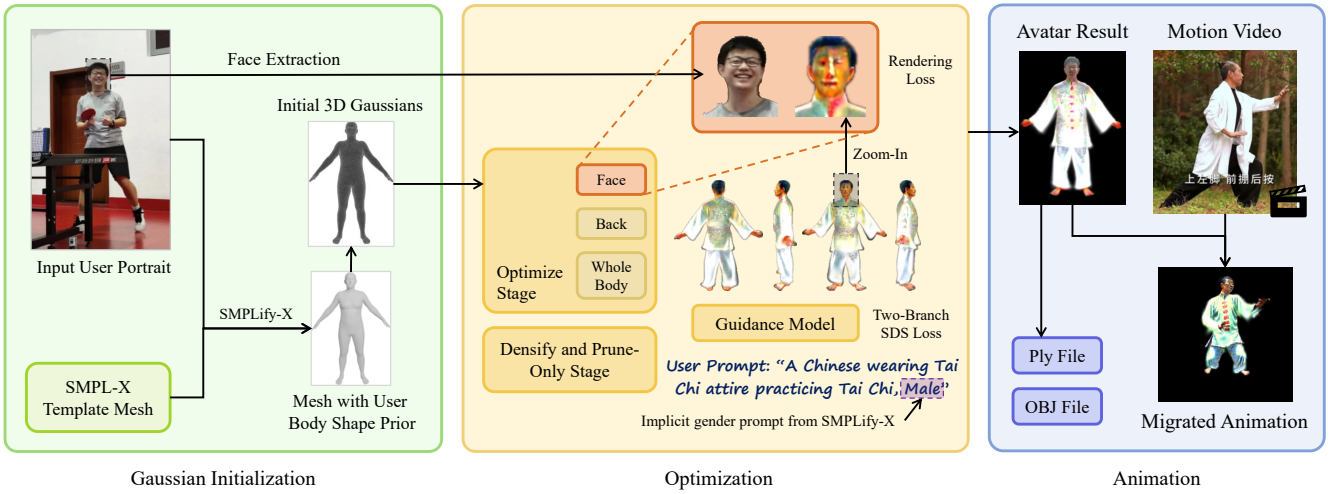

Figure 1: Our Pipeline

to a given tile. 3DGS enhances GPU-based rasterization processes with improved quality and efficiency.

Dual-Branch Score Distillation Sampling(SDS). To mitigate the workload involved in the creation of 3D assets with varied layouts, the seminal work DreamFusion [29] introduces Score Distillation Sampling (SDS), which is used to distill the 2D pre-trained diffusion prior for optimizing 3D representations. Specifically, we represent a 3D scene parameterized by θ and transform it into a corresponding image $x = g(\theta)$ using a differentiable rendering function $g(\cdot)$. By directing samples towards more densely populated areas within the real-data distribution at all noise levels, we enhance the fidelity of renderings from each camera view, aligning them more closely with the credible samples generated by the guidance diffusion model ϕ. DreamFusion uses Imagen [32] as the score estimation function $\epsilon_\phi(x_t; y)$, which predicts the sampled noise ϵ_ϕ using the noisy image x_t, text embedding y, and a specific timestep t. To optimize 3D scenes, SDS applies gradient descent relative to θ:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{\epsilon,t}\left[w_t\left(\epsilon_\phi(x_t; y) - \epsilon\right)\frac{\partial x}{\partial \theta}\right], \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is a Gaussian noise; the noised image is denoted as $x_t = \alpha_t x + \sigma_t \epsilon$. The variables α_t, σ_t, and w_t function as the controlling parameters for the noise schedule decided by the diffusion sampler.

As for 3D human generation, HumanGaussian [20] extends the pre-trained Stable Diffusion [31] to simultaneously denoise the image RGB and depth as SDS source model. HumanGaussian uses 3DGS [13] to represent 3D scenes and such dual-branch design guides 3DGS optimization process from both structural and textural aspects with high efficiency and quality. In addition to RGB rendering, depth map $d(p)$ are also introduced, which is calculated by accumulating depth values overlapping the pixel of N ordered Gaussian instances, using point-based α-blending:

$$d(p) = \sum_{i \in N} d_i \sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j), \quad \sigma_i = \alpha_i G(p, \mu_i, \Sigma_i), \tag{4}$$

where d_i is the depth of the i-th Gaussian center μ_i in the view, and $G(p, \mu_i, \Sigma_i)$ is its Gaussian value at point p. All depth maps d are normalized within $[0, 1]$. Finally, the integrated gradient descent for optimization is as follows:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{SDS} = \lambda_1 \cdot \mathbb{E}_{\epsilon_x,t}\left[w_t(\epsilon_\phi(x_t; p, y) - \epsilon_x)\frac{\partial x}{\partial \theta}\right] \\ + \lambda_2 \cdot \mathbb{E}_{\epsilon_d,t}\left[w_t(\epsilon_\phi(d_t; p, y) - \epsilon_d)\frac{\partial d}{\partial \theta}\right],\end{aligned} \tag{5}$$

where λ_1 and λ_2 balance the structural and textural effects, and $\epsilon_\phi(\cdot)$ are the ϵ-predictions from the extended pre-trained model.

3.2 Body Shape-Aware Modeling

Replacing implicit neural networks with explicit 3DGS [13] for the generation of 3D humans poses a challenge, as the training efficiency of 3DGS heavily relies on the initialization of the Gaussians' center position, even for the comparatively easier 3D reconstruction with dense multi-view image supervision. Besides, Our objective is to generate corresponding virtual avatars by inputting real-life 2D portraits of people and text descriptions specifying their desired appearance. Considering how to fully utilize the priors in the input photos is also a problem worth contemplating.

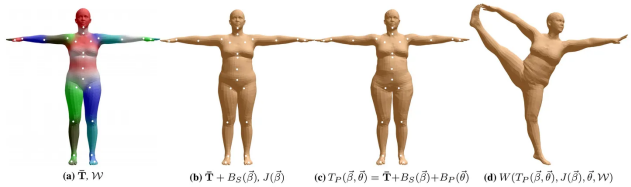

Figure 2: SMPL: Base Model of SMPL-X[1]

Gaussian Initialization with SMPL-X Prior. Previous studies [5, 13, 38, 40] in 3D scene reconstruction primarily used the Structure-from-Motion (SfM) points [34]; some text-to-3D research using

[1]This figure is from the original SMPL [22] paper.

Gaussian splatting initiated Gaussian positions in a fixed manner instead of randomly [35], or employed generic text-to-point-cloud priors like those in Shap-E [12]. However, such methods typically fall short in the human category, resulting in over-sparse points or incoherent body structures. In contrast, SMPL-X, an extension of SMPL [22], exhibits the following characteristics: **1)** It provides a basic human torso mesh shape which can be fine-tuned using shape blend and pose blend functions, making it convenient to utilize the body shape priors present in input human photographs; **2)** It complements the shape topology of the face and hands, which are beneficial for intricate human modeling with fine-grained details.Consequently, we propose to uniformly sample Gaussians on the SMPL-X mesh surface for initialization.

Specifically, We utilize SMPLify-x to regress SMPL-X parameters from input portrait photos, extracting the mesh at the third stage of the model (SMPL is the base model of SMPL-X, the extracted model resembles the mesh in Figure 2(c), augmented with jaw and hand poses θ_f, θ_h introduced by SMPL-X), thereby using the approximate height, body shape, and gender of the user portrait as prior information. On the mesh, Gaussians are instantiated with unit scaling, mean color, and no rotation, with the number of Gaussians significantly exceeding the number of vertices defined by SMPL-X, facilitating the fitting of finer details. Additionally, SMPLify-x also implements a gender classifier to identify the gender of individuals in input images, and in later parts of our pipeline, this gender information is implicitly incorporated into the prompt of the diffusion guiding model (SDS).

3.3 Face Detail-Aware Modeling

Upon testing existing text-to-3D models specialized for human characters [20], we observed that under certain prompts, the generated virtual avatars exhibited facial fragmentation. This issue presented as skin texture distortions in the middle of the face and a deep groove at the neck. In some cases, there were asymmetries in the left and right sides of the face, with discrepancies in skin tone and hairstyle. Additionally, in some anime-style cases, phenomena such as the disappearance of the nose and mouth were observed. Figure 3 distinctly showcases typical examples of the aforementioned errors.

(a) (b) (c)

Figure 3: Fail Cases of HumanGaussian [20]

Facial Rendering Loss. To improve this issue, we introduced a facial rendering loss L_f, where in each regular iteration, there's a certain probability of moving the camera to a fixed position, additionally rendering the current facial image of the Gaussians. A pixel-wise color comparison is then performed with the facial photo extracted from the user portrait input (same size, background replaced to match the Gaussian renderer, and face moved to the center). The color difference is weighted more heavily for pixels near the center than for peripheral pixels, and after normalization, it is used as L_f. The overall loss $\mathcal{L}$ is expressed as:

$$\mathcal{L} = \lambda_1 L_{SDS1} + \lambda_2 L_{SDS2} + \lambda_3 L_f + \lambda_4 L_s \tag{6}$$

Here, L_{SDS1} and L_{SDS2} are the dual-branch SDS from the texture-structure joint pre-trained model (its gradient calculation formula refers to Equation 5); L_f is the previously mentioned facial loss; L_s is the sparsity loss.

When only the facial loss is added, it is observed that only the frontal view shows the user's facial features. After exporting the model, it is found that some Gaussians created a layer of artefacts in front of the face, with a certain distance from the character model, potentially being eliminated during the pruning phase. Thus, we also include the distance from the Gaussians to the original SMPL-X human model along the line of sight in the facial loss calculation.

3.4 Video-to-Animation Motion Transfer

Even though the entire framework is trained using a single pose, the generated avatar mesh is capable of animating unknown poses. By inputting a motion video where a character remains centered, each frame is processed separately using OpenPose and SMPLify-X to obtain a sequence of SMPL body pose parameters. These parameters are then transformed in a zero-shot manner to create smoothly transitioning, transferable animations.

4 EXPERIMENTS

4.1 Implementation Details

3D Gaussian Splatting Setups. Prior to the training process, 100,000 points are uniformly sampled on the surface of the exported SMPL-X mesh and initialized as 3D Gaussians. Each Gaussian is initially assigned an opacity of 0.1. The color is encoded using Spherical Harmonics (SH) coefficients [6, 13] of degree 0, as delineated in reference. The optimization of the 3D Gaussian Spheres (3DGS) is conducted using the Adam optimizer [14], characterized by beta values of [0.9, 0.99]. The learning rates are set at $5e-5$, $1e-3$, $1e-2$, $1.25e-2$, and $1e-2$ for the Gaussion parameters of the center position μ, scaling factor s, rotation quaternion q, color c, and opacity α, respectively. During the prune-only phase [20], the threshold for the scaling factor is exclusively set to 0.8.

Training Framework Setups. The overall training framework is an enhanced implementation based on PyTorch [25], ThreeStudio [7], and GaussianDreamer [40], building upon the HumanGaussian [20] pipeline. We use the prompt processor from *Stable Diffusion 2.0 base* [31], integrating annealing with negative prompt guidance [20]. The camera settings include a distance range of [1.5, 2.0], a field of view (fovy) range of [40°, 70°], and an elevation range of [-30°, 30°]. The training process spans a total of 3600 iterations, The dual-branch SDS loss weights from texture-structure joint model and the sparsity loss weights λ_1, λ_2, λ_3 are set as 0.5, 0.5 and 1. Between 300 and 2100 iterations, augmentation and pruning of Gaussians occur every 300 epochs. Between 2400 and 3300 iterations, a prune-only phase is executed every 300 epochs. During the 1200 to 3600

(a) Prompt: An Asian cute football girl with curly hair, female

(b) Prompt: A Chinese male wearing Tai Chi attire practicing Tai Chi, male

Figure 4: Experiment Result of Avatar Generation

iterations, we zoom into the head region with a camera distance range of [0.4, 0.6], at a 25% probability, to enhance facial quality with loss weights λ_1, λ_2, λ_3, λ_4 in Equation 6 being respectively adjusted to 0.5, 0.5, 1, and 1. The extraction of user facial images is currently done manually, with plans to automate this process using scripts in future developments. When trained with a batch size of 6 and a resolution of 1024×1024, the process requires approximately two hours on a single NVIDIA V100 (32GB) GPU.

Animation Migrator Setups. For the input motion videos, they are segmented at approximately 33 frames per second (meaning a frame is captured from the video every 0.03 seconds). After compressing the images, they are processed separately through OpenPose and the SMPLify-X method to obtain a sequence of SMPL-X poses. To simplify the process, we use only the parameters shared with SMPL, utilizing 21 body pose parameters and discarding additional SMPL-X parameters like facial expressions, hand gestures, and jaw movements. The animations rendered based on these pose sequences are output as mp4 videos at a frame rate of 30 frames per second. For an 8-second vertical motion video (with a total of 253 frame images split, and the image size compressed to 320×568), this entire process takes approximately 3 hours on a single NVIDIA GTX 1070 GPU.

4.2 Human Avatar Generation

We have tested our framework with various inputs. Figure 4 demonstrates two typical use cases: the first portrait input is derived from a piece generated by Stable Diffusion [31], with the 2D image prompt being "a full-body photo of a young Chinese girl," and the second portrait input is a real photograph from a classmate. The effect prompt inputs and the resulting outputs are shown in Figure 4(a) and (b), respectively. As can be observed, our framework effectively handles portraits in different styles, including those containing real photographs, across various effect prompts. The generated virtual avatars exhibit precise body structure and facial textures, realistic appearance, and align well with the expected outcomes of the effect prompts.

(a) (b) (c) (d)

Figure 5: Comparison of Facial Effects

Analysis of Facial Loss's effects. As mentioned in Section 3.3, one of the primary objectives of designing the facial rendering loss was to address the issues of facial fragmentation or asymmetry in existing text-to-3D models. Therefore, we conducted comparative experiments on facial modeling effects, as shown in Figure 5. subfigure (a) displays the facial photo input by the user, (b) shows the results from the existing model [20], (c) presents the effects of introducing facial rendering loss with the same seed initialization in SDS, and (d) depicts the effects of increasing the facial loss by 100 times (with only trained after 1200 epochs). Observation reveals that (c), compared to (b), effectively mitigates the problem of facial fragmentation and reduces incorrect effects in the neck region. The face shape and hairstyle become wider, similar to the user's facial input, enhancing naturalness; in group (d), by amplifying the loss, the generated avatar's face rapidly converges to the user's facial

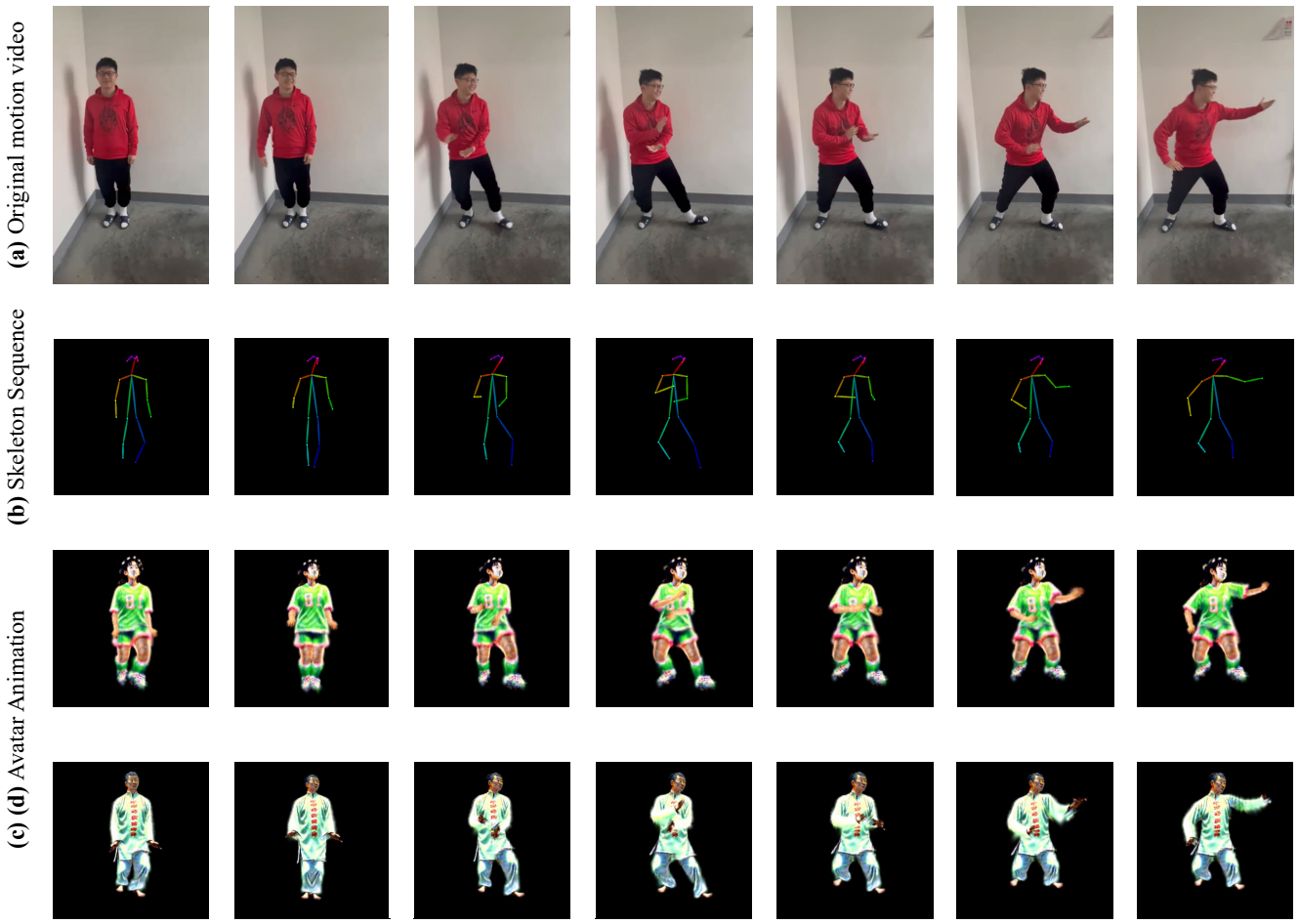

Figure 6: Experiment Result of Video-to-Animation Motion Transfer

input. Thus, we can discern that the introduction of facial loss plays a significant role in enhancing the naturalness of the image's face. Moreover, by adjusting the weight of the facial loss, we can control whether the model learns only the face shape from the user's portrait or all textures.

4.3 Avatar Animation

We utilized previously experimented subjects - the avatar of a football player (Person 1) and a Tai Chi practitioner (Person 2) - to demonstrate the animation effects of Tai Chi movements. The Tai Chi movement video was shot with an iPhone 14 using a mobile perspective. This video lasts approximately 8 seconds (comprising 253 frames of segmented image inputs, with the image size compressed to 320*568). The specific animation effects are shown in Figure 6. It is observable that the input Tai Chi movement video successfully drives the posture animation of the virtual avatar, resulting in smooth and fluid effects.

5 DISCUSSION

Limitations. Our method, capable of creating virtual avatars from combined image and text inputs, does have certain limitations: 1) The guiding model still faces performance bottlenecks, particularly in detailing hands, feet, and facial consistency. We have observed issues with effects in some scenarios. 2) The current model requires substantial computational resources, particularly demanding high GPU memory, which may not be feasible for all users.3) The added facial loss technique, while generally effective for face swapping, has inconsistencies in skin tone. Future work could explore integrating a controlnet for img2img tasks into SDS to address this.

Ethical Considerations. While our method is capable of generating realistic 3D human avatars, there exists a potential for misuse in harmful contexts. We, therefore, emphasize our commitment to ethical practices and strongly advocate for the responsible application of this technology. Our belief is that, when used appropriately, this technique can significantly contribute to advancements in both the research community and various industrial sectors.

6 CONCLUSION

In this paper, we propose a framework, **AAGaussian**, designed to create virtual human avatars with intricate structures and realistic appearances. We first use SMPL-X to receive user portrait inputs for Gaussian initialization, followed by proposing a facial rendering loss, which, in conjunction with SDS, guides the generation of the virtual avatar, ameliorating the issue of facial fragmentation in existing models. Experiments demonstrate that our framework can produce efficient and high-quality results for various inputs, sculpting realistic and diverse virtual avatars for users. This framework also includes an animation generator that can add smooth animations to the generated virtual avatars based on motion video inputs.

ACKNOWLEDGMENTS

This work was conducted as a group assignment for the course CS3310 Computer Graphics at Shanghai Jiao Tong University. The computations in this work were run on the π 2.0 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

Special thanks are also extended to Zhiye Wang for his contributions in the testing data capture and animation processing scripts.

REFERENCES

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5461–5470.

[2] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. 2007. Multi-scale capture of facial geometry and motion. *ACM transactions on graphics (TOG)* 26, 3 (2007), 33–es.

[3] Bowen Cai, Jinchi Huang, Rongfei Jia, Chengfei Lv, and Huan Fu. 2023. NeuDA: Neural Deformable Anchor for High-Fidelity Implicit Surface Reconstruction. *arXiv preprint arXiv:2303.02375* (2023).

[4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. arXiv:2303.13873 [cs.CV]

[5] Zilong Chen, Feng Wang, and Huaping Liu. 2023. Text-to-3D using Gaussian Splatting. arXiv:2309.16585 [cs.CV]

[6] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5501–5510.

[7] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.

[8] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022).

[9] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20374–20384.

[10] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. arXiv:2112.01455 [cs.CV]

[11] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*. 3334–3342.

[12] Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. arXiv:2305.02463 [cs.CV]

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4, Article 139 (jul 2023), 14 pages. https://doi.org/10.1145/3592433

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. 2023. DreamHuman: Animatable 3D Avatars from Text. arXiv:2306.09329 [cs.CV]

[16] J. P. Lewis, Matt Cordner, and Nickson Fong. 2000. Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-Driven Deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 165–172. https://doi.org/10.1145/344779.344862

[17] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

[18] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 300–309.

[19] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)* 40, 6 (2021), 1–16.

[20] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. 2023. HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting. arXiv:2311.17061 [cs.CV]

[21] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. arXiv:2309.03453 [cs.CV]

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (oct 2015), 16 pages. https://doi.org/10.1145/2816795.2818013

[23] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (dec 2021), 99–106. https://doi.org/10.1145/3503250

[24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (jul 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[27] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.

[28] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9054–9063.

[29] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. arXiv:2209.14988 [cs.CV]

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487 [cs.CV]

[33] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation. arXiv:2110.02624 [cs.CV]

[34] Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2006. Photo Tourism: Exploring Photo Collections in 3D. *ACM Trans. Graph.* 25, 3 (jul 2006), 835–846. https://doi.org/10.1145/1141911.1141964

[35] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. arXiv:2309.16653 [cs.CV]

[36] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS* (2021).

[37] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. arXiv:2305.16213 [cs.LG]

[38] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. arXiv:2310.08528 [cs.CV]

[39] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2023. 4k4d: Real-time 4d view synthesis at 4k resolution. *arXiv preprint arXiv:2310.11448* (2023).

[40] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. arXiv:2310.08529 [cs.CV]

[41] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.

[42] Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4574–4584.