Realistic 3D Generation from a Single 2D Image using Gaussian Splatting and Diffusion

Syed Muhammad Uzair Ahmed, Aizaz Usman, Aun Ali Kazi, Abbas Ghouri, Zayam Zia Malik
FAST NUCES, Karachi, Pakistan
uzair2005.bz@gmail.com, Aghazohaib14@gmail.com

***Abstract*—The generation of realistic 3D models from a single 2D image has long been a significant challenge in computer vision. Traditional methods often rely on multi-view datasets or high-resolution data, which limits their practical applications. In this paper, we present a novel approach that leverages 3D Gaussian Splatting and diffusion-based denoising to reconstruct 3D shapes progressively from 2D images. This technique eliminates the dependency on multiple views and high-resolution data, offering a more accessible solution. Additionally, we explore the trade-offs between using Variational Autoencoders (VAEs) and StyleGANs for generating initial 2D features. Our experiments show that while StyleGAN offers higher realism, it is computationally more complex. The proposed method effectively reconstructs 3D models from a single 2D image, demonstrating significant improvements in visual quality and accuracy.**

***Index Terms*—3D Reconstruction, Gaussian Splatting, Diffusion, Single Image, Generative Models, VAE, StyleGAN**

I. Introduction

The task of generating 3D models from 2D images has been a challenging area of research for decades. Traditionally, 3D reconstruction required multi-view images, typically captured in controlled studio settings, or high-resolution datasets to create accurate 3D models. These approaches, while effective, suffer from limitations in terms of accessibility, computational complexity, and the need for multiple viewpoints.

Recent advancements in generative modeling and neural rendering techniques have led to more feasible solutions that bypass these limitations. In particular, Gaussian splatting and diffusion models offer new possibilities for realistic 3D reconstruction from a single 2D image. This paper introduces a method that uses Gaussian splatting, a technique that represents 3D scenes as a collection of particles (3D Gaussians), combined with a diffusion-based denoising process, to generate high-quality 3D shapes from 2D inputs. The ability to generate realistic 3D models from a single 2D image opens up new possibilities in various applications such as virtual reality, augmented reality, and 3D content creation.

II. Methodology

A. 3D Gaussian Splatting

The foundation of our approach lies in the concept of 3D Gaussian splatting, which represents a 3D scene using discrete geometric primitives known as 3D Gaussians. These particles are used to form the final 3D structure by projecting and blending them in image space through a process called splatting. Each Gaussian particle is characterized by its location, orientation, and scale. The projection of these particles in image space is carefully managed to ensure that the resulting 3D representation is consistent with the input 2D image.

In the initial stages of the process, random noise is introduced into the particles, and through iterative refinement, the particles are adjusted to better represent the desired 3D structure. The key advantage of using 3D Gaussian splatting is that it allows for more flexibility and adaptability compared to traditional polygonal meshes, especially when dealing with complex surfaces and varying densities.

B. Progressive Denoising with Diffusion

Once the initial Gaussian splats are generated, a diffusion-based denoising process is employed to progressively refine the 3D structure. This step involves iteratively reducing noise while aligning the 3D Gaussians with the provided 2D image projections. The process ensures that the 3D structure not only becomes more refined but also maintains visual coherence with the input 2D view.

Through the diffusion process, the model gradually learns to preserve important details and smooth out inconsistencies in the 3D structure. The ability to iteratively refine the model allows for high-quality, realistic reconstructions even from low-resolution 2D images.

C. FLAME Mesh Integration

To further improve the realism and detail of the generated 3D models, the refined 3D Gaussians are integrated into a mesh structure. In particular, we use the FLAME (Faces Learned with an Active Model of Expression) mesh, a parametric model that provides a flexible and detailed representation of human faces. By attaching the 3D Gaussians to the triangles of the FLAME mesh, we can animate the model, capturing facial expressions and subtle deformations with high fidelity.

The integration of the FLAME mesh allows for more precise control over the generated 3D model and helps preserve the facial identity and expression details throughout the rendering process.

D. VAE vs. StyleGAN for 2D Feature Generation

The initial 2D features required for the reconstruction process can be generated using either Variational Autoencoders (VAEs) or StyleGANs. Each approach has its advantages and trade-offs. VAEs are simpler and less computationally

expensive, making them a suitable option for lightweight applications. However, VAEs often produce smoother and less realistic outputs due to their tendency to blur fine details.

On the other hand, StyleGANs are capable of generating highly realistic and detailed 2D features, but they require significantly more computational resources and training time. In this study, we compare the performance of both approaches and demonstrate that while StyleGAN offers superior realism, VAEs are more efficient in terms of computational requirements.

III. Results

A. Qualitative Results

Our method produces high-quality 3D reconstructions even from low-resolution 2D inputs. The iterative denoising process ensures that the 3D structure aligns well with the target view, maintaining geometric consistency while adding fine details. The FLAME mesh integration further enhances the realism of the reconstructed avatars, allowing for accurate facial expressions and identity preservation across different viewpoints.

B. Quantitative Results

We evaluated our method on a set of benchmark datasets, comparing our results to previous state-of-the-art techniques in terms of metrics like LPIPS, PSNR, and SSIM. Our approach consistently outperforms existing methods in novel view synthesis and expression synthesis, achieving higher fidelity reconstructions and more consistent identity preservation across novel perspectives.

C. Comparison with Existing Techniques

Compared to methods like NeRF and other Gaussian splatting-based approaches, our method significantly reduces computational time and increases rendering speed, allowing for real-time performance without sacrificing quality. By eliminating the need for multi-view datasets and high-resolution images, our approach provides a more accessible and efficient solution for 3D model generation.

IV. Discussion and Future Work

While our method demonstrates promising results in reconstructing realistic 3D avatars from monocular images, there are several avenues for future work. One area for improvement is the optimization of the diffusion process to further reduce computational complexity and accelerate the denoising steps. Additionally, incorporating more advanced generative models, such as GANs, could further enhance the realism and detail of the generated 3D models.

Another potential avenue is the integration of additional information sources, such as depth maps or multi-modal inputs, to improve the accuracy and robustness of the reconstruction process. Furthermore, the development of real-time 3D reconstruction systems for virtual reality and augmented reality applications would make this technology more practical for interactive and immersive environments.

V. Conclusion

We have presented a novel approach to generating realistic 3D models from a single 2D image using Gaussian splatting and diffusion-based denoising. Our method removes the dependency on multi-view and high-resolution data, making 3D generation more accessible and efficient. Through experiments and comparisons with existing techniques, we have demonstrated the effectiveness of our approach in producing high-fidelity 3D reconstructions that maintain consistency across different viewpoints and facial expressions.

Our method opens up new possibilities in 3D content creation, especially in applications such as virtual reality, gaming, and interactive media, where realistic and customizable avatars are crucial.

Acknowledgements

We would like to thank the authors of [1] for their foundational work on Gaussian splatting, which served as the inspiration for this study. Additionally, we acknowledge the contributions of the FAST NUCES team for their continuous support throughout this project.

References

[1] J. Gao, J. Chen, and X. Zhang, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *CVPR*, 2023.

[2] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *CVPR*, 2019.