CoGSD: Fast Consistency Generation Based on 3D Gaussian Splatting

Zhenghui Sun[1] , Xianglong Li[2] and Shuyuan Chang[3]

[1] Harbin Engineering University, School of Computer Science and Technology, 145 Nantong Street, Nangang District, Harbin, Heilongjiang Province, China
tangchaoyuezi@outlook.com

[2] Ocean University of China, School of Computer Science, 238 Songling Road, Laoshan District, Qingdao City, Shandong Province, China
lxl2015ly@163.com

[3] Nanjing University of Science and Technology, School of Intelligent Manufacturing, 8 Fuxing Road, Jiangyin City, Wuxi City, Jiangsu Province, China
shuyuanchang@njust.cn

Abstract. This research presents a method called CoGSD (Consistent Gaussian Splatting Dreamer) for rapid construction of 3D models with multi-view consistency through 3D Gaussian Splatting. However, the traditional approach of 3D Gaussian Splatting encounters challenges in effectively constructing 3D assets due to the absence of stable ground truth. Furthermore, the inherent expansion characteristics of 3D Gaussian Splatting result in abnormal expansion of saturated Gaussian points and inconsistencies across multiple views. Additionally, the lack of reliable ground truth exacerbates these multi-view inconsistency issues. To solve this problem, we use a pre-trained consistent diffusion model to generate consistent viewpoints. In our framework, instead of generating diffusion with a single a priori perspective, the 2D image generation method of SDS uses a controlnet-tuned pre-trained model to generate 2D images with coherent viewpoints, resulting in high-quality 3D model generation. The method in this paper provides an effective solution for 3D modeling and is expected to be widely used in the field of 3D modeling and visual effects.

Keywords: 3D Model Generation, Multi-view Consistent Image Generation, 3D Gaussian Splatting, Score Distillation Sampling, Diffusion Model

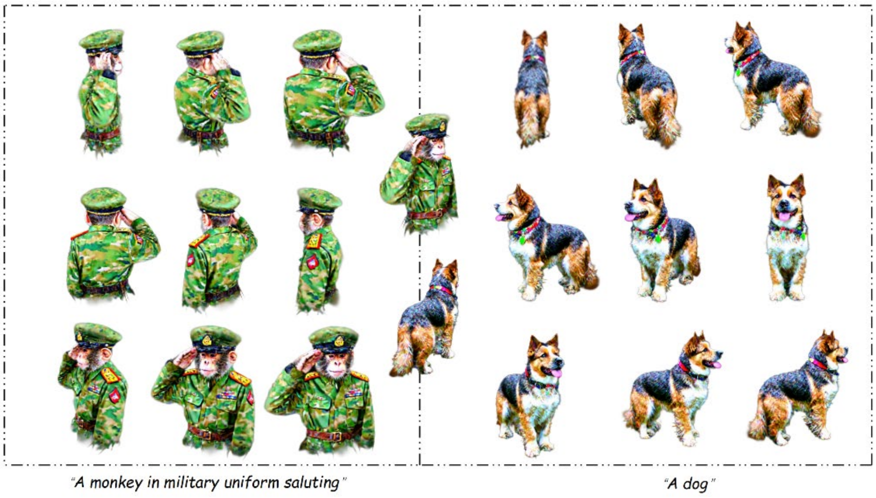

Fig. 1. Our CoGSD framework achieves high fidelity and excellent consistency across multiple viewpoints without showing high saturation of colors and shapes.

1 Introduction

The realm of 3D digital content creation is undergoing a rapid and significant transformation, expanding its influence across various sectors like digital gaming, advertising, film production, and the metaverse. This evolution is driven by advanced technologies such as image-to-3D and text-to-3D conversions, which are revolutionizing how professionals and amateurs alike create and interact with 3D assets. The impact of these technologies is far-reaching, significantly enhancing user engagement and immersion in digital experiences, especially in the film and gaming industries. However, the process of converting 2D content into 3D models presents significant challenges, primarily due to the limitations in existing methods.

The current method has two main paths: text is converted into a 3D model through 2D images [1-4] and an end-to-end method that directly generates a 3D model based on text description [5,6].

The existing text to 3D technology routes are not all end-to-end methods[2,3,7,8] , most of them first convert text input into images [9-11] , and then proceed to convert these images into 3D models.

Methods for image generating 3D usually use the input individual images as base view for multi-view image generation[12-17] , and then 3D reconstruction using multiple images with different viewpoints [18,19] Despite these advancements, current methods still struggle to meet the demands for high-quality 3D reconstruction, particularly in terms of maintaining image consistency during multi-view image generation. The reliance on intermediate steps often hampers the fidelity and consistency of the final 3D model. Consequently, this results in the emergence of two key issues: the multi-view inconsistency problem, which pertains to maintaining consistency in object shapes across different perspectives, and the Janus problem, wherein AI-generated 3D objects may exhibit multiple heads or faces [2,20].

The reason is that the 2D diffusion model does not have all the three-dimensional spatial information, so multi-view information cannot be integrated more effectively. A lot of work has been devoted to solving this problem in recent years. Zero123 [12] uses synthetic datasets to learn relative camera viewpoint control, which allows the generation of new images of the same object under specified camera transformations. Some models based on zero123, such as One-2-3-45 [14], SyncDreamer [16], and Consistent 123 [21] are also based on one-time perfect image group generation model. However, this will result in the consistent diffusion model being unable to be iteratively optimized to maximize its performance.

To tackle these challenges, we introduce a new method called CoGSD (Consistent Gaussian Splatting Dreamer). Our approach aims to maintain consistency throughout the iterative optimization process, ensuring that the final 3D renderings closely match the intended viewpoints.The key concept behind CoGSD is to utilize pre-trained models to generate images with consistent viewing angles. These images are then used to create coherent 3D contours and produce high-fidelity 3D models.

CoGSD uses the Shap-E prior to construct the underlying point cloud structure, and introduces a pre-trained consistency diffusion model during the iteration process, by which the consistency of the same viewpoint before and after the iteration is maintained while forming the consistency of multiple views.

On this basis, CoGSD uses the 3D a priori knowledge as existing knowledge to generate 3D high-fidelity images, using a 2D diffusion model to deeply optimize the Gaussian parameters of the 3D model. Score Distillation Sampling (SDS) method is used as a loss function for fine-tuning the parameters to ensure that the generated novel views are consistent with the base viewpoints while maintaining high quality and fidelity.

Our experiments, conducted on RTX 3090 GPUs with stable-diffusion-2-1-base [22] and Shap-e [23] models, demonstrate the efficacy of our approach.

We used advanced image quality assessment metrics such as PSNR, SSIM, LPIPS, and FID for a comprehensive evaluation.

Controlled by a priori constraints, the final results demonstrate the validity of our method.The CLIP metric illustrates the degree of fit between the 3D model and the cues. Also, we show perspective images of the model to demonstrate the results of this study in generating consistent models.

The results show that CoGSD significantly outperforms existing methods, establishing new benchmarks in the field of 3D content creation.

The contributions of this study are as follows:

- We implemented a new framework that recycles consistent diffusion models (such as zero123) during the iterative generation process and uses ControlNet [24] to maintain stable front and rear control during each iteration step. This can maximize the potential for consistent generation.
- The reconstruction method of Gaussian Splatting is highly sensitive to ground truth and can quickly generate, delete, rotate angles, and split changes of points.

- This work also focuses on exploring the initial point density information required for Gaussian Splatting to generate 3D assets, which will have a reference impact on subsequent work.

2 Related Works

2.1 3D Reconstruction

Since the introduction of Neural Radiation Field (NeRF) [25], it has been widely applied in the fields of novel view synthesis, light field estimation, and three-dimensional reconstruction with impressive performance. DreamFusion [2] utilizes pre-trained image diffusion a priori with a specialized image space loss function to optimize a 3D model represented by a neural radiation field (NeRF). However, 2D image generation models lack 3D spatial information. To alleviate this problem, studies such as Score Jacobian Chaining [26] use view cues and additional regularization strategies. However, 3D generation based SDS [2] needs to run at unusually high bootstrap scales, and ProlificDreamer [3] solves these challenges with variational distillation (a general form of SDS.). Recently, 3D Gaussian Splatting [27] has been proposed as an alternative 3D representation of NeRF, which has shown impressive quality and speed in 3D reconstruction [28].

2.2 3D Generative Models

Some projects leverage a few existing large 3D datasets for end-to-end generation, as seen in [29-31]. However, the challenge lies in the absence of a standardized representation of the 3D data. DreamFusion [2] achieves groundbreaking text-to-3D generation using pre-trained image diffusion priors. Nonetheless, this method of reconstructing nerf is notably slow. To address this issue, [4] devised a two-stage optimization framework, acquiring a rough model before conducting specific generative reconstruction efforts.Moreover, the straightforward approach of guiding them through potential fraction distillation necessitates encoding the potential space at each guidance step. Latent-NeRF [32] proposes incorporating NeRF into the potential space to yield Latent-NeRF. With the introduction and remarkable application of 3D Gaussian Splatting, GaussianDreamer [8] uses 3D Gaussian Splatting [27] and takes the step of densifying the point cloud to obtain a 3D model. DreamGaussian～\cite[33] and GaussianDreamer have the same basic reconstruction method, but the kernel of the mechanism is completely different, through Generative Gaussian Splatting [27], Efficient Mesh Extraction [34], and UV-space Texture Refinement [35] in three steps. Our work also uses 3D Gaussian Splatting for reconstruction.

3D consistency generation methods have also been studied in recent years, and consistency generation methods not only include the above mentioned Shap-e and so on through the end-to-end approach to the generation of methods, recently especially hot is [12] proposed Zero-123, which uses a synthetic dataset to learn the control of the relative camera point of view, which allows for the generation of a new image of the same object under the specified camera transformations. Whereas Magic123 [1] uses a single prior, text inversion, and monocular depth regularization method to accomplish a two-stage strategy to complete a coherent generative model with good results; One-2-3-45 [14] employs diffusion pre-training of Zero-123 [12] to generate multiview images for the input view, and then boosts them into 3D space, accomplishing a relatively good coherent model construction at a small cost; Recently SyncDreamer [16] likewise uses Zero-123 [12] model weights, and they propose a simultaneous multiview diffusion model that models a joint probability distribution of multiview images such that multiview-consistent images are generated in a single inverse process that addresses the geometric and color consistency of the generated images.

3 Preliminaries

In this section, we will briefly review the representation and rendering process of 3D-GS [27] in 3.1 and elaborate on the diffusion model in 3.2.

3.1 3D Gaussian Splatting

3D Gaussian Splatting [27] introduces an explicit three-dimensional scene representation using a point cloud format, employing Gaussian functions for scene modeling. Each gaussian particle is characterized by a covariance matrix Σ and a central point X,, the latter being the Gaussian's mean. The covariance matrix Σ can be decomposed into a scale matrix S and a rotation matrix R for differential optimizations:

$$\Sigma = RSS^T R^T \tag{1}$$

Furthermore, the Gaussian representation is defined as:

$$G(X) = e^{-\frac{1}{2}x^T \Sigma^{-1} x} \tag{2}$$

For rendering novel views, Gaussians on the camera plane utilize splatting techniques [36]. The transformed covariance matrix Σ' in camera coordinates, calculated using the Jacobian affine view transform W and the projection transform J, is given by:

$$\Sigma' = JW\Sigma W^T J^T \tag{3}$$

Each Gaussian particle has the following attributes: position $X \in R^3$, color defined by the spherical harmonic coefficient $C \in R^k$ (where k represents degrees of freedom), opacity $\alpha \in R$, rotation factor $r \in R^4$, and scale factor $s \in R^3$. For each pixel, all Gaussian colors and opacities are computed using Equation 2. The blending formula for N-ordered points overlapping pixels is as follows:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_i) \tag{4}$$

Here, c_i and α_i denote the density and color of the point, calculated by multiplying the Gaussian with covariance Σ by the opacity of each point and the SH color coefficient, which can be optimized.

3.2 Score Distillation Sampling

DreamFusion [2], one of the most representative approaches to upgrading 2D diffusion models to 3D, proposes to optimize the 3D representation by using a pre-trained 2D diffusion model ϕ using the Score Distillation Sampling (SDS) method. Specifically, the 3D representation MipNeRF [37] is used as the parameter θ to be optimized, and the rendering method is used as the parameter g to obtain the rendered image $\mathbf{x} = g(\theta)$. In order to make the rendered image $\mathbf{x}$ similar to the samples obtained from the diffusion model ϕ, Dreamfusion uses a score estimation function: $\hat{\epsilon}_\phi(\mathbf{z}_t; y, t)$, which is used to estimate the score of the rendered image $\mathbf{x}$, the text embedding y, and the noise embedded in the rendering method, given a noisy image z_t , the text embedding y, and the noise ϵ embedded in the rendered image $I_{R,T}$ the text embedding y and the noise level t predicts the sampling noise $\hat{\epsilon}_\phi$. The score estimation function provides the direction used to update the parameter θ. The formula for computing the gradient is as follows.

$$\nabla_\theta \mathcal{L}_{\mathrm{SDS}}(\phi, \mathbf{x} = g(\theta)) E_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon\right)\frac{\partial \mathbf{x}}{\partial \theta}\right] \tag{5}$$

where ϵ represents Gaussian noise, and $w(t)$ is the weighting function.

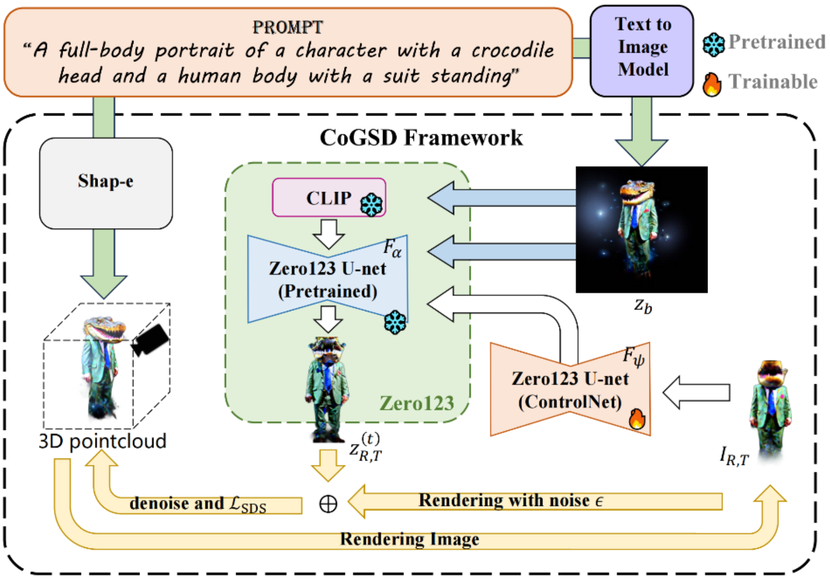
PROMPT
“A full-body portrait of a character with a crocodile head and a human body with a suit standing”
Text to Image Model
Pretrained
Trainable
CoGSD Framework
Shap-e
CLIP
Zero123 U-net (Pretrained)
F_α
Zero123
3D pointcloud
z_b
F_ψ
Zero123 U-net (ControlNet)
$z_{R,T}^{(t)}$
$I_{R,T}$
denoise and $\mathcal{L}_{\mathrm{SDS}}$
⊕
Rendering with noise ϵ
Rendering Image

Fig. 2. Our framework.

4 Method

In this section, we introduce CoGSD (Consistent Gaussian Splatting Dreamer) in detail, as shown in the **Fig. 2**. The basic viewpoint is derived from the prompt, with ControlNet and Zero123 aiming to align the remaining viewpoint distribution as closely as possible to this basic viewpoint. Meanwhile, the initial point cloud of the 3D model is obtained for subsequent processing. An image from a specific viewpoint, denoted as $I_{R,T}$, is then rendered from this point cloud to guide Zero123 in diffusing the generation of a novel view.

4.1 Novel View Generation

When addressing the issue of high variability in 3D Gaussian Splatting, we encounter the challenge of producing highly saturated Gaussian particles. To tackle this problem, we've implemented an innovative strategy that incorporates a shape prior. This prior ensures that the model possesses a fundamental point cloud structure at the initial stage, thereby laying a solid foundation for subsequent generative reconstruction.

Further, in order to generate a novel view image that is consistent with the base viewpoint, we adopt an approach that combines the target camera information with the base viewpoint data. Specifically, we first introduce an a priori model rendering image $I_{R,T}$ of the target viewpoint into a ControlNet F_{ψ}. The ControlNet is a zero123-oriented diffusion model, which replicates the weights of that model such that F_{ψ} obtains trainable weights based on $I_{R,T}$.

Next, we introduce the base view z_b and cue y into this framework to generate novel view images. The generation strategy for the novel view can be described by the following equation: $F_{\alpha} \oplus F_{\psi}$, where $\oplus$ denotes processing using the Zero-1-to-3 model adjusted by ControlNet. This approach not only retains the core features of the underlying perspective but also allows us to finely control the generation of image details by adjusting model weights.

By combining the Shap-E prior, the target viewpoint prior, and the ControlNet-based weight adjustment strategy, we achieve effective control of highly saturated Gaussian particles , as well as the ability to generate a novel view image that is highly consistent with the original base view.

$$z_{R,T}^{(t+1)} = F_{\beta} \oplus F_{\psi}\left(I_{R,T}, z_{R,T}^{(t)}, z_b, y\right) \tag{6}$$

4.2 Optimization

In order to significantly improve the detail richness and overall quality of 3D assets, we adopt the point cloud prior to the 3D diffusion model (Shape-E [23]) when initializing the 3D Gaussian parameter θ. After this step was completed, we further used a two-dimensional diffusion model to optimize these Gaussian parameters deeply. We use Score Distillation Sampling (SDS) as the loss function to fine-tune the three-dimensional Gaussian parameters.

In terms of specific implementation strategy, we first generate the rendered image $\mathrm{x} = g(\theta)$ through the 3D Gaussian Splatting method g. Here g refers to the photometric rendering technique we elaborated on in 3.1, which allows us to generate high-quality rendered images from three-dimensional Gaussian parameters. Next, in order to achieve parameter optimization, we calculate the SDS loss, which is specially designed to guide the two-dimensional diffusion model $F_{\alpha} \oplus F_{\psi}$ to update the Gaussian parameter θ gradient. The continuous iterative process aims to maximize the distribution of newly generated perspectives while maintaining consistency and aligning as closely as possible with the distribution of basic perspectives, thereby producing higher-quality new perspective images. SDS leverages the direct correlation between Gaussian parameters and the target image, and also includes comprehensive improvements to the generated images.

Table 1. Comparison with Zero123, Syncdreamer, Shap-e, Gaussiandreamer baseline models

	Zero123	SyncDream	Shape-E	Gaussian-Dreamer	CoGSD(ours)
PSNR ↑	17.57	19.83	12.66	20.23	**25.68**
SSIM ↑	0.868	0.883	0.701	0.891	**0.904**
LPIPS ↓	0.093	0.085	0.152	0.084	**0.079**
FID ↓	0.029	0.026	0.054	0.028	**0.024**

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) E_{t,\epsilon}\left[w(t)\big(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon\big)\frac{\partial \mathbf{x}}{\partial \theta}\right] \quad (7)$$

Among them, $\hat{\epsilon}_\phi(\mathbf{z}_t; y, t)$ is the image generated by diffusion model with noise. Its definition is as follows. We are The consistency perspective is introduced here:

$$\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) = I_{R,T} + \epsilon + z_{R,T}^{(t)} \quad (8)$$

After a series of short optimization cycles using the 2D diffusion model $F_\alpha \oplus F_\psi$, the resulting 3D instance not only maintains the 3D value provided by the 3D diffusion model F_{3D} consistency, and also achieves higher quality and fidelity.

5 Experiment

5.1 Compare with Baseline

In order to validate the multi-view consistency of our model, we performed evaluations at Google Scan Objects (GSO), which is a dataset containing a large number of 3D scanned objects. These objects were acquired through advanced scanning techniques that ensure high-quality and high-resolution 3D images. We verified the image similarity between a particular viewpoint (camera parameters randomized) of the generated 3D assets and the corresponding viewpoint ground truth of the GSO by acquiring one of the viewpoints (camera parameters randomized).

We conducted experiments on RTX 3090 GPUs, setting the number of training iteration steps to 1200. For the 2D diffusion model, we used stable-diffusion-2-1-base [22] with a guidance scale of 7.5, and for the 3D diffusion model Shap-E [23], with a guidance scale of 7.5, and learning rates of initial opacity and position of 10^{-2} and 10^{-5} respectively.

In order to provide a comprehensive quantitative assessment of the performance of the novel view synthesis technique, we employ four state-of-the-art image quality assessment metrics that cover multiple key dimensions of image similarity. Specifically, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [38], Learned

Perceptual Image Patch Similarity(LPIPS) [39], and Fréchet Inception Distance (FID) [40]. Together, these evaluation metrics form a comprehensive evaluation framework that not only examines the differences between the benchmark algorithms and the proposed method in this study in terms of traditional pixel-level accuracy, but also provides in-depth comparisons and analyses from the perspectives of structural preservation, perceptual similarity, and statistical feature distribution. The results are shown in the **Table 1.** Comparison with Zero123, Syncdreamer, Shap-e, Gaussiandreamer baseline models. Through this multi-dimensional evaluation approach, we are able to measure and present the performance of the novel view generation technique in various aspects more accurately, thus providing valuable references for further research and applications.

This article compares the effects with Dreamfusion [2], ProlificDreamer [3], zero123 [12], and SyncDreamer [16]. The results are shown in the **Fig. 3**.

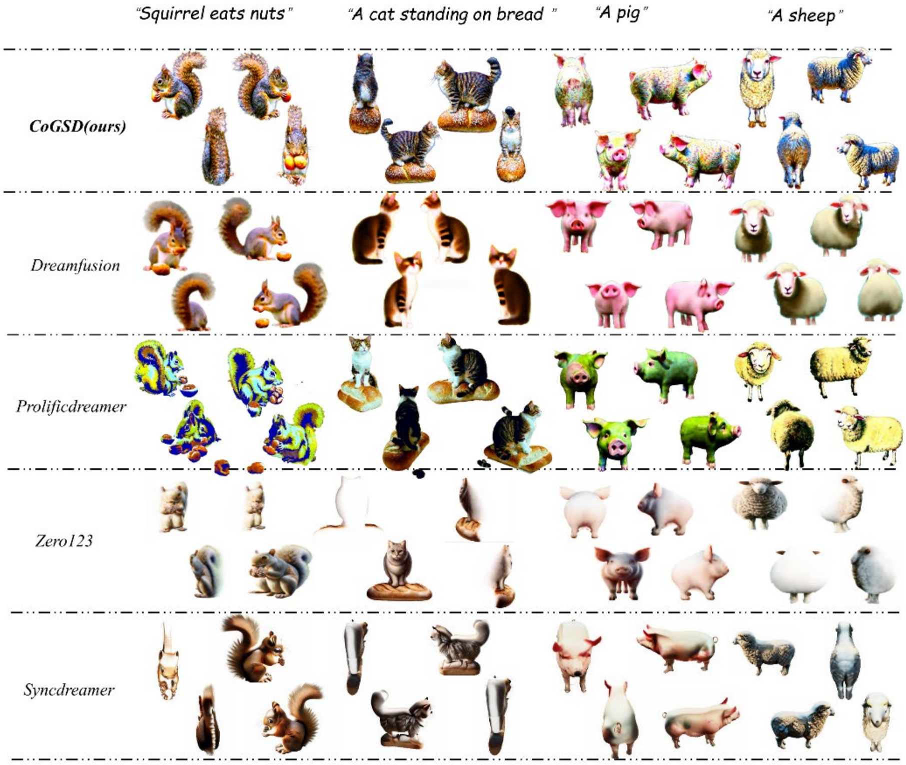

Fig. 3. Examples for exhibiting.

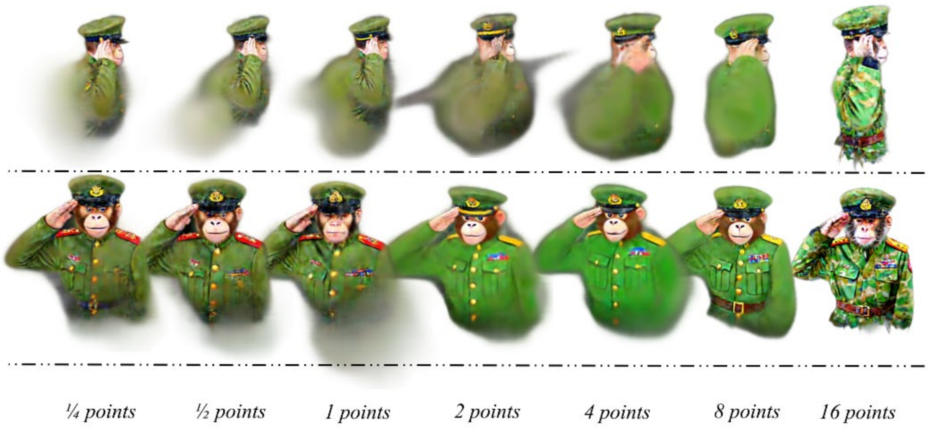

Fig. 4. This is the result of a test in the "A monkey in military uniform saluting" prompt, from a quarter of the number of points to 16 times the number of points in the original a priori model.

5.2 Ablation Study

The impact of the initial prior point cloud number on the results. Since Gaussian point clouds have rapid variability, in order to suppress the occurrence of this variability, GaussianDreamer introduced a densification method to achieve this purpose. However, there is no exact explanation of the appropriate point cloud density, which was carried out in our study. In order to experimentally explore the appropriate parameters for this densification, we performed experiments on intervals of $^{1}/_{4}$ to 16 times the number of point clouds. Using the GSO data set as the test object and PSNR as the indicator to explore the optimal densification effect, The results are shown in the **Fig. 5**. We have found that densification is best produced at around 16x, and that too much increase in the number of points can lead to slow computing and memory limitations that outweigh the benefits. This greatly reduces the anomalous generation of Gaussian point clouds around the subject.

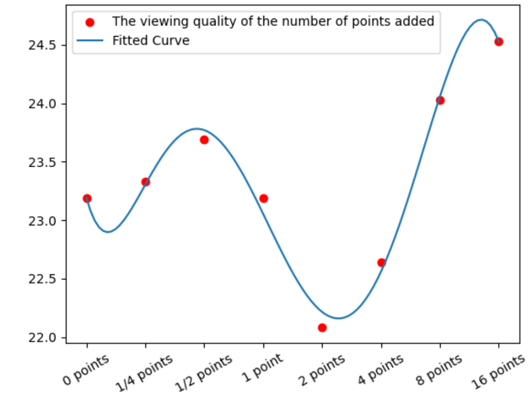

Fig. 5. Variation of the quality of the generated 3D model with the number of additions, the horizontal axis is the number of additions and the vertical axis is the PSNR metrics.

Analysis of opacity distribution of initial point cloud. In the 3D Gaussian representation, an accurate representation of the scene can be achieved by interleaving the optimization of the 3D Gaussian model and carefully adjusting the density parameters, especially the precise optimization of the anisotropic covariance. It also produces an effect on OPACITY, when the model finishes generating the 3D resource, the distribution of its OPACITY number is shown in **Fig. 6**. After calculating the numerical quantities, the mean is -2.186, the variance is 0.290, the statistic of the Shapiro-Wilk Test is 0.976, and the pvalue is 0.25. We cannot reject the hypothesis that this is a normal distribution. This result is consistent with our expectations for Gaussian Splatting technology. Due to the characteristics of its normal distribution, we can also perform some flexible transformations on the opacity distribution of the point cloud. Migrating it to other reconstruction methods may have surprising effects, which requires further work.

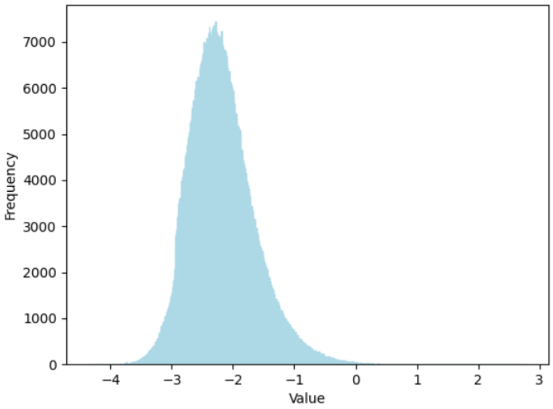

Fig. 6. Quantitative distribution of opacity values of point clouds

6 Conclusion

In this study, we propose a consistent 3D asset generation method. Unlike the consistency-based view one-time reconstruction 3D generation method, our method incorporates consistency information during the iterative optimization process to ensure that the 3D rendering results are close to the consistency information of the novel view. The framework not only remains efficient in generating consistent information, but also ensures relative control of the current viewpoint. In this paper, we also explore the impact of the point cloud prior to the particle opacity distribution, providing a reference metric for subsequent research on 3D asset generation using Gaussian particles. The experimental results we employed show that the method performs excellently in generating high-quality 3D models, and the evaluation metrics of the generated 3D models are substantially ahead of baseline.However, due to the limitations of the Zero-1-to-3 pretraining framework, there is still room for improvement in our method for generating finer 3D models. In the future, the method can be further enhanced and optimized by using more advanced diffusion models based on this study.

References

1. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.-Y., Skorokhodov, I., Wonka, P., Tulyakov, S., others: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)

2. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)

3. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. arXiv preprint arXiv:2305.16213 (2023)

4. Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., Lin, T.-Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 300-309. (Year)

5. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)

6. He, Z., Wang, T.: OpenLRM: Open-Source Large Reconstruction Models. \url{https://github.com/3DTopia/OpenLRM} (2023)

7. Guo, Y.-C., Liu, Y.-T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.-H., Zou, Z.-X., Wang, C., Cao, Y.-P., Zhang, S.-H.: threestudio: A unified framework for 3D content generation. \url{https://github.com/threestudio-project/threestudio} (2023)

8. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. arXiv preprint arXiv:2310.08529 (2023)

9. Lab, D.: DeepFloyd IF. (2023)

10.OpenAI: DALL·E 2. (2023)

11.OpenAI: DALL·E 3. (2023)

12.Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9298-9309. (Year)

13.Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)

14.Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. arXiv preprint arXiv:2311.07885 (2023)

15.Liu, M., Xu, C., Jin, H., Chen, L., Varma, M., Xu, Z., Su, H.: One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. In: Thirty-seventh Conference on Neural Information Processing Systems. (Year)

16.Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. arXiv preprint arXiv:2309.03453 (2023)

17.Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: MVDream: Multi-view Diffusion for 3D Generation. arXiv:2308.16512 (2023)

18.M\"uller, T., Evans, A., Schied, C., Keller, A.: Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans. Graph. 41, 102:101-102:115 (2022)

19.Tang, J.: Stable-dreamfusion: Text-to-3D with Stable-diffusion. (2022)

20.Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12663-12673. (Year)

21.Weng, H., Yang, T., Wang, J., Li, Y., Zhang, T., Chen, C.L., Zhang, L.: Consistent123: Improve consistency for one image to 3d object synthesis. arXiv preprint arXiv:2310.08092 (2023)

22.Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.o.: High-Resolution Image Synthesis With Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695. (Year)

23.Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)

24.Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836-3847. (Year)

25.Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65, 99-106 (2021)

26.Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. (2022)

27.Kerbl, B., Kopanas, G., Leimkhler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42, (2023)

28.Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023)

29.Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: LION: Latent point diffusion models for 3D shape generation. arXiv preprint arXiv:2210.06978 (2022)

30.Cheng, Y.-C., Lee, H.-Y., Tulyakov, S., Schwing, A.G., Gui, L.-Y.: Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4456-4465. (Year)

31.Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., others: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)

32.Wang, B., Dutt, N.S., Mitra, N.J.: ProteusNeRF: Fast Lightweight NeRF Editing using 3D-Aware Image Context. arXiv preprint arXiv:2310.09965 (2023)

33.Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)

34.Zhang, C., Chen, T.: Efficient feature extraction for 2D/3D objects in mesh representation. In: Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), pp. 935-938. (Year)

35.Chen, Z., Yin, K., Fidler, S.: Auv-net: Learning aligned uv maps for texture transfer and synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1465-1474. (Year)

36.Yifan, W., Serena, F., Wu, S., ztireli, C., Sorkine-Hornung, O.: Differentiable surface splatting for point-based geometry processing. ACM Transactions on Graphics (TOG) 38, 1-14 (2019)

37.Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5855-5864. (Year)

38.Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600-612 (2004)

39.Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586-595. (Year)

40.Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30, (2017)