Repaint123: Fast and High-quality One Image to 3D Generation with Progressive Controllable Repainting

Junwu Zhang[1⋆] Zhenyu Tang[1⋆] Yatian Pang[1,3] Xinhua Cheng[1]
Peng Jin[1] Yida Wei[4] Xing Zhou[5] Munan Ning[1,2] Li Yuan[1,2⋆⋆]
[1]Peking University [2]Pengcheng Laboratory [3]National University of Singapore
[4]Wuhan University [5]RabbitPre

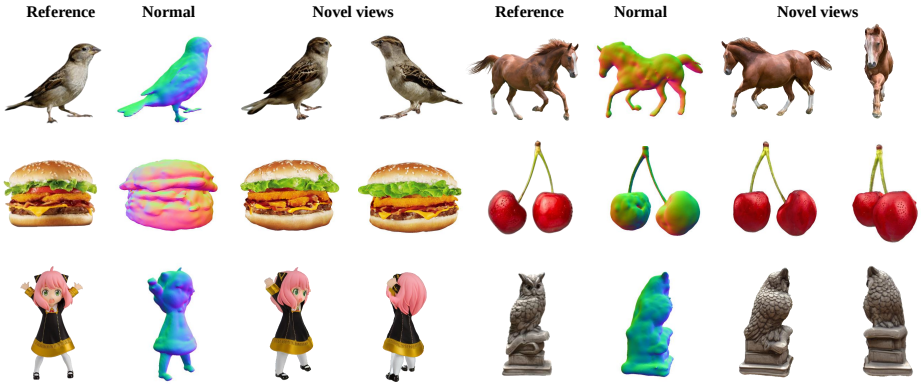

Fig. 1: Repaint123 generates high-quality 3D content with detailed texture from a single image in **2 minutes**. Repaint123 adopts 3D Gaussian Splatting in the coarse stage and then utilizes a progressively repainting strategy with the diffusion model for high-quality 3D content generation with efficiency.

Abstract. Recent image-to-3D methods achieve impressive results with plausible 3D geometry due to the development of diffusion models and optimization techniques. However, existing image-to-3D methods suffer from texture deficiencies in novel views, including multi-view inconsistency and quality degradation. To alleviate multi-view bias and enhance image quality in novel-view textures, we present **Repaint123**, a fast image-to-3D approach for creating high-quality 3D content with detailed textures. Repaint123 proposes a progressively repainting strategy to simultaneously enhance the consistency and quality of textures across different views, generating invisible regions according to visible textures, with the visibility map calculated by the depth alignment across views.

⋆ Equal contribution.
⋆⋆ Corresponding author.

Furthermore, multiple control techniques, including reference-driven information injection and coarse-based depth guidance, are introduced to alleviate the texture bias accumulated during the repainting process for improved consistency and quality. For novel-view texture refinement with short-term view consistency, our method progressively repaints novel-view images with adaptive strengths based on visibility, enhancing the balance of image quality and view consistency. To alleviate the accumulated bias as progressively repainting, we control the repainting process by depth-guided geometry and attention-driven reference-view textures. Extensive experiments demonstrate the superior ability of our method to create 3D content with consistent and detailed textures in 2 minutes.

Keywords: Image-to-3D· Texture Enhancement · Diffusion model

1 Introduction

Generating 3D content from one given reference image plays a key role at the intersection of computer vision and computer graphics [11, 17, 24, 25, 33, 35], which is desired by users for innovative applications across fields including robotics, virtual reality, and augmented reality. However, the image-to-3D task is quite challenging due to the generated 3D content is expected with reasonable geometry and consistent textures, while the single image input is insufficient in multi-view information. Recent studies [24, 28, 35, 50, 51] employ 2D diffusion models [14, 39] to guide 3D generation from the reference image with Score Distillation Sampling (SDS) [34]. While preliminary attempts [28, 51] leveraging view-independent diffusion models (e.g. Stable Diffusion [39]) faced challenges with over-saturated textures and multi-face geometry, more recent efforts [24, 26, 35, 46, 50] utilize view-conditioned diffusion models [24, 26, 46]. Although these methods utilizing view-conditioned diffusion models mitigate the multi-face problem and generate 3D content with plausible geometry, they still encounter challenges in generating high-quality textures. This limitation stems from training these models on small-scale synthetic 3D datasets, hindering their ability to produce detailed images from novel views.

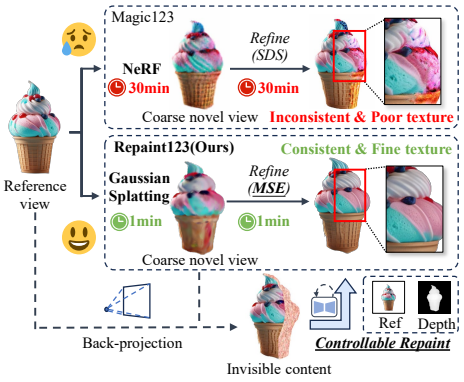

Fig. 2: Motivation. Current image-to-3D methods adopt SDS loss for generating, resulting in inconsistent and poor textures. Repaint123 proposes a progressively repainting strategy based on visible content to efficiently create invisible content with MSE loss, achieving consistent and high-quality textures.

To address textural deficiencies including inconsistency and low quality mentioned above, we propose a novel approach named Repaint123 for fast and high-quality image-to-3d generation by generating consistent and high-quality novel-view images with a controllable content repaint process. As shown in Figure 2, the core of Repaint123 is to progressively repaint the invisible content with a view-independent diffusion model for adjacent-view consistency, incorporating multiple controls of geometry and texture for enhanced quality and multi-view consistency. Specifically, our method adopts a two-stage optimization framework. In the coarse stage, we utilize 3D Gaussian Splatting [18] as the representation to efficiently obtain a coarse 3D model in 1 minute. In the fine stage, we progressively sample the camera views around and repaint invisible content with control techniques including using novel-view depth map as geometry prior, injecting reference-view attention features as texture prior [61] and employing reference-view image prompt as semantic prior. To further enhance image quality and preserve multi-view consistency, we propose a visibility-based adaptive strategy that refines the visible regions with different strengths. After achieving high-quality repainted images with multi-view consistency, we directly leverage simple MSE loss to efficiently refine 3D content textures instead of the time-consuming and over-smoothed SDS.

We conduct extensive experiments on multiple datasets and demonstrate that our method significantly advances image-to-3D generation, producing high-quality, multi-view consistent 3D content with detailed textures in approximately 2 minutes from scratch (as shown in Figure 1), outperforming current state-of-the-art NeRF-based and Gaussian-Splatting-based methods in both consistency and texture quality. Our contributions can be summarized as follows:

- We propose a progressively repainting strategy based on visibility for the image-to-3D task to generate 3D content with consistency and high quality by repainting invisible parts according to visible textures.
- We introduce multiple control techniques including reference-driven information injection and coarse-based depth guidance to alleviate the texture bias accumulated during the repainting process for further quality enhancement.
- Our experiments demonstrate Repaint123 efficiently generates high-quality 3D content with consistent and detailed textures in **2 minutes**.

2 Related Works

2.1 Diffusion Models for 3D Generation

The recent notable achievements in view-independent 2D diffusion models [14,39] have brought about exciting prospects for generating 3D objects. Pioneering studies [34, 53] have introduced the concept of distilling a 2D text-to-image generation model to generate 3D shapes. Subsequent works [1, 6, 7, 10, 16, 22, 28, 35, 37, 42–44, 51, 52, 55–57, 60, 61, 64] have adopted a similar per-object optimization approach, building upon these initial works. Nevertheless, the majority of these techniques consistently experience low efficiency and multi-face

issues. Unlike the previous study HiFi-123 [61], which employed similar inversion and attention injection techniques for image-to-3D generation, our approach diverges primarily in our proposed repainting strategy, as well as many variations in optimization objectives, image prompting, and 3D representations. Recently, some works [23, 25, 26, 48] extend view-independent diffusion model to view-conditioned diffusion model to generate multi-view images for reconstruction, while these methods usually suffer from low-quality textures as the multi-view diffusion models are trained on limited and synthesized data. To optimize both consistency and quality, we integrate a view-conditioned diffusion model for initial consistency in the coarse stage and a view-independent model to enhance texture quality in the fine stage.

2.2 Controllable Image Synthesis

A major challenge in the field of image generation is achieving controllability. Many works have been done recently to increase the controllability of images generated by diffusion models. ControlNet [62] and T2I-Adapter [31] attempt to control the creation of images by utilizing data from different modalities. Besides, some optimization-based methods [12, 30, 40] learn new parameters or fine-tune the diffusion model in order to enhance control over the generation process. Other methods [3, 58] leverage multi-view attention to introduce information from other-view images for gaining superior control. We incorporate multiple control techniques into the view-independent diffusion model for maintaining view consistency in the fine stage.

2.3 3D Representations

Neural Radiance Fields (NeRF) [29], as a volumetric rendering method, has gained popularity for its ability to enable 3D optimization [2, 4, 8, 13, 21] under 2D supervision, while NeRF optimization can be time-consuming. Numerous efforts [32, 41] for spatial pruning have been dedicated to accelerating the training process of NeRF on the reconstruction setting. however, they fail in the generation setting. Recently, 3D Gaussian Splatting [5, 9, 18, 50, 59] has emerged as an alternative 3D representation to NeRF and has shown remarkable advancements in terms of both quality and speed, offering a promising avenue. Therefore, we adopt Gaussian Splatting for efficiently generating coarse 3D model.

3 Preliminary

3.1 DDIM Inversion

DDIM [47] transforms random noise $\boldsymbol{x}_T$ into clean data $\boldsymbol{x}_0$ over a series of time steps, by using the deterministic DDIM sampling in the reverse process, i.e., $\boldsymbol{x}_{t-1} = (\alpha_{t-1}/\alpha_t)(\boldsymbol{x}_t - \sigma_t\epsilon_\phi) + \sigma_{t-1}\epsilon_\phi$. On the contrary, DDIM inversion progressively converts clean data to a noisy state $\boldsymbol{x}_T$, i.e., $\boldsymbol{x}_t = (\alpha_t/\alpha_{t-1})(\boldsymbol{x}_{t-1} -$

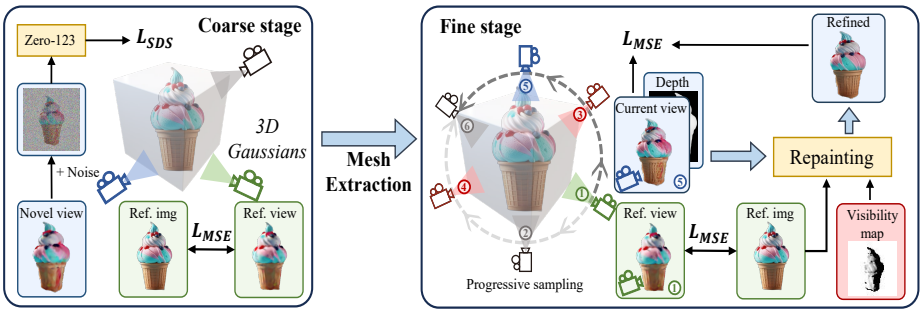

Fig. 3: Overview of Repaint123. In the coarse stage, we leverage 3D Gaussian Splatting to represent the coarse content, which is optimized by MSE loss at the reference and SDS loss at the invisible views. In the fine stage, we convert coarse content to mesh and progressively sample and generate invisible views from the reference view bidirectionally. Concretely, the current view is repainted based on the visibility map from bidirectional neighbor views with control techniques including attention injection and depth guidance, leading to consistent and detailed textures.

$\sigma_{t-1}\epsilon_\phi) + \sigma_t\epsilon_\phi$, here ϵ_ϕ is the predicted noise by the UNet. This method can precisely reconstruct the original clean data while greatly speeding up the process by skipping many intermediate diffusion steps.

3.2 3D Gaussian Splatting

Gaussian Splatting [18] presents a novel method for synthesizing new views and reconstructing 3D scenes, achieving real-time speed. Unlike NeRF, Gaussian Splatting uses a set of anisotropic 3D Gaussians defined by their locations, covariances, colors, and opacities to represent the scene. To compute the color of each pixel $\mathbf{p}$ in the image, it utilizes a typical neural point-based rendering [19, 20], The rendering process is as follows:

$$C(\mathbf{p}) = \sum_{i\in\mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1-\alpha_j)\,, \quad \text{where, } \alpha_i = o_i e^{-\frac{1}{2}(\mathbf{p}-\mu_i)^T \Sigma_i^{-1}(\mathbf{p}-\mu_i)}, \tag{1}$$

where c_i, o_i, μ_i, and Σ_i represent the color, opacity, position, and covariance of the i-th Gaussian respectively, and $\mathcal{N}$ denotes the number of related Gaussians.

4 Method

In this section, we introduce our two-stage framework named Repaint123 for fast and high-quality image-to-3D generation, as illustrated in Figure 3. In the coarse stage, we adopt 3D Gaussian Splatting [18] representation optimized by

SDS loss [34] with the view-conditioned diffusion model [24] to learn a coarse geometry and texture (see Section 4.1). Inspired by [50], we convert the coarse 3D Gaussians to textured mesh to facilitate subsequent texture enhancement in UV space. In the fine stage, we progressively repaint invisible content in novel views by rotating camera viewpoints bi-directionally from the reference view (see Section 4.2). To mitigate texture bias accumulated during the gradual repainting across wide angles, we incorporate multiple control techniques including depth guidance, injection of reference-view attention, and the use of reference-view image prompts (see Section 4.2). Additionally, we propose a visibility-based adaptive strategy to refine visible regions when observed from a superior perspective (see Section 4.2). After obtaining a high-quality repainted image that remains consistent across different views, we efficiently optimize the texture map by simple Mean Square Error (MSE) loss (see Section 4.2). This process is repeated for each sampled view, alternating between repainting and optimization, until the texture is completely reconstructed to accommodate a 360-degree perspective.

4.1 Coarse Stage: Gaussian Splatting with 3D Diffusion Prior

In the coarse stage, we adopt 3D Gaussian Splatting for efficient initialization. Through SDS [34], 3D diffusion priors are back-propagated to the 3D Gaussians. At each diffusion step, we sample a random current camera view v_c and its camera pose p_c, which orbits around the center of the object. Subsequently, the RGB image, denoted as I^c_{rgb}, and the alpha image, represented by I^c_{a}, are rendered from the current viewpoint p. Given an image $\tilde{I}^r_{\text{rgb}}$ and an alpha mask $\tilde{I}^r_{\text{a}}$ as input to the reference view, the SDS loss can be formulated as:

$$\mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,c,\epsilon}\left[(\epsilon_\phi(I^c_{\text{rgb}}; t, \tilde{I}^r_{\text{rgb}}, \Delta p_c) - \epsilon)\frac{\partial I^c_{\text{rgb}}}{\partial \Theta}\right] \tag{2}$$

where $\epsilon_\phi(\cdot)$ is the predicted noise by the diffusion ϕ, Δp_c is the relative camera pose and Θ represents the optimizable Gaussian parameters. Additionally, we optimize the RGB image and alpha image rendered from the reference view to align with the input reference image using reconstruction loss:

$$\mathcal{L}_{\text{Ref}} = \lambda_{\text{rgb}}||I^r_{\text{rgb}} - \tilde{I}^r_{\text{rgb}}||^2_2 + ||I^r_{\text{a}} - \tilde{I}^r_{\text{a}}||^2_2. \tag{3}$$

The final loss function is the weighted sum of $\mathcal{L}_{\text{SDS}}$ and $\mathcal{L}_{\text{Ref}}$.

4.2 Fine Stage: Visibility-based Controllable Repainting

Progressive Invisible Content Repainting. We repaint previously unseen content with progressively sampling novel camera viewpoints, beginning from the reference view and increasing a certain angular interval with each iteration. To obtain invisible regions in the image from a newly sampled view v_c, we use the back-projection technique based on the depth map $D^p \in \mathbb{R}^{H\times W}$ in the previously seen view v_p. Formally, given the camera intrinsics $K \in \mathbb{R}^{3\times 3}$, we first remap the

pixels of the image rendered from v_p into 3D points X^p expressed in the camera coordinate frame, which can be obtained as $X_{i,j} = K^{-1}\left[iD_{i,j}, jD_{i,j}, D_{i,j}\right]^\top$. With $P_p, P_c \in \mathbb{R}^{3\times4}$ the world-to-camera poses for images from v_p and v_c, we can transform X^p into camera coordinate frame of v_c to obtain $X^{p,c}$:

$$X^{p,c} = P_p P_c^{-1} h(X^p), \tag{4}$$

where $h : (x, y, z) \rightarrow (x, y, z, 1)$ is the homogeneous mapping. Then we can render a depth map $D^{p,c}$ from $X^{p,c}$ as $D^{p,c}_{i,j} = KX^{p,c}_{i,j,2}$. Regions with dissimilar depth values between D^c and $D^{p,c}$ are recognized as invisible regions, forming a repainting mask M where areas with zeros are invisible and areas with ones are visible. Therefore, we repaint invisible content and maintain visible content by:

$$x_t = x_t^{\text{inv}} \odot M + x_t^{\text{rev}} \odot (1 - M), \tag{5}$$

where $\odot$ is element-wise matrix multiplication, t is the current denoising timestep, x_t^{inv} is the DDIM [47] inverted latent, and $x_t^{\text{rev}} \sim \mathcal{N}\left(\mu_\phi\left(x_{t+1}, t\right), \Sigma_\phi\left(x_{t+1}, t\right)\right)$ is the diffusion model output.

Additionally, to mitigate accumulated bias for distant camera views during gradual repainting, as shown in the fine stage in Figure 3, we sample camera viewpoints bi-directionally, alternating between clockwise and counterclockwise directions. This bidirectional sampling allows us to repaint up to 180 degrees from two opposite directions, instead of repainting across the entire 360 degrees from one direction. Moreover, to repaint bidirectionally invisible content and ensure smooth transitions at the junction of the two directions, we replace the previous unidirectional repainting mask M by a bidirectional repainting mask. This mask merges two unidirectional masks from bidirectional neighbor views (red cameras in Figure 3) by taking the maximum value of each pixel.

Control-based Content Enhancement. While the progressive repainting maintains consistency between adjacent views, views that are farther apart may become inconsistent and experience a decline in quality due to cumulative errors. To tackle these issues, we improve the repainting process by incorporating various control techniques to enhance the consistency and quality of synthesized images. Specifically, for geometry control, we employ ControlNet [62] to condition the diffusion model with coarse depth maps, ensuring geometric consistency across images. To transfer reference textures and mitigate the cumulative texture bias during the progressive repainting, inspired by HiFi-123 [61], we incorporate an attention injection mechanism [3] that injects reference attention features into the diffusion denoising process. By replacing the novel-view content features (Key features K_t and Value features V_t) with reference-view attention features (K_r and V_r) during each denoising step, the novel-view image features can directly query the high-quality reference features by:

$$\text{Attention}(Q_t, K_r, V_r) = \text{Softmax}\left(\frac{Q_t K_r^T}{\sqrt{d}}\right) V_r, \tag{6}$$

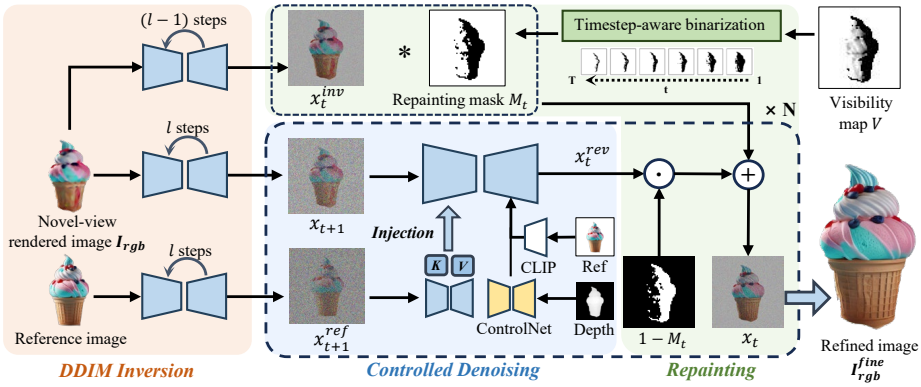

Fig. 4: Visibility-based Controllable Repainting Pipeline. Our method repaints invisible content (black areas in M_t) of novel-view rendered image I_{rgb} according to the inverted latent x_t^{inv} and the visibility map V to obtain refined image $I_{\text{rgb}}^{\text{fine}}$. We introduce multiple control techniques to enhance the consistency and quality of generated textures, including attention feature injection, depth guidance by ControlNet, and semantic prompt by CLIP encoder. Noticing that the repainting mask M_t is simultaneously changed with denoising timesteps for adaptive refinement.

where Q_t represents the novel-view query features. This enhancement facilitates the transfer of texture details and improves the consistency between the reference-view image and novel-view images.

During our experiments, we observed a decline in image quality during repainting, likely due to not using text prompts for performing classifier-free guidance [15], crucial for producing high-quality images in diffusion models. To improve this, we adopt IP-Adapter [58], which uses a CLIP encoder to convert reference images into semantic features. These features are then projected into image prompts, providing visual cues that enhance image generation through classifier-free guidance.

Adaptive Visible Content Refinement. Repainting only the invisible content often leads to distortion in the visible area As shown in Figure 5, when the previous view is an oblique view, it leads to a low-resolution update on the texture maps, resulting in high distortion when rendering from a better observing view. Therefore, the visible content currently observed in a superior view should be refined to replenish details. However, the selection of refinement strength for these visible content is tricky [38, 54], as excessive strength produces unfaithful results leading to inconsistency with previously repainted views, while insufficient strength limits quality improvement. We propose a visibility-based adaptive refinement strategy to refine these previously seen regions with different strengths, aiming to achieve improved quality-consistency trade-off. As visualized in Figure 5, we view repainting for visible content as a process similar to super-resolution that replenishes detailed information. According to the Orthographic

Projection Theorem, which asserts that the projected resolution of a surface is directly proportional to the view obliqueness ($cos\theta$ as shown in Figure 5), we can assume that the repainting strength is equal to $(1 - cos\theta_1/cos\theta_2)$, where $cos\theta_2/cos\theta_1$ represents the upsampling scale in terms of super-resolution and visibility variations in the context of 3D rendering. Therefore, we can define a visibility map V as $V_{i,j} = cos\theta^{p,c}_{i,j}/cos\theta^{c}_{i,j}$, where $cos\theta^{p,c}_{i,j}$ and $cos\theta^{c}_{i,j}$ are obtained by multiplying the current-view normal map with camera rays of the previous view v_p and the current view v_c respectively. Higher visibility value indicates less repainting strength, while invisible areas require full strength.

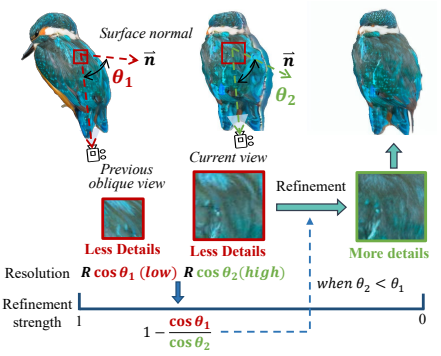

Fig. 5: Relation between view obliqueness and refinement strength. Red box shows the same texture part in different views. Texture details in generated views might be insufficient due to the variation of angles.

By this association between repainting strength and the visibility map V, we can binarize the visibility map to the timestep-aware adaptive repainting mask M_t during each denoising step, visualized in the green box "Timestep-aware binarization" in Figure 4:

$$M_t^{i,j} = \begin{cases} 1, & \text{if } V^{i,j} > 1 - t/T \\ 0, & \text{else,} \end{cases} \tag{7}$$

where i, and j are the 2D position of pixels in visibility map V, and T is the total number of timesteps of the diffusion model. By doing this, we can adaptively refine visible content with a proper strength.

Efficient Optimization. Using a progressive repainting strategy and various control techniques to refine the inverted latent, we obtain a high-quality, view-consistent refined image, $I^{\text{fine}}_{\text{rgb}}$. This process allows us to use simple MSE loss for efficient texture map optimization:

$$\mathcal{L}_{\text{fine}} = ||I^{\text{fine}}_{\text{rgb}} - I_{\text{rgb}}||^2_2, \tag{8}$$

where I_{rgb} represents the rendered image in the fine stage. MSE loss is faster and more deterministic to optimize than the traditional SDS loss, speeding up the optimization process. After sampling a novel view, we first conduct image repainting and then proceed with efficient texture optimization, repeating this process until we cover a 360-degree view range.

5 Experiment

5.1 Implementation Details

In our experiment, we consistently apply the same hyperparameters across all our method's results.We progressively turn around the viewpoints by 40 degrees each time, consequently obtaining 8 views per object. We use a 50-step DDIM schedule and perform a 30-step latent diffusion inversion. Stable Diffusion 1.5 is utilized for all methods. During the coarse stage training, we set λ_{rgb} to 10. Generating a single 3D object takes only 2 minutes on a single 40G A100 GPU, approximately 1 minute for the coarse stage and mesh extraction, and another 1 minute for refinement.

5.2 Baselines

We adopt RealFusion [28], Make-It-3D [51], and Zero123-XL [24], Magic123 [35] as our NeRF-based baselines and DreamGaussian [50] as our Gaussian-Splatting-based baseline. RealFusion presents a single-stage algorithm for NeRF generation leveraging 2D SDS loss for novel views. Make-It-3D is a two-stage approach that shares similar objectives with RealFusion but employs a point cloud representation for refinement at the second stage. For Zero123-XL, we adopt the implementation [49] and add a mesh fine-tuning stage for fair comparison. Integrating Zero123 and RealFusion, Magic123 incorporates a 2D SDS loss with 3D SDS loss provided by Zero123 to balance geometry and texture and adopts DMTet [45] representation at the second stage. DreamGaussian integrates 3D Gaussian Splatting into 3D generation and greatly improves the speed.

5.3 Evaluation Protocol

Datasets. Based on previous research, we utilized the Realfusion15 dataset [28] and test-alpha dataset collected by Make-It-3D [51], which comprises many common objects from diverse styles.

Evaluation metrics. An effective 3D generation approach should produce 3D content which closely resemble the reference view, and maintain consistency of semantics and textures with the reference image when observed from new views. Therefore, to evaluate the overall quality of the generated 3D object, we choose the following metrics from two aspects: 1) PSNR and LPIPS [63], which measure pixel-level and perceptual reconstruction quality respectively at the reference view; 2) CLIP Similarity [36] and Contextual Distance [27], which assess the similarity of semantics and detailed textures respectively between the novel perspective and the reference view.

Fig. 6: Qualitative comparisons on image-to-3D generation. Zoom in for details.

Dataset	Metrics\Methods	NeRF-based				Gaussian-Splatting-based	
		RealFusion	Make-it-3D	Zero-123-XL*	Magic123	DreamGaussian	**Repaint123**
RealFusion15	CLIP-Similarity↑	0.71	0.81	0.83	0.82	0.77	**0.85**
	Context-Dis↓	2.20	1.82	1.59	1.64	1.61	**1.55**
	PSNR↑	19.24	16.56	19.56	**19.68**	18.94	<u>19.00</u>
	LPIPS↓	0.194	0.177	0.108	0.107	0.111	**0.101**
Test-alpha	CLIP-Similarity↑	0.68	0.76	0.84	0.84	0.79	**0.88**
	Context-Dis↓	2.20	1.73	1.52	1.57	1.62	**1.50**
	PSNR↑	22.91	17.21	24.39	**24.69**	22.33	<u>22.38</u>
	LPIPS↓	0.105	0.237	0.050	**0.046**	0.057	<u>0.048</u>
	Optimization time	20min	1h	30min	1h (+2h)	**2min**	**2 min**

Table 1: We show quantitative results in terms of CLIP-Similarity↑ / Contextual-Distance↓ / PSNR↑ / LPIPS↓. The results are shown on the RealFusion15 and test-alpha datasets, while **bold** reflects the best for all methods and the **underline** represents the best for Gaussian-Splatting-based methods. * indicates that Zero123-XL incorporates a mesh fine-tuning stage for further quality improvement. The time required by textual inversion is indicated in parentheses.

5.4 Comparisons

Comparisons with NeRF-based Methods. As shown in Table 1, we evaluate the quality of generated 3D objects across various NeRF-based methods. Our method achieves superior 3D consistency in generating 3D objects, as evidenced by the best performance of CLIP-similarity and Contextual-distance. Regarding reference-view reconstruction quality, there is a gap compared with NeRF-based approaches as shown in Table 1, which we attribute to the immaturity of current Gaussian-Splatting-based methods. Compared to NeRF-based methods for the optimization time, our approach reaches a significant acceleration of over 10 times and simultaneously achieves high quality, due to the 3D Gaussian-Splatting representation in coarse stage and efficient texture refinement in fine stage. As shown in Figure 6, Repaint123 achieves the best visual results in terms of texture consistency and generation quality as opposed to other NeRF-based methods. From the visual comparison, our method achieves consistent and detailed textures in invisible areas, while Zero123-XL results in over-smooth textures, Magic123 produces inconsistent and oversaturated colors, Realfusion and Make-It-3D fail to generate full geometry and consistent textures.

Comparisons with GS-based methods. We conduct comparisons with other works based on 3D Gaussian Splatting in the last column of Table 1. Within a comparable generation time, our method demonstrates superiority over existing Gaussian-based approach in aspects of texture consistency and the quality of reference-view reconstruction. The superiority of our proposed Repaint123 is evidenced by the evaluation of four distinct metrics from the Table 1. As shown in Figure 6, DreamGaussian usually leads to over-smooth texture inconsistent with reference view, while our method can produce high-quality 3D content with view-consistent and detailed textures.

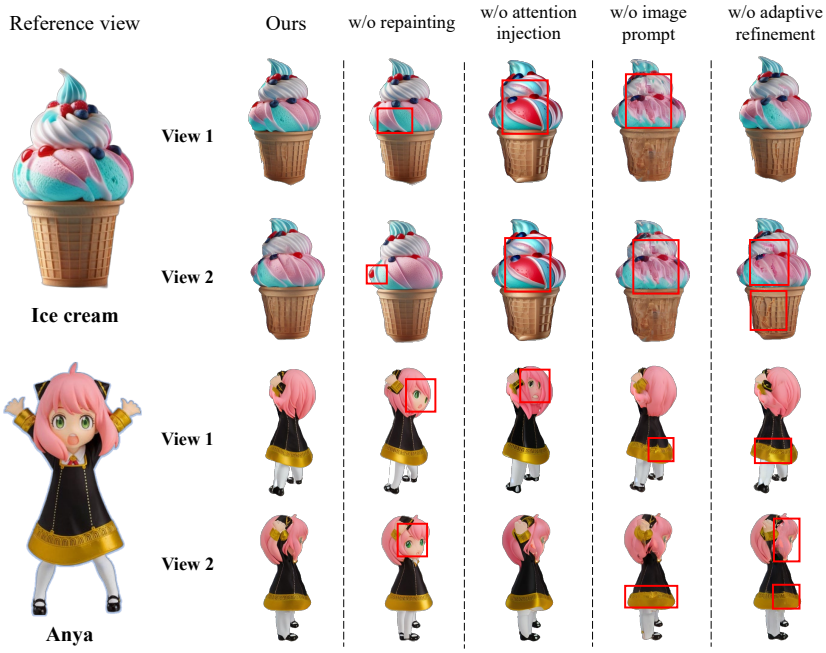

Fig. 7: Qualitative ablation study by removing one component at a time from the overall method. The presence of artifacts is highlighted by red boxes. Inconsistencies such as multi-face problems and mismatches in content and style are evident without using repainting and attention injection strategies. Absence of image prompting and adaptive refinement notably degrades quality.

5.5 Ablation and Analysis

This section details both qualitative and quantitative analyses to highlight the effectiveness of our proposed methods, as illustrated in Figure 7 and Table 2.

Effectiveness of Repainting. Figure 7 illustrates that without repainting strategy, the generated novel-view image tends to resemble the reference image, leading to inconsistencies in the shared visible areas between different views, such as the content misalignment (observed in the Ice Cream example) and issues with multiple faces (evident in the Anya example). These issues stem from the absence of alignment constraints for shared visible regions, leading to conflicts and quality degradation of the reconstructed 3D texture.

Impact of Attention Injection. Table 2 shows that injecting attention features from the reference-view image markedly enhances the consistency of both semantics and fine-grained textures, as evidenced by increased CLIP similarity and reduced Contextual distance. Without this injection strategy, as depicted in Figure 7, the synthesized images retain basic semantics but fail to accurately transfer detailed textures and styles from the reference view, resulting in inconsistencies across multiple views.

Method \ Metric	CLIP↑	Contextual↓	PSNR↑	LPIPS↓
Coarse	0.71	1.78	21.17	0.133
vanilla repainting	0.71	1.62	22.41	0.049
+attention injection	0.78	1.56	**22.42**	**0.048**
+image prompt	0.84	1.52	22.40	**0.048**
+adaptive refinemnet (**Ours**)	**0.88**	**1.50**	22.38	**0.048**

Table 2: Quantitative ablation study on Test-alpha dataset by *progressively* adding our proposed components. The last three lines show the cumulative effects of the proposed modules. Our approach, utilizing all strategies, delivered the best performance.

Role of Image Prompt. The introduction of using the reference image as an image prompt, as shown in Table 2 and Figure 7, significantly boosts both multi-view consistency and image quality of the generated images. Without this technique to perform classifier-free guidance, the generation of detailed and consistent textures across views is compromised.

Benefits of Adaptive Refinement. The necessity of adaptive refinement is clear from Figure 7, where its absence leads to artifacts and blurriness in obliquely viewing areas of previous views due to low-resolution updates, as mentioned in Section 4.2. Table 2 also demonstrates its benefits through improved CLIP similarity and Contextual distance.

6 Limitations

Despite the promising results, our method still has some limitations. Our method is based on 3D Gaussian Splatting representation in the coarse stage during training. While 3D Gaussian Splatting accelerates the training process, it may exhibit geometry artifacts during mesh extraction, such as holes, and inferior results compared to NeRF-based methods for the reconstruction of reference view, due to immaturity in generation tasks. This is supported by results in the Appendix, where NeRF is considered as an alternative to 3D Gaussian Splatting for the coarse stage. We expect these limitations can be mitigated with the development of 3D Gaussian Splatting in the future.

7 Conclusion

This work presents Repaint123 for generating high-quality 3D content from a single image in about 2 minutes. By leveraging progressive controllable repaint, our approach overcomes the limitations of existing studies and achieves state-of-the-art results in terms of both texture quality and multi-view consistency, paving the way for future progress in 3D content generation from one image. Furthermore, we validate the effectiveness of our proposed method through a comprehensive set of experiments.

Acknowledgements

This work was supported in part by Natural Science Foundation of China (No. 62332002, 62202014), and Shenzhen Basic Research Program (No.JCYJ20220813151736001)

References

1. Armandpour, M., Zheng, H., Sadeghian, A., Sadeghian, A., Zhou, M.: Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. arXiv preprint arXiv:2304.04968 (2023)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
3. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22560–22570 (October 2023)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022)
5. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2024)
6. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
7. Chen, Y., Zhang, C., Yang, X., Cai, Z., Yu, G., Yang, L., Lin, G.: It3d: Improved text-to-3d generation with explicit view synthesis. arXiv preprint arXiv:2308.11473 (2023)
8. Chen, Z., Funkhouser, T., Hedman, P., Tagliasacchi, A.: Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. arXiv preprint arXiv:2208.00277 (2022)
9. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. arXiv preprint arXiv:2309.16585 (2023)
10. Cheng, X., Yang, T., Wang, J., Li, Y., Zhang, L., Zhang, J., Yuan, L.: Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. arXiv preprint arXiv:2310.11784 (2023)
11. Dou, Z., Wu, Q., Lin, C., Cao, Z., Wu, Q., Wan, W., Komura, T., Wang, W.: Tore: Token reduction for efficient human mesh recovery with transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15143–15155 (2023)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
13. Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P.: Baking neural radiance fields for real-time view synthesis. ICCV (2021)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

16. Huang, Y., Wang, J., Shi, Y., Qi, X., Zha, Z.J., Zhang, L.: Dreamtime: An improved optimization strategy for text-to-3d content creation. arXiv preprint arXiv:2306.12422 (2023)
17. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
18. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (ToG) **42**(4), 1–14 (2023)
19. Kopanas, G., Leimkühler, T., Rainer, G., Jambon, C., Drettakis, G.: Neural point catacaustics for novel-view synthesis of reflections. ACM Transactions on Graphics (TOG) **41**(6), 1–15 (2022)
20. Kopanas, G., Philip, J., Leimkühler, T., Drettakis, G.: Point-based neural rendering with per-view optimization. In: Computer Graphics Forum. vol. 40, pp. 29–43. Wiley Online Library (2021)
21. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: CVPR (2023)
22. 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: CVPR (2023)
23. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)
24. 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 (2023)
25. 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)
26. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
27. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European conference on computer vision (ECCV). pp. 768–783 (2018)
28. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: CVPR (2023)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
30. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
31. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM TOG (2022)
33. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)

34. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
35. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al.: Dreambooth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508 (2023)
38. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
41. Sara Fridovich-Keil and Alex Yu, Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR (2022)
42. Seo, H., Kim, H., Kim, G., Chun, S.Y.: Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. arXiv preprint arXiv:2304.02827 (2023)
43. Seo, J., Jang, W., Kwak, M.S., Ko, J., Kim, H., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. arXiv preprint arXiv:2303.07937 (2023)
44. Shen, Q., Yang, X., Wang, X.: Anything-3d: Towards single-view anything reconstruction in the wild. arXiv preprint arXiv:2304.10261 (2023)
45. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
46. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
47. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020), https://arxiv.org/abs/2010.02502
48. Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. arXiv preprint arXiv:2306.07881 (2023)
49. Tang, J.: Stable-dreamfusion: Text-to-3d with stable-diffusion (2022), https://github.com/ashawkey/stable-dreamfusion
50. 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)
51. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: ICCV (2023)
52. Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., Tombari, F.: Textmesh: Generation of realistic 3d meshes from text prompts. arXiv preprint arXiv:2304.12439 (2023)

53. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
54. Wang, T., Kanakis, M., Schindler, K., Van Gool, L., Obukhov, A.: Breathing new life into 3d assets with generative repainting. arXiv preprint arXiv:2309.08523 (2023)
55. 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)
56. Wu, J., Gao, X., Liu, X., Shen, Z., Zhao, C., Feng, H., Liu, J., Ding, E.: Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. arXiv preprint arXiv:2307.16183 (2023)
57. Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 views. arXiv e-prints pp. arXiv–2211 (2022)
58. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023)
59. 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)
60. Yu, C., Zhou, Q., Li, J., Zhang, Z., Wang, Z., Wang, F.: Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. arXiv preprint arXiv:2307.13908 (2023)
61. Yu, W., Yuan, L., Cao, Y.P., Gao, X., Li, X., Quan, L., Shan, Y., Tian, Y.: Hifi-123: Towards high-fidelity one image to 3d content generation. arXiv preprint arXiv:2310.06744 (2023)
62. 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 (2023)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
64. Zhu, J., Zhuang, P.: Hifa: High-fidelity text-to-3d with advanced diffusion guidance. arXiv preprint arXiv:2305.18766 (2023)