VISIBILITY-UNCERTAINTY-GUIDED 3D GAUSSIAN INPAINTING VIA SCENE CONCEPTIONAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

3D Gaussian Splatting (3DGS) has emerged as a powerful and efficient 3D representation for novel view synthesis. This paper extends 3DGS capabilities to inpainting, where masked objects in a scene are replaced with new contents that blend seamlessly with the surroundings. Unlike 2D image inpainting, 3D Gaussian inpainting (3DGI) is challenging in effectively leveraging complementary visual and semantic cues from multiple input views, as occluded areas in one view may be visible in others. To address this, we propose a method that measures the visibility uncertainties of 3D points across different input views and uses them to guide 3DGI in utilizing complementary visual cues. We also employ the uncertainties to learn a semantic concept of the scene without the masked object and use a diffusion model to fill masked objects in the input images based on the learned concept. Finally, we build a novel 3DGI framework, VISTA, by integrating VISibility-uncerTainty-guided 3DGI with scene conceptuAl learning. VISTA generates high-quality 3DGS models capable of synthesizing artifact-free and naturally inpainted novel views. Furthermore, our approach extends to handling dynamic distractors arising from temporal object changes, enhancing its versatility in diverse scene reconstruction scenarios. We demonstrate the superior performance of our method over state-of-the-art techniques using two challenging datasets: the SPIn-NeRF dataset, featuring 10 diverse static 3D inpainting scenes, and an underwater 3D inpainting dataset derived from UTB180, which includes fast-moving fish as inpainting targets.

1 INTRODUCTION

3D representation effectively models a scene and has the ability to synthesize new views of the scene (Barron et al., 2021; Mildenhall et al., 2021; Wang et al., 2021; Kerbl et al., 2023). 3D Gaussian splatting (3DGS) methods have been demonstrated as efficient and effective ways to represent the scene from a set of images taken from different viewpoints (Kerbl et al., 2023; Tang et al., 2023; Wu et al., 2024). Further, enabling editability of 3D scene representations is a cornerstone of technologies like augmented reality and virtual reality Tewari et al. (2022). 3D Gaussian inpainting task is one of the key editing techniques, aiming to replace specified objects with new contents that blend seamlessly with the surroundings. This capability allows us to: *(1) Remove objects from static scenes:* given multi-view images, we can create a 3D representation that generates novel views with specific objects removed and believably filled in (Figure 1 (Upper)). *(2) Clean up dynamic scenes:* for scenes with moving elements like fish in the water (see Figure 1 (Bottom)), we can build a 3D representation that excludes these transient objects, enabling clear, consistent novel view synthesis.

However, such an important task is non-trivial and the key challenge is how to leverage the complementary visual and semantic cues from multiple input views. Intuitively, for a synthesized view, the ideal approach is to replace the targeted erasure region with the occluded content, which naturally completes the inpainting. The key information for this process lies within the other view images, where the obscured areas may be visible from different angles. However, how to utilize multi-view information effectively is still an open question. State-of-the-art works first remove the targeted erasure region-related Gaussians and fill the regions via 2D image inpainting method (Ye et al., 2024; Wang et al., 2024), which, however, neglects the complementary cues from other views. The latest work (Liu et al., 2024) leverages depth maps of different views to involve the cross-view complementary cues implicitly. However, depth maps cannot fully represent complementary cues, such as the texture pattern from adjacent perspectives, and the depth project can hardly get high-quality depth

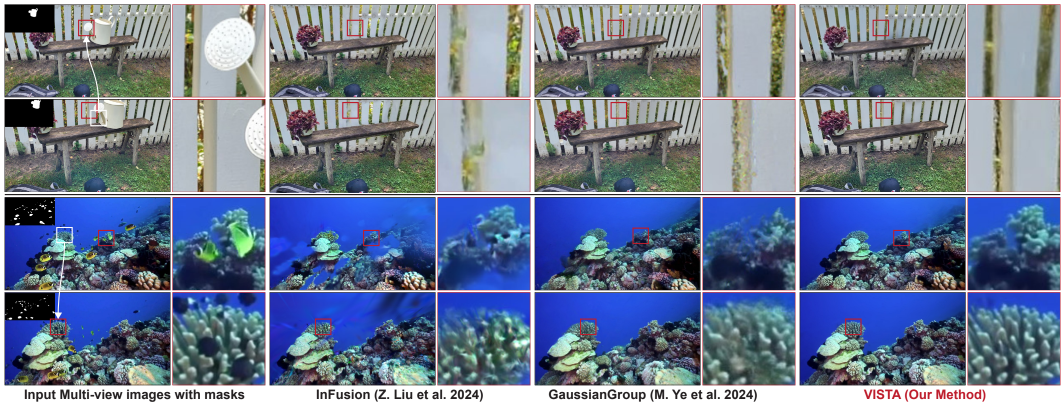


Figure 1: Two examples demonstrating the application of two state-of-the-art methods, namely InFusion (Liu et al., 2024) and GaussianGroup (Ye et al., 2024), alongside our proposed method for 3D Gaussian inpainting to fill masked static and dynamic objects, respectively. The red boxes highlight the advantages of our method and are enlarged on the right side of each image for better visibility. The white boxes and arrows indicate complementary visual cues between two different viewpoints

maps when moving objects across different views. As the two cases shown in Figure 1, InFusion synthesizes new views with obvious artifacts.

In this work, we propose VISibility-uncerTainty-guided 3DGI via scene conceptuAl Learning (VISTA), a novel framework for 3D Gaussian inpainting that leverages complementary visual and semantic cues. Our approach begins by measuring the visibility of 3D points across different views to generate visibility uncertainty maps for each input image. These maps indicate which pixels are most valuable for the inpainting task, based on the principle that pixels visible and consistent from multiple views contribute more significantly. We then integrate these visibility uncertainty maps into the 3D Gaussian splatting (3DGS) process. This enables the resulting Gaussian model to synthesize new views where masked regions are seamlessly filled with visual information from complementary perspectives. To address scenarios where large masked regions lack complementary visual cues from other views, we propose learning the concept of the scene without the masked objects. This conceptual learning is guided by the prior inpainting mask and the visibility uncertainty maps derived from the input multi-view images. The learned concept is then utilized to refine the input images, effectively filling the masked objects through a pre-trained Diffusion model. Furthermore, we implement an iterative process alternating between visibility-uncertainty-guided 3DGI and scene conceptual learning, progressively refining the 3D representation. As illustrated in Figure 1 (Upper), our method successfully reconstructs high-quality 3D representations of static scenes, naturally filling masked object regions with contextually appropriate content. Additionally, VISTA demonstrates its versatility by effectively removing distractors in dynamic scenes (see Figure 1 (Bottom) for examples).

We demonstrate the superior performance of our method over state-of-the-art techniques using two challenging datasets: the SPIn-NeRF dataset, featuring 10 diverse static 3D in-painting scenes, and an underwater 3D inpainting dataset derived from UTB180, which includes fast-moving fish as inpainting targets. In summary, the contributions of our work are as follows:

1. We propose VISibility-uncerTAinty-guided 3D Gaussian inpainting (VISTA-GI) that explicitly leverages multi-view information through visibility uncertainty, achieving 3D Gaussian inpainting for more coherent and accurate scene completions.

2. We propose VISibility-uncerTAinty-guided scene conceptual learning (VISTA-CL) and leverage it for diffusion-based inpainting. VISTA-CL fills masked regions in input images using learned scene concepts, addressing the inpainting task at its core. This approach enhances the fundamental understanding of the scene, leading to more accurate and contextually appropriate inpainting results.

3. We introduce VISTA (VISibility-uncerTainty-guided 3D gaussian inpainTing via scene conceptuAl learning), a novel framework that iteratively combines VISTA-GI and VISTA-

CL. This approach simultaneously leverages complementary visual and semantic cues, enhancing 3D Gaussian inpainting with geometric and conceptual information.

4. We extend VISTA to handle dynamic distractor removal in 3D Gaussian splatting, significantly improving its performance on scenes with temporal variations and outperforming state-of-the-art methods.

2 Related Work

2.1 NeRF and 3D Gaussian Splatting

The challenge of reconstructing a scene from 2D images to obtain suitable new viewpoints is a complex and worthy topic of exploration in computer vision and computer graphics (Lombardi et al., 2019; Kutulakos & Seitz, 2000). Recently, NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023) have emerged as two distinct approaches to 3D reconstruction, continuously improving the quality of the reconstructions.

Neural Radiance Fields (NeRF) is an implicit representation method for 3D reconstruction. It utilizes deep learning techniques to extract the geometric shapes and texture information of objects from images taken from multiple viewpoints, and it uses this information to generate a continuous 3D radiance field, allowing for highly realistic 3D models to be presented from any angle and distance (Barron et al., 2021). However, their excessively high training and rendering costs (Barron et al., 2022; 2023) often result in poor performance in practical applications. To resolve these issues, 3D Gaussian splitting (3DGS) is promoted as an explicit representation method that achieves state-of-the-art real-time rendering of high-quality images (Lu et al., 2024). 3DGS explicitly models the space as multiple Gaussian blobs, each with specific 3D positions, opacity, anisotropic covariance, and color features. Through training, it achieves an explicit representation of the three-dimensional space, enabling real-time synthesis of high-quality viewpoint images.

2.2 2D inpainting and 3D inpainting

2D inpainting is an elemental task in image generation. The task aims to use the pre-generated mask to create appropriate content for the masked area. Traditional patch-based methods Ružić & Pižurica (2014) and later GAN-based (Goodfellow et al., 2014) methods Yu et al. (2018) could somewhat inpaint regular and small mask areas, but they fail in complex scenes or when there are significant content omissions. Recently, diffusion models Ho et al. (2020); Sohl-Dickstein et al. (2015); Song et al. (2020) have become the most powerful technology in inpainting (Lugmayr et al., 2022; Suvorov et al., 2022; Li et al., 2022) for their ability to generate new, semantically plausible content.

Meanwhile, 3D inpainting to edit the scene reconstructed by NeRF or 3DGS is still a challenging task because of the complexity of spatial representation. NeRF-based inpainting Liu et al. (2022); Mirzaei et al. (2023); Weder et al. (2023) succeed in inpainting the static objects in the implicit representation. However, their performance is limited because of NeRF's obstacles. 3DGS-based inpainting methods such as Gaussian Grouping (Ye et al., 2024), InFusion (Liu et al., 2024), and GaussianEditor (Wang et al., 2024) focus on inpainting an existing static Gaussian Splatting scene, but neglecting the dynamic distractors that may appear before obtaining the static scene. GScream (Wang et al., 2025) focuses on removing objects by introducing monocular depth estimation and employing cross-attention to enhance texture. It remains a method focused on static objects. SpotLessSplats (Sabour et al., 2024) notices the dynamic distractors and repairs these areas using the pre-predicted masks, but it fails to repair occluded and completely unseen areas.

3 Preliminaries: 3D Gaussian Splatting and Inpainting

3.1 3D Gaussian Splatting

Given a set of images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ captured from various viewpoints and timestamps, 3D Gaussian splatting (3DGS) aims to learn a collection of anisotropic Gaussian splats $\mathcal{G} = \{\mathbf{g}_j\}_{j=1}^{M}$ from these multi-view images. Each splat $\mathbf{g}_j$ is characterized by a Gaussian function with mean μ_j, a positive semi-definite covariance matrix $\sum_j$, an opacity α_j, and view-dependent color coefficients $\mathbf{c}_j$. Once the parameters of the 3D Gaussian splats $\mathcal{G}$ are determined, novel view synthesis can be achieved through alpha-blending: $\hat{\mathbf{I}}^{p} = \text{Render}(\mathcal{G}, \mathbf{p})$. We can use $\mathcal{I}$ to supervise the optimization of $\mathcal{G}$

$$\arg\min_{\mathcal{G}} \lambda_1 \sum_{i=1}^{N} \|(\mathbf{I}_i - \hat{\mathbf{I}}^{p_i})\|_1 + \lambda_2 \sum_{i=1}^{N} \text{D-SSIM}(\mathbf{I}_i, \hat{\mathbf{I}}^{p_i}), \tag{1}$$

where $\mathbf{p}_i$ denotes the camera perspective of the image $\mathbf{I}_i$, $\hat{\mathbf{I}}^{p_i} = \text{Render}(\mathcal{G}, \mathbf{p}_i)$, and $\lambda_1 + \lambda_2 = 1$. For novel view synthesis, given a camera perspective $\mathbf{p}$, the process involves the following steps: projecting each 3D Gaussian onto a 2D image plane, sorting the Gaussians by depth along the view direction, and blending the Gaussians from front to back for each pixel. A key advantage of 3DGS (Kerbl et al., 2023) is its ability to synthesize a new view in a single pass, whereas NeRF requires pixel-by-pixel rendering. This efficiency makes 3DGS particularly well-suited for time-sensitive 3D representation applications, offering a significant performance boost over NeRF.

3.2 3D Gaussian Inpainting

Given a set of captured images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ and corresponding binary mask maps $\mathcal{M} = \{\mathbf{M}_i\}_{i=1}^{N}$ delineating objects for removal (See Figure 1), 3D Gaussian Inpainting (3DGI) constructs a new 3D Gaussian splatting (3DGS) representation. This representation eliminates specified objects and replaces them with content that integrates with the environment. The resulting 3DGS representation can synthesize arbitrary views where the specified objects are imperceptibly absent, maintaining visual coherence across viewpoints while effectively 'erasing' targeted objects. We can use the segment anything model (SAM) (Kirillov et al., 2023) with few manual annotations to generate mask maps, aligning with methods like (Ye et al., 2024) for precise object delineation.

SOTA methods and limitations. An intuitive approach to 3D Gaussian Inpainting (3DGI) involves deriving a 3D mask for the specified objects based on the provided 2D masks. The process of new view synthesis then follows a two-step procedure: first, generating the specified view and its corresponding mask, and then applying existing 2D image inpainting techniques to achieve the desired 3DGI effect. This methodology has been adopted in recent works by Wang et al. (2024) and Ye et al. (2024). However, this approach does not leverage the complementary information available across multiple viewpoints during the inpainting process. A key example is the failure to utilize information from regions that may be occluded in one view but visible in another. Consequently, this method struggles to maintain consistency with the surrounding environment, particularly when dealing with large masked regions. This limitation underscores the need for more sophisticated techniques to effectively integrate and synthesize information from multiple perspectives to achieve more coherent and realistic 3D inpainting results. Beyond this solution, the latest work Liu et al. (2024) utilizes the cross-view complementary cues through depth perception. It formulates the 3D Gaussian inpainting as two tasks, *i.e.*, 2D image inpainting and depth inpainting, and the complementary cues in multiple views are implicitly utilized via depth projection. However, depth maps cannot fully represent complementary cues, such as the texture pattern from adjacent perspectives, and the depth project can hardly get high-quality depth maps when moving objects across different views. As case 2 shown in Figure 1, InFusion synthesizes new views with obvious artifacts.

4 Methodology

This section details the proposed framework called VISibility-uncerTainty-guided 3D Gaussian inpainting via scene concepTional learning (VISTA). The core principle is to identify the visibility of 3D points across different views and utilize this information to guide the use of complementary visual and semantic cues for 3D Gaussian inpainting.

To elucidate this concept, we introduce the visibility-uncertainty-guided 3D Gaussian inpainting (VISTA-GI) in Section 4.1, where we define the visibility uncertainty of 3D points and employ it to guide the use of complementary visual cues for 3DGI. In Section 4.2, we propose leveraging the visibility uncertainty to learn the semantic concept of the scene without specified objects. We then perform concept-driven Diffusion inpainting to process the input images, harnessing complementary semantic cues. To fully utilize complementary visual and semantic cues, we propose in Section 4.3 an iterative combination of VISTA-GI and VISTA-CL. Finally, in Section 4.4, we extend our VISTA

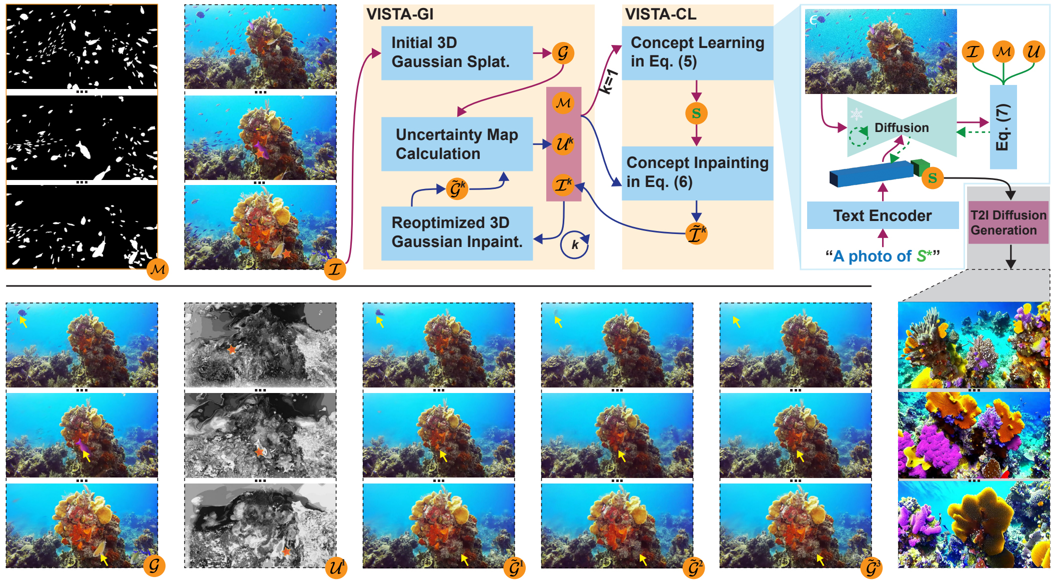


Figure 2: Framework of VISTA comprising two modules: VISTA-GI (described in Section 4.1) and VISTA-CL (detailed in Section 4.2). Results from three views are displayed for key variables in the framework. Note that $\mathcal{G}$, $\tilde{\mathcal{G}}^1$, $\tilde{\mathcal{G}}^2$, and $\tilde{\mathcal{G}}^3$ are 3DGS representations, and the displayed examples are rendered from these representations. The last column shows generated images derived from the learned scene concept. In the uncertainty map, we use ☆ to highlight areas of high uncertainty, which denote points (e.g., dynamic fishes) visible from only a few views. Yellow arrows demonstrate the progressive improvement in inpainting quality achieved by our method.

framework to address the challenge of dynamic distractors in captured images. This extension excludes transient objects, resulting in clearer and more consistent novel view synthesis.

4.1 VISTA-GI: Visibility-uncertainty-guided 3D Gaussian Inpainting

Initial 3D Gaussian Splatting. Given the input images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$, we employ the original 3DGS method in Section 3.1 and Equation (1) to construct a 3D representation $\mathcal{G}$. This representation can then be utilized to render novel views. However, as illustrated in Figure 2, this initial representation fails to exclude dynamic objects (such as fish) and exhibits noticeable artifacts, including blurring.

Visibility uncertainty of 3D Points. We define a set of adjacent camera perspectives/views denoted as $\mathcal{P} = \{\mathbf{p}_v\}_{v=1}^{V}$, where V is the number of adjacent views. For a 3D point $\mathbf{X}$ in the scene, we can project it to different camera perspectives in $\mathcal{P}$ via the built 3DGS $\mathcal{G}$ and get their colors under V views, *i.e.*, $\{\mathbf{x}_v\}_{v=1}^{V}$. Then, we calculate the variations of colors of the point under different views

$$u_{\mathbf{x}} = \mathrm{var}(\{\mathbf{x}_v\}_{v=1}^{V}), \tag{2}$$

where $\mathrm{var}(\cdot)$ is the variation function. We denote the result $u_{\mathbf{x}}$ as the *visibility uncertainty* of the 3D point $\mathbf{X}$. Intuitively, $u_{\mathbf{x}}$ represents the visibility and consistency of the point across the V views. For example, if the point $\mathbf{X}$ can be seen at all views, the colors under different views are consistent and $u_{\mathbf{x}}$ is small. If the point can be only seen by a few views or its color deviates between different views, the visibility uncertainty tends to be significantly high.

Reoptimized 3D Gaussian inpainting. With the 3D point's visibility uncertainty, we aim to calculate the visibility uncertainty map of the input image and measure the visibility of each pixel at other views. Specifically, for an image $\mathbf{I}_i$ in $\mathcal{I}$, we first calculate its depth map $\mathbf{D}_i$ based on the $\mathcal{G}$. Then, we project each pixel of $\mathbf{I}_i$ to a 3D point and calculate its visibility uncertainty via Equation (2) under V adjacent views. Then, we obtain a pixel-wise visibility uncertainty map, which is normalized by dividing each pixel's uncertainty value by the standard deviation computed across all uncertainty values. The resulting normalized map is denoted as $\mathbf{U}_i$. For the N input images, we have N visibility uncertainty maps $\mathcal{U} = \{\mathbf{U}_i\}_{i=1}^{N}$. Then, we use them to update the original mask maps $\mathcal{M}$ and uncertainty maps $\mathcal{U}$ by

$$\mathbf{M}_i' = \mathbf{U}_i \odot (1 - \mathbf{M}_i) + \vartheta \cdot \mathbf{M}_i, \tag{3}$$

where the first term weights the unmasked regions via the visibility uncertainty map: the points other views cannot see should be assigned low weights during optimization. The ϑ controls the constraint degrees of the original masks. Then, we obtain the finer mask maps $\{\mathbf{M}'_i\}_{i=1}^N$ and re-optimize the 3D representation by adding the guidance of mask maps to the objective function in Equation (1):

$$\arg\min_{\mathcal{G}} \lambda_1 \sum_{i=1}^{N} \|(1-\mathbf{M'}_i) \odot (\mathbf{I}_i - \hat{\mathbf{I}}^{p_i})\|_1 + \lambda_2 \sum_{i=1}^{N} \text{D-SSIM}(\mathbf{I}_i, \hat{\mathbf{I}}^{p_i}, 1-\mathbf{M'}_i), \tag{4}$$

where we have $\hat{\mathbf{I}}^{p_i} = \text{Render}(\mathcal{G}, \mathbf{p}_i)$ and $\lambda_1 + \lambda_2 = 1$. Intuitively, the objective function is to ignore the mask and high-uncertainty regions during the optimization. As a result, we get an updated counterpart $\tilde{\mathcal{G}}$. Similar strategies have been also adopted in recent works (Sabour et al., 2024; 2023).

Intuitively, with the visibility uncertainty maps, we can exclude the pixels that other views cannot see to build the 3D representation, which explicitly leverages the complementary visual cues. As the $\mathcal{U}$ shown in Figure 2 (Bottom) , the pixels with high uncertainty denote the corresponding points (e.g., dynamic fishes) visible from only a few views. This is reasonable since the dynamic fishes are at different locations across different views. We also display the updated 3D representation $\tilde{\mathcal{G}}^1$, showing that the dynamic objects and some artifacts are removed.

4.2 VISTA-CL: Visibility-uncertainty-guided Scene Conceptual Learning

VISTA-GI can reconstruct masked objects when complementary visual information is available from alternative viewpoints. However, for masked regions lacking such cues, we need a more sophisticated approach to comprehend the scene holistically and generate plausible new content to fill these gaps. To achieve this, we propose to learn a conceptual representation $\mathbf{s}$ of the scene through textual inversion (Gal et al., 2022; Zhu et al., 2024), which can be formulated as

$$\mathbf{s} = \text{ConceptLearn}(\mathcal{I}, \mathcal{U}, \mathcal{M}), \tag{5}$$

The learned concept $\mathbf{s}$ is a token and encapsulates the scene's essence without the masked objects. We then leverage $\mathbf{s}$ to process the input images, eliminating the masked objects

$$\tilde{\mathbf{I}}_i = \text{ConceptInpaint}(\mathbf{s}, \mathbf{I}_i, \mathcal{U}, \mathcal{M}), \forall \mathbf{I}_i \in \mathcal{I}, \tag{6}$$

Scene conceptual learning. We formulate the scene conceptual learning, *i.e.*, as the personalization text-to-image problem (Ruiz et al., 2023) based on textual inversion (Gal et al., 2022), and we add the guidance of the visibility uncertainty maps in Section 3.2. Specifically, we have a pre-trained text-2-image diffusion model containing an image autoencoder with ϕ and ϕ^{-1} as encoder and decoder, a text encoder φ, and a conditional diffusion model ϵ_θ at latent space. Then, we learn the scene concept $\mathbf{s}$ by optimizing the following objective function

$$\mathbf{s} = \arg\min_{\mathbf{s}^*} \mathbb{E}_{\mathbf{I}_i \in \mathcal{I}, \mathbf{z}=\phi(\mathbf{I}), \mathbf{y}, \epsilon \in \mathcal{N}(0,1), t}(\|(1-\mathbf{M}'_i) \odot (\epsilon_\theta(\mathbf{z}_t, t, \Upsilon(\varphi(\mathbf{y}), \mathbf{s}^*)) - \epsilon)\|_2^2), \tag{7}$$

where $\mathbf{y}$ is a fixed text (*i.e.*, 'a photo of S^*') and the function $\Upsilon(\Gamma(\mathbf{y}), \mathbf{s}^*)$ is to replace the token of 'S^*' within $\Gamma(\mathbf{y})$ with $\mathbf{s}^*$. The tensor $\mathbf{M}'_i$ is calculated via Equation (3) based on the visibility uncertainty map and the given mask map. Intuitively, we use the Equation (7) to force the learned concept to mainly contain the unmasked scene regions. To validate the learned concept, we can feed 'a photo of S^*' to the T2I diffusion model to generate images about the learned concept. As shown in Figure 2, the images in the lower right are created directly by the T2I diffusion model and illustrate a concept similar to the original scene without any dynamic objects.

Scene conceptual-guided inpainting. We use the learned concept $\mathbf{s}$ to inpaint all input images through the pre-trained T2I diffusion model. Given one image $\mathbf{I}$ from $\mathcal{I}$, we can extract its latent code by $\mathbf{z} = \phi(\mathbf{I})$. Then, we perform the forward diffusion process by iteratively adding Gaussian noise to the $\mathbf{z}$ over T timesteps, obtaining a sequence of noisy latent codes, *i.e.*, $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_T$, where $\mathbf{z}_0 = \mathbf{z}$. At the tth step, the latent is obtained by

$$q(\mathbf{z}_t|\mathbf{z}_0) = \sqrt{\overline{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t, \ \epsilon_t \sim \mathcal{N}(0, \mathbb{I}), \tag{8}$$

where $\overline{\alpha}_t = \prod_{\tau=1}^{t}(1-\beta_\tau)$. $\mathcal{N}(0,\mathbb{I})$ represents the standard Gaussian distribution. As we set the time step as T, the complete forward process can be expressed as $\mathbf{z}_T \sim \text{q}(\mathbf{z}_{1:T}|\mathbf{z}_0) = \prod_{t=1}^{T} \text{q}(\mathbf{z}_t|\mathbf{z}_{t-1})$.

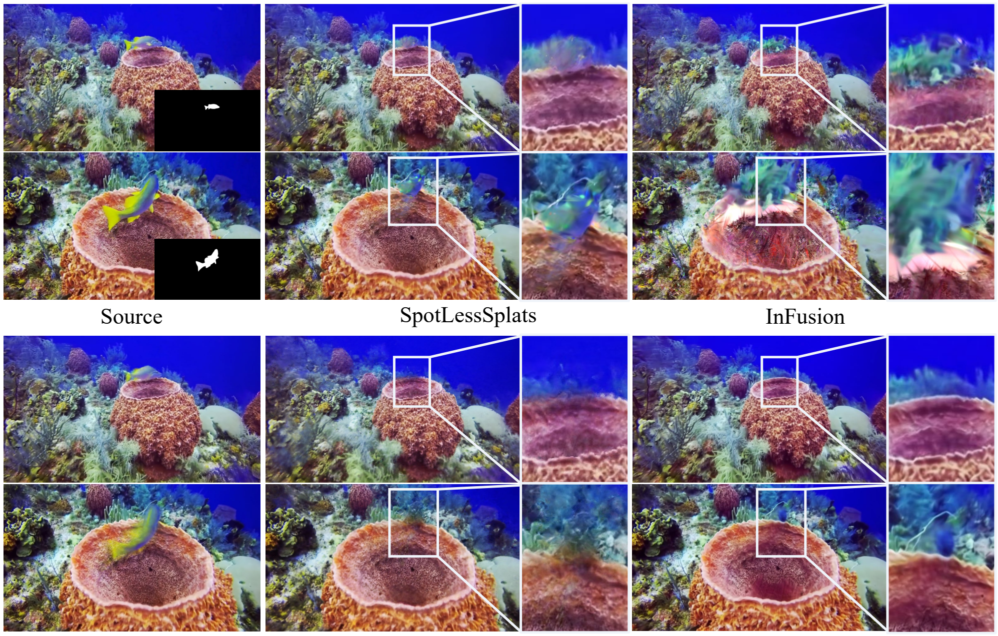


Figure 3: Example of dynamic inpainting on the Underwater 3D Inpainting Dataset.

At the reverse denoising process, we follow the strategy of RePaint (Lugmayr et al., 2022) but embed the guidance of visibility uncertainty maps and the learned concept $\mathbf{s}$. Intuitively, at the time step $t > 1$ during denoising, we only denoise the masked regions conditioned on the scene concept $\mathbf{s}$ while maintaining the unmasked regions with the same content in Equation (8), that is, we have

$$\tilde{\mathbf{z}}_{t-1} = (1 - \mathbf{m}') \odot \mathbf{z}_{t-1} + \mathbf{m}' \odot \hat{\mathbf{z}}_{t-1}, \tag{9}$$

where $\mathbf{z}_{t-1} \sim q(\mathbf{z}_t | \mathbf{z}_0)$ and $\mathbf{m}'$ is the downsampled $\mathbf{M}' \in \{\mathbf{M}'_i\}_{i=1}^N$ calculated by Equation (3) and has the exact resolution as the latent code $\mathbf{z}_{t-1}$. $\hat{\mathbf{z}}_{t-1}$ is denoised from the $\tilde{\mathbf{z}}_t$ with the guidance of the learned concept $\mathbf{s}$, that is,

$$\hat{\mathbf{z}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\tilde{\mathbf{z}}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\tilde{\mathbf{z}}_t, t, \mathbf{s})) + \sigma_t \xi, \text{s.t.}, \xi \sim \mathcal{N}(0, \mathbb{I}), \tag{10}$$

If $t = 1$, $\tilde{\mathbf{z}}_0 = (1 - \mathbf{m}') \odot \mathbf{z} + \mathbf{m}' \cdot \hat{\mathbf{z}}_0$. Then, we can get the inpainted image via decoder $\tilde{\mathbf{I}} = \phi^{-1}(\tilde{\mathbf{z}}_0)$. We can use the above ConceptInpaint to process each image within $\mathcal{I}$ and get a new image set $\tilde{\mathcal{I}}$.

4.3 VISTA: Combining VISTA-GI and VISTA-CL

Given the input images $\mathcal{I}$ and their corresponding mask maps $\mathcal{M}$, VISTA-GI generates visibility uncertainty maps $\mathcal{U}$ as the visual cues and refines the 3DGS representation $\tilde{\mathcal{G}}$. VISTA-CL takes $\mathcal{I}, \mathcal{U}$, and $\mathcal{M}$ as inputs and produces processed input images $\tilde{\mathcal{I}}$ as the semantic cues. Intuitively, we can combine the raw images $\mathcal{I}$ and $\tilde{\mathcal{I}}$, feed them back into VISTA-GI, where $\tilde{\mathcal{I}}$ serve as better views. This allows for an iterative process between VISTA-GI and VISTA-CL. We denote the k-th iteration's 3D representation from VISTA-GI as $\tilde{\mathcal{G}}^k$ and the processed images from VISTA-CL as $\tilde{\mathcal{I}}^k$.

In practice, three iterations are typically sufficient to achieve smooth convergence of the training metrics. The hyperparameter ϑ is initialized by 0 and increases by 0.1 with each iteration. We show an example in Figure 2. The synthetic views $\tilde{\mathcal{G}}^1$, $\tilde{\mathcal{G}}^2$, and $\tilde{\mathcal{G}}^3$ gradually contain fewer distractors, and the results of the final iteration $\tilde{\mathcal{G}}^3$ demonstrate clean and clear views, which means better 3D inpainting under the guidance of the visual and semantic cues.

4.4 VISTA for Dynamic Distractor Removal

VISTA could be easily extended to remove dynamic distractors across multi-view images $\mathcal{I}$ by identifying the dynamic regions in $\mathcal{I}$ and obtaining the mask maps $\mathcal{M}$. In our implementation, we use the tracking method and MASA (Li et al., 2024) to automatically get the mask maps for dynamic objects in the scene. MASA is an open-vocabulary video detection and segmentation model introducing coarse pixel-level information to our method. This plays a similar role as DEVA (Cheng et al., 2023) used in Gaussian Grouping (Ye et al., 2024). However, the masks used in Gaussian Grouping are limited to static objects, while we mask static and dynamic objects that need to be inpainted. For dynamic objects, the uncertainty map can complement the coarse mask that excludes those dynamic distractors from the reconstruction. As shown in Figure 2, the synthetic view $\mathcal{G}$ obtained without masks fairly removes those fish moving greatly but ignores those objects without significant movement. The semantic information in the coarse masks $\mathcal{M}$ identifies these distractors, which the uncertainty map $\mathcal{U}$ cannot detect, and then these distractors can be eliminated by VISTA-CL. As a result, VISTA can remove both static and dynamic distractors in the scene by combining these two mask maps in Equation (3).

5 Experiments

5.1 Datasets and Metrics

To evaluate our method, we conduct experiments on the SPIn-NeRF Dataset for 3D inpainting in general scenes and the Underwater 3D Inpainting Dataset for scene repairing in challenging scenes. More details can be found in Appendix A.

Underwater 3D inpainting dataset. This dataset is derived from the underwater object tracking dataset UTB180 (Alawode et al., 2022), from which we selected multiple videos for resampling, ultimately forming 10 underwater scene datasets. We resample the video in certain FPS to fulfill the motion requirements of initial reconstruction. Each scene contains dozens of images from various viewpoints, and the initial Structure from Motion point cloud and camera intrinsics are obtained via COLMAP (Schonberger & Frahm, 2016). Each viewpoint image undergoes object detection using the open-source method MASA (Li et al., 2024) to obtain rough object masks.

Method	UCIQE ↑	URanker ↑	CLIP Score ↑
SPIn-NeRF	0.49	1.59	0.70
InFusion	0.50	1.52	0.71
SpotLess	0.50	1.59	0.70
Ours	**0.51**	**1.64**	**0.72**

Table 1: Quantitative results of dynamic inpainting on the Underwater 3D Inpainting Dataset.

SPIn-NeRF dataset. The SPIn-NeRF dataset was proposed in Mirzaei et al. (2023). It contains 10 general 3D inpainting scenes, divided into 3 indoor and 7 outdoor scenes. Each scene includes 100 images from various viewpoints, along with corresponding masks. In these datasets, the ratio of the training set to the testing set is 6 to 4. We compare our method with other approaches using the provided camera intrinsics and initialized SfM point cloud.

Metrics. Following SPIn-NeRF, we evaluate the experimental results in two quantitative terms: one for static scenes with ground truth using PSNR, SSIM, LPIPS, and Fid for Reference-based IQA (Image Quality Assessment), and the other for dynamic scenes without ground truth using UCIQE (Yang & Sowmya, 2015), URanker (Guo et al., 2023) and CLIP Score (Hessel et al., 2021) for the underwater Non-Reference IQA. Following the typical comparison methods mentioned in SPIn-NeRF (Mirzaei et al., 2023) and RefFusion (Mirzaei et al., 2024), LPIPS, and Fid are calculated around the masked region by considering the bounding box of the mask. UCIOE is a generally used underwater metric that utilizes a linear combination of chroma, saturation, and contrast for quantitative assessment, quantifying uneven color casts, blurriness, and low contrast. URanker is a transformer-based metric to assess the quality of underwater images. Meanwhile, the CLIP Score measures the relation between image and text. As a result, we serve ‘An underwater scene without fish’ as the caption to evaluate the effects of fish removal.

Method	LPIPS ↓	Fid ↓	PSNR ↑	SSIM ↑
Masked Gaussians	0.594	278.32	10.77	0.29
SPIn-NeRF	0.465	156.64	15.80	0.46
Gaussians Grouping	0.454	123.48	14.86	0.27
InFusion	0.567	118.26	15.59	0.53
Ours	**0.418**	**113.58**	**16.48**	**0.59**

Table 2: Quantitative results of static inpainting on the SPIn-NeRF Dataset.

5.2 Experimental Results

We compare our method with several state-of-the-art open-source 3D inpainting methods, such as Infusion (Liu et al., 2024), SPIn-NeRF (Mirzaei et al., 2023), Gaussian Grouping (Ye et al., 2024),

Source Gaussian Grouping InFusion Ours

Figure 4: Example visualization of static inpainting on the SPIn-NeRF Dataset.

and SpotLessSplats (Sabour et al., 2024). SpotLessSplats is only designed for scenarios with dynamic distractors, while others are the latest static inpainting methods. Infusion is retrained using its publicized code (Liu et al., 2024).

Results on underwater 3D inpainting dataset. We compare our method with baseline methods on the underwater dataset. Figure 3 illustrates the performance of various inpainting methods on our dataset, especially for dynamic objects. Some perspectives can compensate for some areas that need repair, while others require direct inpainting from the algorithm. This scene represents a scenario that can effectively reflect real-world inpainting task datasets. The figure shows that our method presents the most stable and consistent inpainting scene without artifacts and blurriness. SpotLessSplats removes part of these Gaussians representing the moving fish but fails to repair the missing area hiding behind the fish. The results of Infusion are obtained from a single inpainted reference image, which distorts the images in other viewpoints, although the views rendered near the reference image are relatively clear. Additionally, the results of SPIn-NeRF show 3D consistency, but some synthetic images exhibit artifacts and blurriness in certain viewpoints. Table 1 shows the quantitative metrics of the image quality after inpainting. For the UCIQE and URanker, our method outstrips other methods by utilizing the uncertainty map to reduce the weight of blurry areas caused by underwater floating objects during reconstruction. Besides, the CLIP Score of our method outperforms other methods for better removal of the target objects.

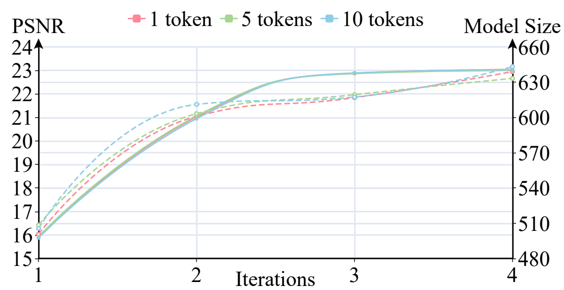


Figure 5: Relationship between model performance (PSNR) and model size (MB) with different token numbers. The dashed and solid lines represent the model size and performance variations respectively. The model performance (solid lines) under different token numbers almost overlaps.

Results on SPIn-NeRF dataset. Figure 4 depicts an example scene from the SPIn-NeRF Dataset masking a stationary box that requires inpainting. The results of Gaussian Grouping are fairly realistic at the 2D image level, but there are significant inconsistencies between perspectives, such as distortion at the edges of stairs. The results of InFusion appear more realistic from one certain perspective. Still, its approach of optimizing one single view compromises the performance of other perspectives, leading to unpredictable artifacts in those views. Our method benefits from an iterative progressive optimization approach, ensuring consistency across perspectives through multiple inpainting and reconstruction, resulting in more stable outcomes.

Method	UCIQE ↑	URanker ↑	CLIP Score ↑
Ours w/o VISTA-GI	0.48	1.52	0.70
Ours w/o VISTA-CL	0.50	1.59	0.69
Ours	**0.51**	**1.64**	**0.72**

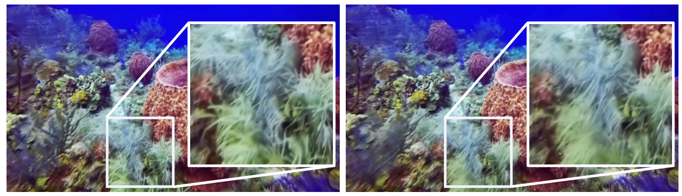
Table 3: Quantitative ablation study of VISTA-GI and VISTA-CL on the Underwater 3D Inpainting Dataset.

Ablation study on VISTA-GI and VISTA-CL. We conducted ablation experiments on the underwater 3D inpainting dataset by removing the VISTA-GI and VISTA-CL from our final version, respectively. The specific results are shown in Table 3. Our experiments demonstrate two key findings. First, attempting reconstruction using only a 2D generative model without VISTA-GI leads to significantly degraded image quality metrics. This validates that VISTA-GI's uncertainty guidance effectively mitigates multi-view inconsistencies during 3D reconstruction, resulting in higher-quality outputs. Second, while omitting VISTA-CL maintains image quality comparable to existing methods like SplotLess and SPIn-NeRF, the lack of concept-guided learning significantly reduces CLIP-Score metrics. This indicates that without conceptual constraints, the inpainting process produces results that are visually plausible but semantically inconsistent with the scene context.

5.3 DISCUSSIONS

This section will combine experimental results to discuss the reasons behind some hyperparameter settings in Section 4, further demonstrating our approach. More details can be found in Appendix A.

Effects of token numbers to depict one scene. The quantity of tokens needed to describe a scene is important since each count corresponds to a specific color. The Figure 5 shows how the inpainting results change as the descriptive tokens change. Too many tokens to depict the scene do not increase the model performance and may even increase the model size. The effect of textual inversion is to focus on learning the rough semantic features of the scene rather than the detailed object features, thereby not necessarily requiring very detailed tokens. We also observe that the training PSNR becomes smooth after three iterations, inspiring us to set three iterations. More iterations cause a larger model size, which means excessive Gaussians to fit the noise introduced by the diffusion model.



w/ raw images　　w/o raw images

Figure 6: Reconstruction results with and without raw images. Involving the raw images in our method will improve the inpainting performance.

Reasons for combining raw images $\mathcal{I}$ and $\tilde{\mathcal{I}}$ rather than substituting raw images $\mathcal{I}$ with $\tilde{\mathcal{I}}$ in Section 4.4 As shown in Figure 6, the reconstruction without raw images could not render the seaweed without ambiguity. The accumulated error from two iterations, caused by 3DGS's inability to fit the scene fully and the uncertainty introduced by the generated model, deteriorates the image quality. Raw images act as an "anchor" for our method, ensuring that the rendered images align closely with the input images and do not deviate significantly.

Time cost analysis and comparison. To quantitatively evaluate performance and computational efficiency, we compare our method against baseline approaches (InFusion, SPIn-NeRF, and SpotLess) on the synthetic scene shown in Figure 8. This scene provides ground truth data, enabling evaluation through reference-based metrics for both rendering quality and computational efficiency during optimization. As shown in Table 4, while our method incurs additional computational overhead compared to vanilla 3DGS due to the integration of iterations and diffusion models, it achieves superior rendering quality while maintaining comparable efficiency to state-of-the-art 3DGS methods (e.g., SpotLess (Sabour et al., 2024)). Furthermore, our approach demonstrates significantly better reconstruction quality while being approximately 10× faster than leading NeRF-based methods such as SPIn-NeRF.

Method	LPIPS ↓	PSNR ↑	Time Cost
InFusion	0.23	19.34	16m 34s
SPIn-NeRF	0.15	23.33	7h 32m 18s
SpotLess	0.14	24.75	30m 26s
Ours	0.10	26.38	33m 34s

Table 4: Quantitative results and time costs on the synthesis data in Figure 8.

6 CONCLUSION

In this work, we presented VISTA, a novel framework for 3D Gaussian inpainting that effectively leverages complementary visual and semantic cues from multiple input views. By introducing visibility uncertainty maps and combining visibility-uncertainty-guided 3D Gaussian inpainting (VISTA-GI) with scene conceptual learning (VISTA-CL), our method addresses key challenges in 3D scene editing for static and dynamic scenes. Experimental results on the SPIn-NeRF and UTB180-derived datasets demonstrate VISTA's superior performance over state-of-the-art techniques in generating high-quality 3D representations with seamlessly filled masked regions and effectively removing distractors. The versatility of our approach extends to handling complex inpainting scenarios and dynamic distractor removal, making it a powerful tool for various applications in augmented and virtual reality. By simultaneously leveraging geometric and conceptual information, VISTA represents a significant advancement in 3D Gaussian inpainting, bringing us closer to achieving seamless and realistic 3D scene editing and paving the way for more immersive virtual experiences.

References

Basit Alawode, Yuhang Guo, Mehnaz Ummar, Naoufel Werghi, Jorge Dias, Ajmal Mian, and Sajid Javed. Utb180: A high-quality benchmark for underwater tracking. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3326–3342, 2022.

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 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, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19697–19705, 2023.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1316–1326, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Chunle Guo, Ruiqi Wu, Xin Jin, Linghao Han, Weidong Zhang, Zhi Chai, and Chongyi Li. Underwater ranker: Learn which is better and how to be better. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 702–709, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000.

Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18973, 2024.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10758–10768, 2022.

Hao-Kang Liu, I Shen, Bing-Yu Chen, et al. Nerf-in: Free-form nerf inpainting with rgb-d priors. *arXiv preprint arXiv:2206.04901*, 2022.

Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20669–20679, 2023.

Ashkan Mirzaei, Riccardo De Lutio, Seung Wook Kim, David Acuna, Jonathan Kelly, Sanja Fidler, Igor Gilitschenski, and Zan Gojcic. Reffusion: Reference adapted diffusion models for 3d scene inpainting. *arXiv preprint arXiv:2404.10765*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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.

Tijana Ružić and Aleksandra Pižurica. Context-aware patch-based image inpainting using markov random field modeling. *IEEE transactions on image processing*, 24(1):444–456, 2014.

Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20626–20636, 2023.

Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David J. Fleet, and Andrea Tagliasacchi. SpotLessSplats: Ignoring distractors in 3d gaussian splatting. *arXiv:2406.20055*, 2024.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pp. 703–735. Wiley Online Library, 2022.

Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20902–20911, 2024.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

Yuxin Wang, Qianyi Wu, Guofeng Zhang, and Dan Xu. Learning 3d geometry and feature consistent gaussian splatting for object removal. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2025.

Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel J Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16528–16538, 2023.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, June 2024.

Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, 2024.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

Jiayi Zhu, Qing Guo, Felix Juefei-Xu, Yihao Huang, Yang Liu, and Geguang Pu. Cosalpure: Learning concept from group images for robust co-saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3669–3678, 2024.

A APPENDIX

Experiment Setup Our 3D reconstruction and 2D inpainting method is implemented on a single RTX 4090. We use the default parameters of 3DGS for reconstruction, generating a reconstructed render every 10,000 iterations. Additionally, we employed the commonly used stable-diffusion-v1-5 (Rombach et al., 2022) as the base inpainting model, training it for 3,000 iterations (taking approximately 1.5 hours) using textual inversion for scene representation. Our diffusion model inference consists of a 50-step denoising process, initialized with a noise strength of 1.0 that is progressively reduced by a factor of 0.2 at each iteration.

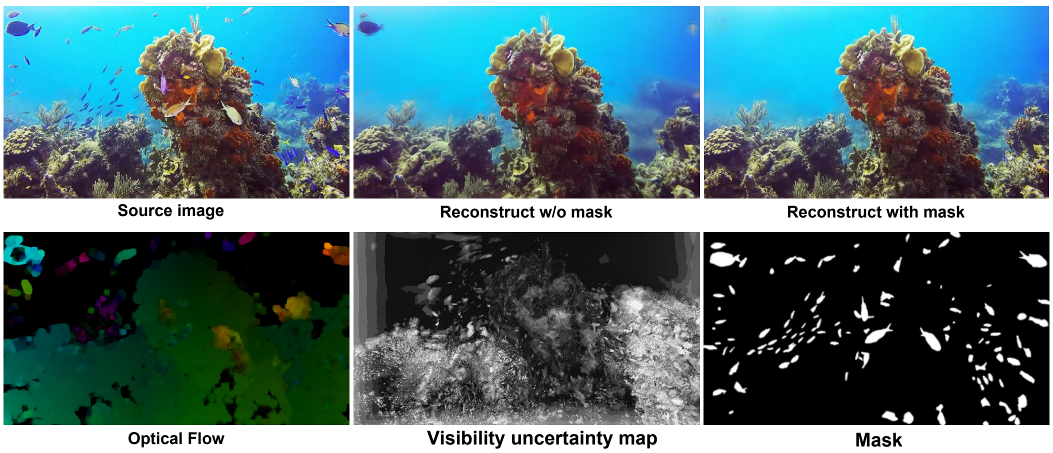


Figure 7: Impact of prior (mask) on inpainting results. Our method will improve inpainting performance by incorporating the mask information. To analyze the results, we also display the optical flow of the source image and the visibility uncertainty map.

A.1 IMPACT OF PRIOR (MASK) ON INPAINTING

Adding the prior (mask) information in our method will significantly improve the inpainting results, especially for those static objects. This is easy to understand because dynamic objects create inconsistencies during the reconstruction process, which our algorithm can detect. In contrast, static object inpainting necessitates the semantic information the detection model identifies.

For instance, in the top-left corner of Figure 7, the fish is retained while the others are removed. This is primarily because the fish remains stationary across different views (as evident in the optical flow map of Figure 7, where the top-left fish exhibits low flow values at its center). Consequently, it has a lower value in the visibility uncertainty map (see the corresponding map in Figure 7). Without using a mask to label this area for repair manually, the fish's geometric characteristics resemble those of a stationary object, such as a rock, making it indistinguishable from our uncertainty detection system.

In contrast, moving fish create significant geometric inconsistencies across viewpoints, enabling our uncertainty detection to flag them as anomalies. This leads to their removal through the inpainting process. To address these challenging scenarios, we introduced mask annotations for fish detection, providing semantic guidance for our inpainting method. As shown in the last column of Figure 7, incorporating the mask ensures the successful removal of the top-left fish.

VISTA in limited scenarios Our uncertainty maps are built by observing a set of adjacent perspectives/views, thus fully utilizing complementary visual cues. However, in some extreme conditions, we don't have enough valid adjacent perspectives/views to get the visual cues. To investigate the performance of our method under such extreme conditions, we manually synthesized an extreme scenario where the camera rapidly changes poses, resulting in very few available adjacent viewpoints. In this case, the VISTA-GI can hardly detect the inconsistency between different views, requiring VISTA-CL to produce better results.

As shown in Figure 8 (a), thanks to the 2D diffusion model, our method utilizes its results to effectively inpaint the scene in such extreme conditions. Meanwhile, as shown in Figure 8 (b), The InFusion

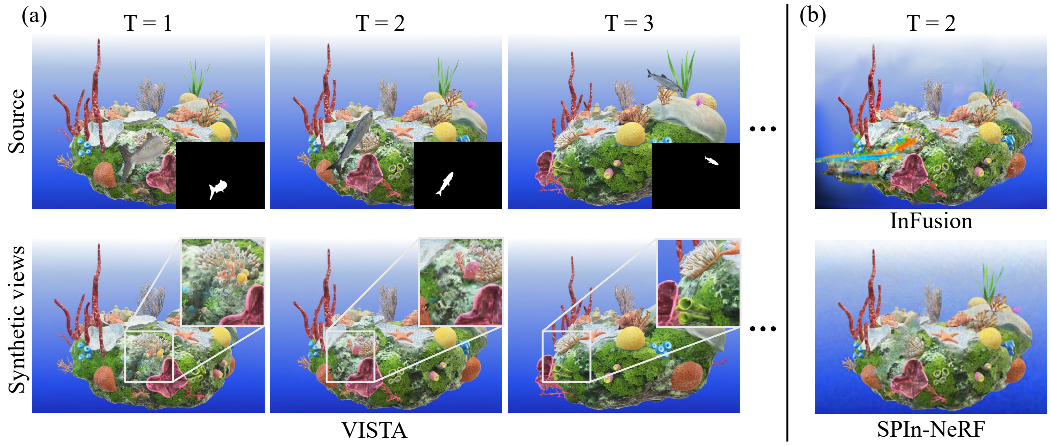


Figure 8: (a) The figure of an artificial synthesis scene in extreme cases. The original views of three adjacent cameras and the inpainting results of our method are demonstrated for comparison. **(b) The results of InFusion and SPIn-NeRF in extreme cases.** Their results are obtained by the camera from 'T = 2' in (a).

result is unrealistic due to neglecting consistency in inpainting. SPIn-NeRF shows a reasonable result but with blurry and indistinguishable inpainting areas. Compared to other methods, our approach benefits from Scene Conceptual Learning, resulting in clearer and more reasonable repairs in the target areas, and the textures and content maintain consistency with the original scene.

A.2 VISUALIZATION OF OUR METHOD

In this part, we visualize more results to demonstrate the effectiveness of our method and the potential failure scenarios that may arise.

Figure 9: Case of real-world pedestrian removal from the nerf-on-the-go dataset.

Real-world case. The underwater dataset we used is derived from real-world diving videos, and due to the effects of the underwater medium and floating debris, these scenes are challenging scenarios in the real world. We also tested our dataset on a scene related to pedestrian removal from the nerf-on-the-go dataset. This scene, called Tree, contains 212 images, with the main distractors from moving pedestrians. As shown in Figure 9, our method achieved high-quality results on this dataset. Due to the abundance of viewpoints in the dataset, there is a lot of supplementary information between perspectives, allowing our method to effectively utilize other viewpoints to repair the blurring caused by moving distractors.

Fail case from our dataset. We provide failure cases of our algorithm in Figure 10. Due to errors in the prior mask, some fish were not detected by the object detection model. Furthermore, since the fish did not move significantly during the shooting process, these areas did not produce inconsistencies across multiple viewpoints during reconstruction, making it difficult for the VISTA-GI component to identify these areas through uncertainty. This also validates our algorithm design approach:

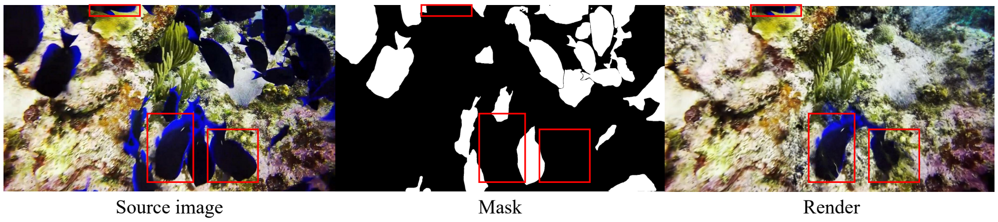


Figure 10: Failure case from our underwater 3D inpainting dataset.

VISTA-CL introduces semantic information through masks, while VISTA-GI incorporates geometric information through uncertainty, complementing each other to remove distractors. However, in this failed case, issues arose in both aspects, resulting in poor reconstruction quality of the final scene.

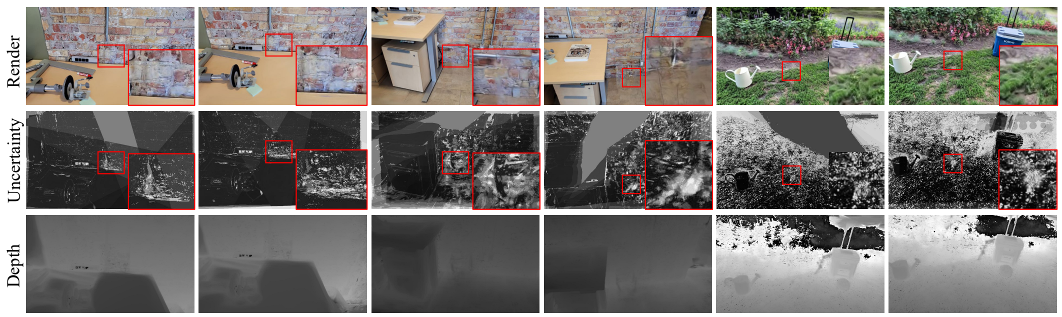


Figure 11: Visualization of the uncertainty map and depth of static scenes.

Uncertainty and depth maps of static scenes. As shown in Figure 11, we further visualize the uncertainty and depth maps of the static scenes. The deeper the color, the closer the depth. It can be observed that our method identifies areas in the rendered image that are inconsistent with other viewpoints and generates reasonable contents.

A.3 Impacts of Hyper-parameters

In this part, we study the influence of the hyper-parameter ϑ in Eq. (3), the initial noise & iterations of diffusion inference, and the threshold of uncertainty map.

Impact of noise reduction ratios in diffusion inference. During diffusion model inference, we investigate how different noise reduction strategies affect reconstruction quality. Starting from an initial noise strength of 1.0, we systematically decrease the noise at each iteration by a fixed ratio. We evaluate four different reduction ratios $\{0.1, 0.2, 0.3, 0.4\}$ and analyze their impact on reconstruction quality across iterations using our dataset. As shown in Figure 12 (a), while all ratios lead to improved PSNR values over iterations, the reduction ratio of 0.2 achieves optimal convergence in the fewest iterations. Based on this empirical analysis, we adopt 0.2 as the noise reduction ratio in our method.

Impacts of ϑ in Eq. (3). We use ϑ to control the prior constraint of the original masks. We investigate how different ϑ increasing strategies affect reconstruction quality. The hyperparameter ϑ is initialized by 0 and increases by 0.1 with each iteration in our paper. We evaluate five different increase ratios $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and analyze their impact on reconstruction quality across iterations using our dataset. As shown in Figure 12 (b), all ratios lead to improved PSNR values over iterations. In the first two iterations, a higher increase ratio improves the reconstruction performance. However, an increase ratio above 0.1 indicates that the algorithm becomes overly confident in the inpainting areas too early, resulting in insufficient interaction of geometric and semantic information between the VISTA-GI and VISTA-CL modules, which subsequently leads to a decline in reconstruction performance in later iterations.

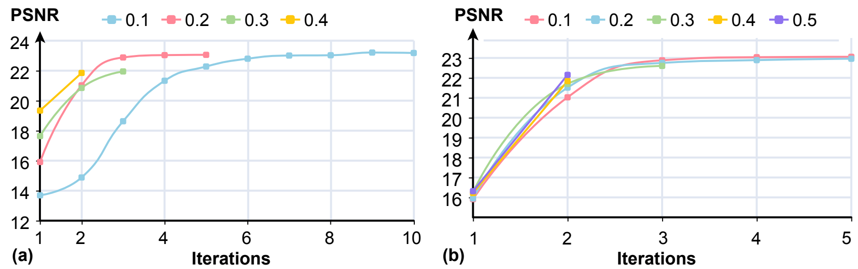


Figure 12: (a) Relationship between 3DGS rendering quality (PSNR) and noise reduction ratio of diffusion inference. (b) Relationship between 3DGS rendering quality (PSNR) and increasing ratio of ϑ in Eq. (3).

Resolution	LPIPS ↓	PSNR ↑	SSIM ↑
64×64	0.51	16.27	0.68
128×128	0.42	18.89	0.69
256×256	0.26	21.33	0.71
512×512	0.11	26.04	0.84
1299×974	0.10	26.38	0.86

Table 5: Quantitative ablation results of different resolutions.

A.4 Impacts of Different Image Resolution

In our experiment setup, we use the stable diffusion v1.5 as the inpainting model and train and test the model following its default setup: if the input image has a resolution higher than 512×512, we crop the image to a new size that is both the closest to the original image size and a multiple of 8; if the input image is smaller than 512×512, we rescale the image to 512×512. To analyze the influence of the strategy on different original resolutions, given an original scene with input images having a size of 1299×974, we downsample these images to four resolutions: 64×64, 128×128, 256×256, and 512×512. Then, for each resolution, we can build a 3D model and evaluate the rendering quality. As shown in Table 5, we observe that: (1) reducing the resolution to 512×512 does not significantly impact any of the metrics, demonstrating our method's robustness to substantial resolution changes. (2) further decreasing the resolution leads to gradual degradation in reference-based metrics, while non-reference metrics remain relatively stable.

A.5 Quantitative Analysis of Large Viewpoint Differences

Ablation study of large viewpoint differences. To evaluate the impact of variants of viewpoint difference, we first capture 34 images from continuously distributed viewpoints around a scene to create a ground truth (GT) 3DGS model. We then systematically reduce the number of viewpoints by sampling them at different intervals $\{2, 3, 4, 5, 6, 7\}$, where larger intervals represent larger viewpoint differences. For each sampling interval, we construct a new 3DGS model and assess its quality by comparing its rendered images against those from the GT model using standard metrics: LPIPS, SSIM, and PSNR. This methodology allows us to analyze how viewpoint difference affects reconstruction quality quantitatively.

Sampling interval	LPIPS ↓	PSNR ↑	SSIM ↑
2	0.09	26.25	0.89
3	0.14	23.42	0.83
4	0.16	22.42	0.80
5	0.27	18.09	0.65
6	0.25	18.71	0.69
7	0.41	15.66	0.57

Table 6: Quantitative results of large viewpoint differences.

Considering that the reduction in available viewpoints for the training leads to decreased 3DGS reconstruction quality, our method still achieves good results even with significant viewpoint variation. This validates that our approach can detect inconsistencies between viewpoints and repair those areas despite the large viewpoint differences. However, in extreme cases, the absence of key viewpoints results in a loss of critical complementary information between viewpoints, leading to a significant decline in the reconstruction metrics of the scene.

Comparisons of different methods in extreme case. To validate our advantages in the extreme case with large viewpoint differences, we conducted a quantitative evaluation of various methods for the extreme case mentioned in Figure 8, and the results are as the following table. It can be seen that our method still outperforms existing methods in removing dynamic distractors under such extreme conditions.

Method	LPIPS ↓	PSNR ↑	SSIM ↑
InFusion	0.23	19.34	0.78
SPIn-NeRF	0.15	23.33	0.82
SpotLess	0.14	24.75	0.84
Ours	**0.10**	**26.38**	**0.86**

Table 7: Quantitative comparison of different methods in extreme case.