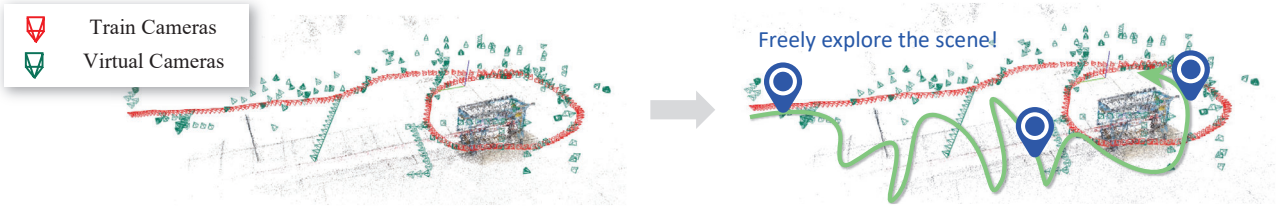
This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ExploreGS: Explorable 3D Scene Reconstruction with Virtual Camera Samplings and Diffusion Priors

Minsu Kim Subin Jeon In Cho Mijin Yoo Seon Joo Kim
Yonsei University
{minsu1206, subinjeon, join, mijin6690, seonjookim}@yonsei.ac.kr



(a) Exploration Results



(b) Virtual Camera Sampling with Diffusion Prior (c) Scene Reconstruction

Figure 1. (a) Explorable 3D scene reconstruction results. Our method renders photorealistic images from any arbitrary viewpoints. (b) Virtual cameras are sampled in regions with missing information, and pseudo-ground truth is generated using diffusion priors. (c) After optimizing 3D Gaussians with virtual viewpoint samples, our method enables real-time scene exploration.

Abstract

Recent advances in novel view synthesis (NVS) have enabled real-time rendering with 3D Gaussian Splatting (3DGS). However, existing methods struggle with artifacts and missing regions when rendering from viewpoints that deviate from the training trajectory, limiting seamless scene exploration. To address this, we propose a 3DGS-based pipeline that generates additional training views to enhance reconstruction. We introduce an information-gain-driven virtual camera placement strategy to maximize scene coverage, followed by video diffusion priors to refine rendered results. Fine-tuning 3D Gaussians with these enhanced views significantly improves reconstruction quality. To evaluate our method, we present Wild-Explore, a benchmark designed for challenging scene exploration. Experiments demonstrate that our approach outperforms existing 3DGS-based methods, enabling high-quality, artifact-free rendering from arbitrary viewpoints.

1. Introduction

The field of novel view synthesis (NVS) has seen significant advancements, enabling photorealistic rendering of scenes from previously unseen viewpoints. Much of this progress is attributed to NeRF [24] and its variants [1, 25]. This advancement has been further accelerated by recent 3D Gaussian Splatting (3DGS) [11], which enables high-quality rendering in real-time. The real-time capability of 3DGS broadens the applications of NVS in domains such as AR, VR, and gaming, leading users to expect the experience of freely exploring reconstructed scenes.

Unfortunately, such an experience is yet to be fully realized, as existing methods suffer from severe degradations in rendering quality when viewpoints deviate significantly from the input observations. These degradations include holes, oversized Gaussians, spiky surface artifacts, and spurious floaters, often arising from occlusions or limited viewpoint coverage (see Fig. 1 (a)). This limitation stems from missing information, since optimization-based approaches cannot synthesize contents beyond the observed data.

We are interested in 3D scene reconstruction that can render more vivid and photorealistic images from arbitrary viewpoints, so that users can *freely explore the reconstructed scenes.* We term this task as "explorable scene reconstruction", which aims to reconstruct complete scenes within the target areas. To avoid generating arbitrary contents in regions far beyond the input observations–which would not reflect actual user experiences–we define the target areas as a bounding box defined by the input views.

One straightforward solution is to take thousands of images to completely cover the target scenes. However, collecting such dense data is time-consuming and labor-intensive, making it impractical for many real-world applications. Thus, we aim to reconstruct complete 3D scenes from input images that only partially cover the target scenes, which better reflects practical capturing scenarios.

In this paper, we introduce ExploreGS, a pipeline that enables explorable scene reconstruction using diffusion priors and 3DGS. ExploreGS leverages video diffusion priors to generate additional training views that complement missing information. These synthesized images at virtual viewpoints serve as pseudo ground truths to refine the initial, incomplete 3D Gaussians, thereby enabling complete, photorealistic rendering at arbitrary viewpoints.

The key challenges of explorable scene reconstruction lie in determining the optimal placement of virtual viewpoints. Unlike previous works that assume evaluation trajectories are known [19, 33], virtual viewpoints in our pipeline need to be sampled to best complement missing information in poorly covered areas. We propose a novel view sampling strategy that strategically positions virtual cameras based on information gain. This approach identifies viewpoints that offer the most novel scene details from the initial 3DGS reconstruction as shown in Fig. 1 (b).

To provide pseudo observations, we generate realistic images along sampled viewpoints. We introduce video diffusion model which converts renderings with degradations into clean images. Although it is motivated by 3DGS-Enhancer [21], we modify the training objectives to fit scene exploration and prepare larger dataset for training. We then finetune 3DGS on both virtual viewpoints and training viewpoints with uncertainty guidance.

To validate our pipeline, we introduce Wild-Explore, a dataset designed to evaluate challenging exploration scenarios with diverse and distant viewpoints. We also assess our method on the Nerfbusters dataset [40], curated to better reflect scene exploration tasks. Experimental results show that our approach achieves the highest scores among 3DGS-based methods. Finally, an interactive exploration demo demonstrates that our method enables broader scene coverage with fewer artifacts compared to existing approaches.

In summary, our contributions can be organized as follows:

- We propose a pipeline for explorable 3D scene reconstruction, which incorporates the real-time rendering of 3DGS, video diffusion priors to complete missing regions, and an uncertainty-based finetuning strategy.
- We propose a novel virtual camera sampling strategy to identify viewpoints that can maximize additional information gain, thereby effectively reducing artifacts and missing regions.
- We introduce Wild-Explore, a new benchmark dataset designed to evaluate challenging exploration scenarios.

2. Related work

2.1. Novel view synthesis

Novel view synthesis aims to generate images of a scene from unseen viewpoints. NeRF [24] and its following works [1, 2] have revolutionized novel view synthesis by representing 3D scenes as continuous volumetric functions, significantly improving rendering quality yet still suffering from slow rendering speed. Recently, 3D Gaussian Splatting [11] has enabled real-time rendering through an explicit Gaussian ellipsoid representation and its differentiable splatting pipeline. Despite these advances, all these methods produce degraded results when the number of images or their view coverage are limited. To address these shortcomings, several works [15, 26, 35, 43] incorporate geometric constraints and depth regularization, leading to less artifacts and improving reconstruction fidelity under sparse-view conditions. However, they remain insufficient for reconstructing explorable scenes, as they lack the ability to fill missing information in the training views.

2.2. Evaluating novel view synthesis

Previous novel view synthesis methods primarily focus on view interpolation, with LLFF [23], Blender synthetic dataset [24], TanksTemples [12] and MipNeRF360 [1] being widely used benchmarks. These datasets place test viewpoints between training viewpoints and thus are not suitable for evaluating performance under large viewpoint changes. To address this issue, Nerfbusters [40] introduces a new dataset with large viewpoint variations by capturing scenes from separate camera paths. Nevertheless, most of the test views in Nerfbusters lie behind the training views, which are largely unconstrained and primarily require gen-

erating arbitrary contents rather than completing the under-reconstructed target scenes.

2.3. Diffusion prior for 3D reconstruction

Since optimization-based NVS methods cannot generate novel contents, recent works [20, 28, 30, 37, 38] have explored methods for exploiting generative diffusion priors. The vast majority of recent research [4–8, 13, 17, 21, 32, 36, 39] train multi-view diffusion models or fine-tune video diffusion models on large multi-view datasets [16, 29, 47] for generating 3D contents. Among these, 3DGS-Enhancer [21] fine-tunes a video diffusion model to enhance a sequence of 3D Gaussian splatting renderings to solve sparse-view reconstruction task. On the one hand, some works [19, 33] apply diffusion prior to address view extrapolation. ExtraNeRF [33] introduces scene-level finetuned image inpainting diffusion and depth-completion model. ViewExtrapolator [19] leverages SVD [3] in a training-free manner. While these methods are most closely related to our work, they assume known evaluation viewpoints and require running the diffusion model for every target view to render. Moreover, their diffusion models lack multi-view consistency, leaving their ability of reconstructing explorable scenes unclear.

2.4. Next-view selection for data acquisition

A number of works studied methods to select the next viewpoint to acquire data when the number of input viewpoints are limited. Early works focus on active learning or information-based strategies for efficient data acquisition [9, 14, 27, 31, 34, 41]. In particular, FisherRF [9] leverages the Fisher information matrix to quantify the information gain or uncertainty of a given viewpoint, serving as a strong baseline for next-best-view selection. However, these methods primarily target single-object or forward-facing scenes, while our method demonstrates superior performance in complex scenes. Recent works extend this line of works to autonomous exploration [10, 18, 22, 45]. Instead of capturing additional inputs, ExploreGS samples virtual viewpoints from initially reconstructed scenes and exploits generative priors to synthesize images from that viewpoints.

3. Method

3.1. Overview

The overall pipeline of ExploreGS is illustrated in Fig. 2. Our pipeline follows three stages: (1) scene initialization, (2) pseudo observation generation with virtual camera sampling and a video diffusion prior, and (3) fine-tuning 3DGS using both the training and the generated observations sets.

In the first stage, we begin by optimizing 3D Gaussians, denoted as $\mathcal{G}$, to represent the 3D scene using the given training viewpoints $\mathcal{V}^T = \{V_i^T\}_{i=1}^{N_T}$ where N_T denotes the number of training viewpoints. Here, each V_i^T has known camera pose and ground truth image I_i^T. Then, we determine the boundary of reconstructable scene based on the input observations and identify occupied regions for virtual viewpoints samplings.

In the second stage, we sample virtual camera trajectories based on information gain of each candiate virtual viewpoint, as shown in Sec. 3.3. N_{Tr} camera trajectories that maximizes the information gain are sampled, and each trajectory Tr_n consists of L progressively shifting virtual viewpoints: $Tr_n = \{V_l^G\}_{l=1}^{L}$. Along these virtual trajectories, we render images to obtain I_i^G and generate pseudo-observations I_i^P from I_i^G by enhancing them using the generative prior of a video diffusion model.

Finally, we reconstruct complete and explorable 3D scenes by fine-tuning the initially optimized 3D Gaussians using both the training and virtual viewpoints, as described in Sec. 3.5. To mitigate the impact of inconsistencies between the generated and the original images during fine-tuning, we propose two techniques: (1) *image-level confidence score*, and (2) *pixel-level confidence map*.

3.2. Scene initialization

Initial reconstruction. Our pipeline begins by optimizing the initial set of 3D Gaussians with the given training set within the 3DGS optimization framework [11]. Then, we estimate the target bounding box of the scene, which is a maximum boundary that can be reconstructed based on the given input images. As previously discussed, reconstructing content beyond this bounding box is highly challenging, as it lacks grounding in the input observations and increasingly resembles unconstrained content generation, which is beyond the scope of this work. This bounding box is computed by using the mesh extracted from the initial 3D Gaussians and the given cameras.

Rasterization-based occupancy estimation Since renderings from cameras inside occupied regions (*e.g.* behind walls or inside solid objects such as an apple) offer redundant images, not relevant to the scene, they should be located in free space. To this end, we introduce a simple rasterization-based algorithm to construct the occupancy grid $\mathcal{O} \in \mathbb{R}^{\mathbb{S}\times\mathbb{S}\times\mathbb{S}}$. We estimate the visibility of each 3D Gaussians using the Gaussian rasterizer by computing transmittance values from the training views. Among these, we average the top-3 visibility scores to obtain the final visibility estimate. A vertex is classified as occupied if the estimated visibility is below threshold τ (we use 0.5); otherwise, it is considered free space.

3.3. Virtual view sampling

After initializing the target scene, our method utilizes video diffusion priors to supplement the missing information from

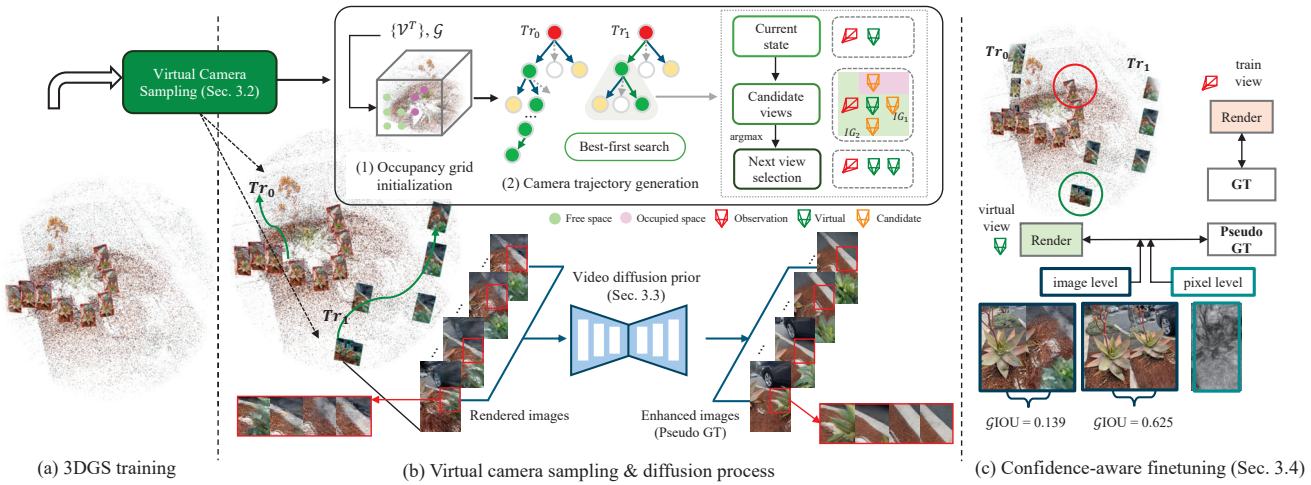


Figure 2. Overview of the proposed framework for scene exploration. (a) The scene is initially optimized using 3DGS on the given training viewpoints. (b) Based on the optimized 3DGS and training viewpoints, we generate virtual camera trajectories, and enhance the rendered views using our diffusion-based enhancement model. (c) Finally, the scene is further optimized using both the original training viewpoints and the newly generated virtual viewpoints.

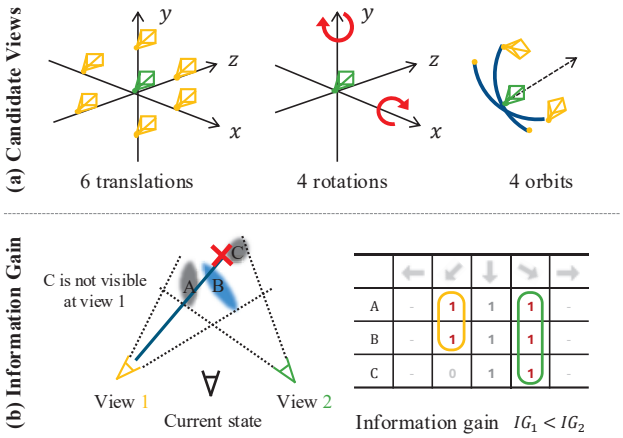


Figure 3. (a) Viewpoint candidates for virtual camera viewpoint generation. (b) Information gain of each viewpoint. Simplified 2D examples of both are presented for clarity.

the input views and removes unwanted artifacts and missing regions. To effectively utilize diffusion priors, our sampling strategy selects virtual camera trajectories that can maximize the information gain of the scene.

As illustrated in Fig. 2, our approach follows a best-first search strategy with depth-first characteristics to construct virtual trajectories. We begin by selecting N_{Tr} training viewpoints at spatially uniform intervals as the initial states for each trajectory and gradually select the next viewpoint. At each step, the current viewpoint expands into K candidate views. The one with the highest information gain becomes the next viewpoint in the trajectory, while the rest are stored for future consideration. Here, the current state refers to the most recently added camera in a trajectory.

Candidate views. We define candidate views as the set of possible camera motion primitives, which determine the transitions from the current state. As shown in Fig. 3, our action space consists of 6 translational moves, 4 pure rotations of the camera, and 4 rotations where the camera orbits around the look-at point, totaling 14 actions. This choice ensure the virtual trajectory efficiently exploring the scene when compared to previous works just using going straight or turning around.

A candidate view is removed if it (1) lies outside the bounding box or in an occupied region, or (2) is close to the 3D Gaussians before fine-tuning, since rendering can be severely disrupted when the camera is placed near the occupied. We utilize the target scene bounding box and the occupancy grid from Sec. 3.2 to achieve this. These constraints prevent the trajectory from diverging away from the reconstructed scene or producing redundant pseudo-observations.

Information gain. We aim to observe 3D Gaussians densely from diverse viewpoints, as sparse or limited-angle observations often lead to inaccurate shape and color estimation. To this end, we measure information gain based on view coverage. We employ a 2D map $\mathcal{M} \in \mathbb{R}^{N_\mathcal{G} \times D}$, where $N_\mathcal{G}$ is the total number of Gaussians and D is the number of discretized viewing directions around 3D Gaussians. The map $\mathcal{M}$ stores binary values where each entry $\mathcal{M}[j,k] = 1$ indicates that the j-th Gaussian has been observed from the k-th viewing direction and initialized with training viewpoints $\mathcal{V}^T$. At each step t, we update the map as follows:

$$\mathcal{M}_t[j,k] = 1, \quad \forall j \in J, \tag{1}$$

where J denotes the set of indices of 3D Gaussians visible from the candidate viewpoint and k indicates the index assigned to the discretized view direction of the candidate viewpoint. This update is performed based on the previous map $\mathcal{M}_{t-1}$, which can be described as

$$IG_t = \sum (\mathcal{M}_t - \mathcal{M}_{t-1}), \tag{2}$$

where IG_t is the number of Gaussians that have either been newly observed or newly viewed from different directions compared to $\mathcal{M}_{t-1}$. 3D Gaussians that are newly observed or seen from a different angle compared to $\mathcal{M}_{t-1}$, ensuring that only newly gained information is counted. To avoid redundant overlaps among trajectories, $\mathcal{M}$ is shared across among all trajectories Tr.

Termination. The search process terminates under two conditions. First, if no valid candidate views can be expanded from the current viewpoint while the number of cameras in the trajectory is still below the predefined length L, the current branch is pruned and the search resumes from the next-best candidate view in the priority queue. Second, if the trajectory reaches the predefined length L, the search concludes. The resulting trajectory is treated as a set of pseudo-observations, and then begins the searching process for the next trajectory Tr_{i+1}. This process is repeated until N_{Tr} trajectories are constructed. We provide visualizations of example results in the supplementary materials.

3.4. Diffusion prior

Since renderings from virtual viewpoints often partially include well-reconstructed parts, it is desirable to leverage these visual cues. To this end, we train a video diffusion model to enhance degraded rendered images I^V to be realistic images I^G. Our model is built on [21] but introduces several key modifications. First, unlike [21], which adopts a frame interpolation setup, we address extreme view extrapolation in scene exploration by performing next-frame prediction using only the observation closest to the virtual camera as the sole visual cue. Second, to mitigate this ill-posed task, we condition the model on the scene text description and the ground truth image I^T from the nearest training view to each pseudo-observation. These are injected into the cross-attention layers of the diffusion model. At last, we create a dataset specifically designed to model artifacts arising from extreme extrapolation. The details about the dataset and the training process for the diffusion model are provided in the supplementary materials.

3.5. Finetuning

Previous work [21] observed that applying an image-based loss (e.g., L1 loss) to a rendered image from a virtual camera and its pseudo ground truth often degrades well-reconstructed regions, as 3DGS is sensitive to minor inaccuracies from diffusion model. It mitigates this issue by using two confidence levels, where higher confidence values indicate fewer conflicts between the contents of training and virtual viewpoints, thereby guiding 3D Gaussians to receive stronger gradients. It adopts the image level confidence based on the distance between a virtual viewpoint and a training viewpoint and pixel level confidence based on the scale of 3D Gaussians.

Higher image-level confidence is applied when virtual viewpoints far from training viewpoints since they have less conflicts between them. While effective, using distance alone often fails to account for pure camera rotations, as the distance remains unchanged even when the view frustum overlap between viewpoints decreases. As our generated trajectories involve both translations and rotations, this results in suboptimal supervision.

To address this issue, we introduce a novel image-level confidence metric based on intersection-over-union (IoU) of Gaussians, denoted as G-IOU. Our key insight is that a high view frustum overlap between a training viewpoint and a virtual viewpoint implies many co-visible 3D Gaussians, leading to more conflicts. Using the Gaussian rasterizer, we efficiently extract the indices of visible Gaussians from each view, and define their intersection as the co-visible set:

$$\mathcal{G}_{\text{covis}}(\mathcal{V}^V, \mathcal{V}^{\text{ref}}) = \mathcal{G}_{\text{vis}}(\mathcal{V}^V) \cap \mathcal{G}_{\text{vis}}(\mathcal{V}^{\text{ref}}), \tag{3}$$

where $\mathcal{G}_{vis}(\cdot)$ denotes the set of 3D Gaussians visible from a given viewpoint, and $\mathcal{V}_{ref}$ refers to the training viewpoint nearest to $\mathcal{V}^V$. Based on this, G-IOU is defined as

$$G\text{IOU}(\mathcal{V}^V, \mathcal{V}^{\text{ref}}) = \frac{|\mathcal{G}_{\text{covis}}(\mathcal{V}^V, \mathcal{V}^{\text{ref}})|}{|\mathcal{G}_{\text{vis}}(\mathcal{V}^{\text{ref}})|}, \tag{4}$$

and the image-level confidence is given by

$$U_{\text{img}} = 1 - G\text{IOU}(\mathcal{V}^V, \mathcal{V}^{\text{ref}}). \tag{5}$$

For the pixel level confidence, rather than solely relying on the scale of 3D Gaussians which still produce artifacts depending on the viewpoint, we define it as the perceptual distance [46] between the rendered image I^V and the output of the diffusion model I^G. Each pixel in the confidence map reflects its contribution to the overall perceptual difference. We upsample the pixel-aligned LPIPS map to the original image resolution and define pixel-level confidence as

$$U_{\text{pixel}}(x, y) = \mathcal{U}(\text{LPIPS}(I^V, I^G))(x, y), \tag{6}$$

where $U_{\text{pixel}}(x, y)$ denotes the confidence value at pixel (x, y) and $\mathcal{U}$ is an upsampling operation to the original image resolution.

Combining the above confidences, the loss for training views is defined as

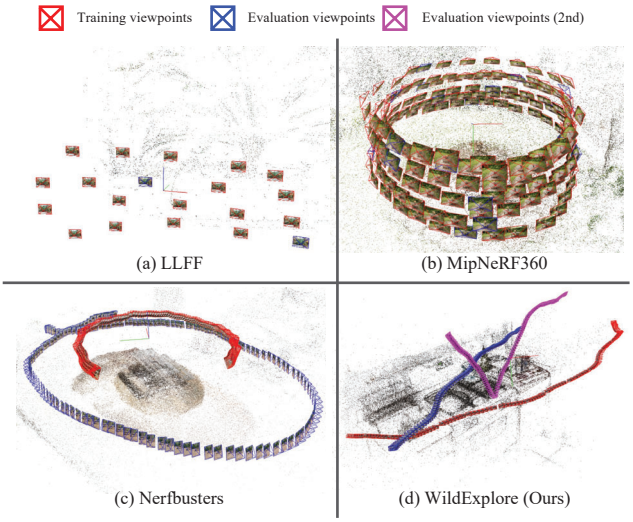


Figure 4. Viewpoint difference comparison with previous 3D reconstruction dataset. Training and evaluation viewpoints are visualized.

$$\mathcal{L}_{\mathcal{V}^T} = ||I^T - I^V||_1 + \mathcal{L}_{\text{D-SSIM}}(I^T, I^V), \tag{7}$$

and the loss for virtual views is defined as

$$\mathcal{L}_{\mathcal{V}^V} = U_{img} \cdot \left(||I^G - I^V||_1 \odot U_{pixel} + \mathcal{L}_{\text{D-SSIM}}(I^G, I^V) \right). \tag{8}$$

The $\odot$ denotes is the pixel-wise multiplication and $\mathcal{L}_{\text{D-SSIM}}$ denotes the dissimilarity variant of the SSIM loss. This fine-tuning process is also illustrated in Fig. 2 (c).

4. Evaluation dataset for scene exploration

Since existing datasets remain inadequate for evaluating scene exploration, as the test cameras are typically positioned within or close to the training trajectories, resulting in interpolation scenarios or only minor viewpoint variations. Although the Nerfbusters dataset [40] introduces extreme novel viewpoints via two separate camera trajectories, its test views are placed far behind the training views. Consequently, the task aligns more with generative outpainting rather than complete scene reconstruction. We introduce two datasets tailored for this task, with additional details provided in the supplementary materials.

4.1. WildExplore

To address the lack of an appropriate benchmark for scene exploration, we introduce *WildExplore*, a new dataset comprising four indoor and four outdoor scenes. As illustrated in Fig. 4, *WildExplore* is designed with two key properties: (1) it explicitly targets the exploration task by placing test viewpoints significantly distant from the training trajectories, ensuring large viewpoint transitions and broad scene coverage, and (2) the evaluation set is captured to ensure sufficient co-visible areas between training and test views, minimizing the need for generating arbitrary contents, in contrast to Nerfbusters.

4.2. Curated Nerfbusters

We curate the Nerfbusters dataset to better align with scene exploration objectives. We selects some scenes where the test views are placed far behind the training views and swap the train/test split, while leaving the others unchanged. This adjustment transforms the task into "moving forward and closely examining the scene", which better reflects realistic exploration scenarios. Our curated version includes seven scenes with swapped splits and two in their original form.

5. Experiments

5.1. Experimental Setup

Implementation details. We first train 3DGS using the given training viewpoints, following the original setup in [11]. The number of virtual trajectories N is set to 20 by default. The occupancy grid resolution S is set to 64. At each step, we select the top 3 candidate views based on computed information gain. We set D as 32 for discretized view direction. For the video diffusion model, we adopt DynamiCrafter [42] as the backbone. After generating the pseudo observation set, we fine-tune 3D Gaussians for 15K iterations, applying densification until 9K, with the same scheduling as 3DGS. Further details will be discussed in the supplements.

Baselines. We evaluate our method in comparison to 3DGS-based methods, including 3DGS [11], 3DGS + Depth, DNGaussians [15], and ViewExtrapolator [19]. 3DGS + Depth applies additional depth regularization using monocular depth estimation [44], following the implementation from the official 3DGS repository [11]. DNGaussians [15] incorporates depth priors during partial optimization of the 3D Gaussian parameters. ViewExtrapolator leverages video diffusion but heuristically selects a single test camera for generating additional viewpoints due to the absence of a dedicated sampling strategy. In our experiments, we follow their protocol and choose the test camera farthest from the training set to encourage broader viewpoint coverage.

5.2. Results

Quantitative comparison. Table 1 reports the quantitative results on the exploration task. To assess reconstruction quality, we measure PSNR, SSIM, and LPIPS [46]. The first row reports the result from 3DGS trained on both the training and test splits, serving as an upper bound. Since our method effectively removes artifacts and fills missing regions, as observed in the qualitative analysis, our model outperforms across all metrics, particularly in PSNR and SSIM.

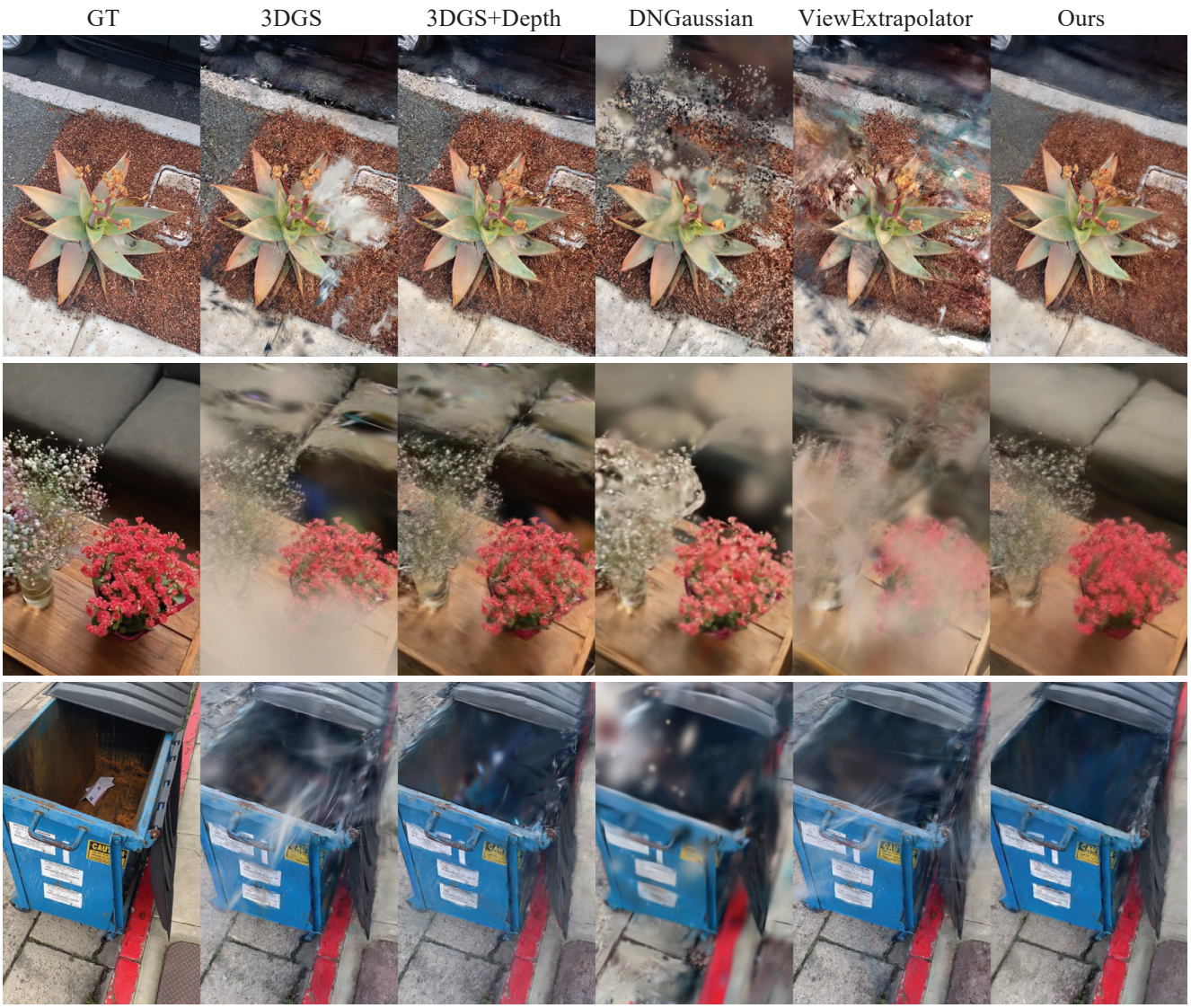


Figure 5. Qualitative comparison on the curated Nerfbusters dataset.

Qualitative comparison. Fig. 5 and Fig. 6 show qualitative comparisons among our method and baseline methods. 3DGS [11] suffers from artifacts, and its variants with depth regularization also meet the same problem, as they lack the capability to fill missing information. Although ViewExtrapolator [19] utilizes a video diffusion prior and thereby shows competitive performance compared to 3DGS variants, it underperforms in challenging scenarios. Its heuristic camera sampling struggles to maximize information gain and results in limited viewpoint coverage, leading to lower visual quality when rendering from distant viewpoints. Its training-free diffusion prior also acts as a bottleneck. In contrast, our method fills missing regions and removes artifacts more effectively, producing images that align closely with the ground truth. See the supplementary materials for additional results.

Method	WildExplore			Curated Nerfbusters		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS Pseudo GT	30.37	0.910	0.144	29.34	0.911	0.128
3DGS	15.81	0.487	0.414	14.51	0.417	0.454
3DGS + Depth	15.78	0.483	0.413	15.00	0.435	0.436
DNGaussians	15.65	0.504	0.450	14.01	0.442	0.525
ViewExtrapolator	16.02	0.510	0.442	14.98	0.450	0.455
Ours	16.96	0.534	0.413	16.29	0.479	0.434

Table 1. Quantitative comparison.

5.3. Ablation study

Information gain. We validate the effectiveness of our information gain design by comparing it with alternative strategies, as shown in Table 2. One alternative samples the 3D space uniformly and evaluate view coverage based on the occupied vertices of an occupancy grid, as proposed in



Figure 6. Qualitative comparison on the WildExplore dataset.

Virtual View Sampling	Curated Nerfbusters		
Info. Gain	PSNR↑	SSIM↑	LPIPS↓
Grid view coverage [22]	15.82	0.4675	0.4451
FisherRF [9]	15.77	0.4593	0.4394
GS view coverage (Ours)	**16.29**	**0.4790**	**0.4336**

Table 2. Comparisons on information gain design.

Method (K = 3)	*Garbage*		
	PSNR↑	SSIM↑	LPIPS↓
3DGS	14.42	0.480	0.526
Ours with Bottom-K	14.22	0.475	0.521
Ours with Top-K	**15.45**	**0.526**	**0.508**

Table 3. Ablation study on information gain. TopK vs BottomK

Finetuning		Curated Nerfbusters		
Image level	Pixel level	PSNR↑	SSIM↑	LPIPS↓
Distance [21]	-	15.00	0.427	0.443
-	Scale [21]	16.18	0.476	0.442
Distance	Scale	15.27	0.432	0.440
-	-	16.30	0.476	0.447
-	pixelLPIPS	16.23	0.478	0.441
GS-IOU	-	16.10	0.473	0.439
GS-IOU	pixelLPIPS	16.29	0.479	0.434

Table 4. Ablations study on finetuning methods.

[22]. Another follows typical active-learning strategies and select viewpoints with the highest uncertainty. We adopt the method proposed in [9] for comparison. As presented in the first and second rows, both show suboptimal results. Grid-based approach often fails to maximize information gain, as it includes the gain from free space, resulting in redundant viewpoint selections. For uncertainty-based method, we observe that it is sensitive to 3D Gaussians with large scales, tending to chase them rather than addressing artifacts from viewpoint changes.

Search strategy. We ablate our best first search strategy to validate its effectiveness. As shown in Table 3, the strategy selecting bottom K makes virtual cameras tend to stay near the trajectory of training viewpoints rather than deviating from it, leading to no performance gain.

Finetuning methods. We validate our finetuning approach by benchmarking it against the methods proposed by [21] , as shown in Table 4. The model without the image-level confidence produces worse results in terms of LPIPS, as generated images whose viewpoints closely match the training set negatively impact already well-reconstructed regions. Ignoring camera rotation further degrades performance, as observed in the second and fifth rows. Although both our image-level confidence map and scale based one are effective, our method leads to slight better performance.

6. Conclusion

This paper presents a framework that enables high-quality rendering of the novel view in arbitrary viewpoints. Our method incorporates the video diffusion priors and the real-time rendering capability of 3DGS. It introduces a sampling strategy to effectively sample trajectories of virtual cameras for maximizing information gain.

Limitations and future works. While our method incorporates video diffusion priors to fill the missing information, we expect that employing more strong diffusion models which can faithfully generate high frequency details may improve the overall performance. In addition, extending the scene bounding box to cover a large scale scene would be an interesting avenue for the future work.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220113, No. RS-2024-00457882), Artificial Intelligence Graduate School Program, Yonsei University, under Grant RS-2020-11201361.

References

[1] 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*, pages 5470–5479, 2022. 1, 2

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields, 2023. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 3

[4] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models, 2024. 3

[5] Jeong gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models, 2023.

[6] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2024.

[7] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024.

[8] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion, 2024. 3

[9] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *arXiv preprint arXiv:2311.17874*, 2023. 3, 8

[10] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11305–11312, 2023. 3

[11] 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. 1, 2, 3, 6, 7

[12] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36:1–13, 2017. 2

[13] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Eschernet: A generative model for scalable view synthesis, 2024. 3

[14] Georgios Kopanas and George Drettakis. Improving NeRF Quality by Progressive Camera Placement for Free-Viewpoint Navigation. In *Vision, Modeling, and Visualization*. The Eurographics Association, 2023. 3

[15] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization, 2024. 2, 6

[16] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision, 2023. 3

[17] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model, 2024. 3

[18] Guangyi Liu, Wen Jiang, Boshu Lei, Vivek Pandey, Kostas Daniilidis, and Nader Motee. Beyond uncertainty: Risk-aware active view acquisition for safe robot navigation and 3d scene understanding with fisherrf, 2025. 3

[19] Kunhao Liu, Ling Shao, and Shijian Lu. Novel view extrapolation with video diffusion priors, 2024. 2, 3, 6, 7

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[21] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors, 2024. 2, 3, 5, 8

[22] Pierre Marza, Laetitia Matignon, Olivier Simonin, Dhruv Batra, Christian Wolf, and Devendra S Chaplot. Autonerf: Training implicit scene representations with autonomous agents. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13442–13449. IEEE, 2024. 3, 8

[23] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, 2019. 2

[24] 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. In *ECCV*, 2020. 1, 2

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 1

[26] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan.

Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs, 2021. 2

[27] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pages 230–246. Springer, 2022. 3

[28] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 3

[29] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction, 2021. 3

[30] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. Zeronvs: Zero-shot 360-degree view synthesis from a single image, 2024. 3

[31] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In *2021 International Conference on 3D Vision (3DV)*, pages 972–981. IEEE, 2021. 3

[32] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2024. 3

[33] Meng-Li Shih, Wei-Chiu Ma, Lorenzo Boyice, Aleksander Holynski, Forrester Cole, Brian Curless, and Janne Kontkanen. Extranerf: Visibility-aware view extrapolation of neural radiance fields with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20385–20395, 2024. 2, 3

[34] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields, 2022. 3

[35] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses, 2023. 2

[36] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion, 2024. 3

[37] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. 3

[38] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation, 2023. 3

[39] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation, 2024. 3

[40] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs, 2023. 2, 6

[41] Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Leo Lebrat. Nerf director: Revisiting view selection in neural volume rendering, 2024. 3

[42] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors, 2023. 6

[43] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360deg $sparseviewsynthesisusinggaussiansplatting$, 2024. 2

[44] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 6

[45] Huangying Zhan, Jiyang Zheng, Yi Xu, Ian Reid, and Hamid Rezatofighi. Activermap: Radiance field for active mapping and planning, 2022. 3

[46] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 5, 6

[47] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. 3