Few-Shot Gaussian-Splatting 3D Reconstruction Enhancement and Application

Paris Zhang
Department of Computer Science
Stanford University
parisz@stanford.edu

Jack Liu
Department of Computer Science
Stanford University
jiayiliu@stanford.edu

Abstract

3D scene reconstruction from limited images is a challenge in computer vision. 3D Gaussian Splatting (3DGS) offers high-quality, rapid reconstruction but relies on numerous images. This project aims to enhance 3DGS performance with limited data and apply it to virtual room tours.

We improve 3DGS by integrating depth information to refine and stabilize the reconstruction process and using diffusion-based novel view synthesis for further regularization. This approach aims to generate accurate and visually coherent reconstructions in a few-shot setting. For real-world application, we developed a system to create high-quality 3D virtual room tours from a single video clip. This system addresses challenges like insufficient and imperfect input data and simplifies the pipeline for non-expert users.

Experiments show our enhancements improve reconstruction quality significantly. The virtual room tour application delivers robust 3D reconstructions efficiently, even from casually shot videos. Our code implementation can be found here.

1. Introduction

3D scene reconstruction using a limited number of images remains a significant challenge in computer vision. A recent advancement that has garnered attention is 3D Gaussian Splatting (3DGS) [5] through its high quality, rapid reconstruction speed, and real-time rendering support. Despite these advantages, one notable limitation is that 3DGS depends on a substantial number of images to maintain high quality. It relies on optimizing independent splats under multi-view color supervision but lacks consideration for global structural coherence. Consequently, when image input is minimal, the model tends to converge to local optima, resulting in optimization failure or floating artifacts [2].

Our project aims to enhance the performance of 3DGS with limited training data and explore its application in real-world contexts.

For performance enhancement, we integrated depth as additional geometric information to refine and stabilize the reconstruction process. Furthermore, we leveraged diffusion-based novel view synthesis techniques to further regularize and improve the reconstruction. By incorporating geometry and diffusion priors, our approach seeks to generate more accurate and visually coherent reconstructions in a few-shot setting.

For the real-world application, we collected our own dataset and optimized Gaussian Splats to generate interactive and high-quality 3D virtual room tours from one video clip of the indoor rooms. We also created an efficient and user-friendly system for the virtual room tour creation.

2. Related Works

Neural Radiance Field (NeRF). The introduction of Neural Radiance Field (NeRF) [7] marked a significant advancement in view synthesis through implicit functions to encode both volumetric density and color observations. Building on this framework, subsequent research has focused on enhancing NeRF's capability in scenarios with limited image data by integrating 3D geometric priors. In particular, [3] employed sparse 3D point clouds derived from structure-from-motion (SfM) solvers and imposed a loss term to align the distribution of ray terminations with 3D keypoints. Similarly, [6] incorporated image correspondences obtained through off-the-shelf models and added both pixel projection and depth loss terms to improve the regularization during NeRF training. There are also works that utilize diffusion priors, for example, ReConfusion [10]. It trains a multi-view conditioned diffusion model to produce plausible single images for novel camera poses and uses it as a prior to help regularize radiance field reconstruction at the novel view.

Regularized 3DGS. 3DGS [5] introduces an innovative scene representation technique using anisotropic 3D Gaussians along with a differentiable renderer, significantly im-

proving both scene construction and novel view synthesis speeds. A further refinement came from [2] that integrated depth information into the 3DGS framework by treating the estimated depth from single input images as ground truths and comparing the rendered depth with it through a training loss term. This method is described in more detail in Section 3 as it serves as our starting point.

3. Approach

We used the official implementation of Depth-Regularized GS (DRGS) [2] as our starting point and baseline.

3DGS optimizes the Gaussian splats based on the rendered image with a color loss $\mathcal{L}_{color}$ and D-SSIM loss $\mathcal{L}_{D-SSIM}$. Prior to 3DGS optimization, DRGS estimates a depth map, $\mathcal{D}_{dense}$, for each image via a monocular depth estimation network, ZoeDepth [1]. This map is then refined using sparse 3D points generated from the Structure-from-Motion solver COLMAP [9] to get $\mathcal{D}^*_{dense}$. During 3DGS optimization, the depth of the Gaussian splats, $\mathcal{D}$, is rendered through rasterization and compared with the adjusted depth map with L1 distance, $\mathcal{L}_{depth} = ||\mathcal{D} - \mathcal{D}^*_{dense}||_1$.

Additionally, to ensure smooth transitions in depth between adjacent pixels, they apply a smoothness constraint. For the depth of a pixel d_i and the depth of its adjacent pixel d_j where d_i and d_j are not on an edge, the smoothness loss term is defined as $\mathcal{L}_{smooth} = \sum_{d_j \in \text{adj}(d_i)} ||d_i - d_i||^2$.

The final loss terms are

$$\mathcal{L} = (1 - \lambda_{ssim})\mathcal{L}_{color} + \lambda_{ssim}\mathcal{L}_{D-SSIM} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{smooth}\mathcal{L}_{smooth}$$

We conduct ablation studies to study the effectiveness of each component of Depth-Regularized GS. Then we explore two methods to improve the reconstruction quality.

3.1. Depth Estimation Enhancement

The performance of DRGS relies heavily on the accuracy of the depth maps generated by ZoeDepth [1]. We replaced it with more advanced depth estimation techniques to generate more accurate "ground-truth" depth, thus potentially enhancing the reconstruction quality.

Depth-Anything [12] is a more robust foundation model trained on a significantly larger dataset, 1.5M labeled images and 62M+ unlabeled images jointly, which produces better zero-shot metric depth estimation than ZoeDepth. This switch has improved the fidelity and consistency of our 3D reconstruction.

3.2. Novel-view Guidance from Diffusion

Inspired by ZeroNVS [8], which trains a 3D-aware diffusion model for single-image novel view synthesis (NVS), then, during reconstruction, uses synthesized novel views to perform SDS-based NeRF distillation, we also leverage a generative diffusion prior for novel view synthesis to regularize a GS-based 3D reconstruction pipeline at novel camera poses.

Similar to how ZeroNVS uses its synthesized novel views for NeRF distillation, our pipeline directly incorporates the diffusion prior, leveraging ZeroNVS to generate novel-view images to augment our training set. With limited training images, we can utilize the diffusion model to generate more images in different views to greatly diversify our training set.

Since we synthesize novel views for each original training image independently of each other, we are concerned that the depth information retrieved from the monocular depth estimation across the novel views might be inaccurate. So we will also explore whether we should incorporate all DRGS loss terms ($\mathcal{L}_{color}$, $\mathcal{L}_{D-SSIM}$, $\mathcal{L}_{depth}$, and $\mathcal{L}_{smooth}$) or only the loss terms based on rendered images without any depth loss ($\mathcal{L}_{color}$ and $\mathcal{L}_{D-SSIM}$).

3.3. Real-World Application: Virtual Room Tour

The largest challenges for applying 3DGS to creating a virtual room tour include the handling of insufficient and imperfect input data—casually shot videos that do not adhere to "photogrammetry best practices". Therefore, we explored the application of DRGS to solve this task. Moreover, the typical pipeline for processing these videos into 3D tours is fraught with multiple error-prone steps that can be intimidating for non-expert users, and the training time usually takes hours. Besides using depth to regularize the training, we employed the following algorithm and system developments:

End-to-End Script. Using a custom dataset is usually poorly supported in existing Gaussian Splatting projects. Initially, users need to upload a video, convert the video into images, estimate the pose of each image, perform image and camera pose conversion, and select training images. Each step requires determining the appropriate tools and passing in arguments, which can be complex for average users. To simplify this process, we have wrapped all steps into an executable script, so users only need to provide the video to receive the output point cloud in a specified folder. We use FFmpeg for converting video to frames and COLMAP for pose estimation. Additionally, we carefully tune parameters such as frames per second and feature matching methods, as these are important for training speed and final quality.

Automatic COLMAP Error Fix. Without carefully shooting videos of the entire house, COLMAP often fails in feature matching of consecutive frames. Specifically, any

poor-quality image, such as a blurred image due to sudden movement or an image with a white wall, can cause consecutive features not to have sufficient overlap, resulting in matching failure. The resulting pose estimation of frames may be broken into separate sets under different global coordinates: $S = \{s_0 = (I_1, I_3, I_4|G_0); s_1 = (I_5, I_6, I_7|G_1)\}$ where I denotes image and G denotes global coordinate. In this case, I_2 is corrupted, and I_4, I_5 do not share sufficient feature overlap. The original pipeline would only proceed with s_0 and discard the rest due to errors caused by non-continuous frame number naming conventions in the set. We have developed a robust pipeline to: 1) detect multiple image sets, 2) separate images into multiple sets and rename them to follow a continuous frame number convention, including those referenced in the camera and image pose binary files, and 3) run training on $s_0, s_1, \cdots, s_n$ to produce a reconstruction of the entire scene captured in the video.

Sample Selection. To accelerate the training process, we strategically select 30-50 images from the input video frames as the training set. This selection process utilizes the camera extrinsic parameters determined by COLMAP. After positioning the cameras in a 3D space relative to a computed central vector, we applied a RANSAC-like algorithm to choose a number of indices that maximize the product of angles between each pair of vectors. This helps in maximizing the spatial diversity of selected frames and enhancing the overall effectiveness of the Gaussian splatting process. We use all the other images and views as a test set to evaluate reconstruction quality.

Training Extension & Early Stopping. Originally, the DRGS method proposed ceasing training after 1,500 iterations to prevent overfitting in scenarios with 2 to 5 available views. Given the more complex inputs, we extend the training up to 5,000 iterations but apply a similar early stopping mechanism. Training is halted if an increase in depth loss is detected alongside satisfactory low RGB and D-SSIM losses. This strategy significantly reduces the training time and prevents overfitting.

4. Experiment / Analysis

For evaluation of reconstruction quality enhancement, we used the NeRF LLFF [7] dataset in 3D scene reconstruction. Following the data preprocessing steps in Depth-Regualarized GS [2], we randomly sample 2-5 images from the training split of the dataset, and evaluate the results on the eval split. For the application of virtual room tours, we collected our own evaluation dataset, which consists of videos of various lengths capturing different types and sizes of apartment rooms.

Method	2-Views			5-Views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o depth adjustment	9.913	0.351	0.769	9.589	0.340	0.765
w/o $\mathcal{L}_{depth}$	8.557	0.130	0.653	12.095	0.320	0.449
w/o $\mathcal{L}_{smooth}$	15.295	0.364	0.402	17.791	0.492	**0.323**
w/o early stop	14.595	0.306	0.422	17.120	0.447	0.333
DRGS	**15.343**	**0.367**	**0.404**	**17.963**	**0.500**	**0.323**

Table 1. **Ablation**. ablation study of different components of our method on 2-view and 5-view renderings

4.1. Ablation Studies

We first perform ablation studies to quantitatively understand the importance of each component in DRGS, as similar components will also be used in our method. As shown in Table 1, all components of DRGS's method contributes to both 2-views and 5-views rendering on all three metrics. As we can see from the experiment, the most significant contributing factor to the reconstruction quality are indeed $\mathcal{L}_{depth}$ and the depth estimation adjustment, where as the $\mathcal{L}_{smooth}$ shows the least significant improvement. This suggests that incorporating and improving high quality depth prior significantly improves reconstruction results.

4.2. Depth Estimation Enhancement

We replaced ZoeDepth with Depth-Anything for a more accurate metric depth estimation and present the quantitative comparison results of 3DGS, DRGS, and our method for NeRF-LLFF scenes in Table 2.

On average, our model outperforms 3DGS and DRGS across 8 datasets and typically demonstrates superior results, especially when the number of images is limited to 2 views. This improvement is likely because more accurate depth information is most helpful when input images are scarce. However, for 2-view inputs, our method performs worse on datasets like Fortress and Trex, possibly because Depth-Anything excels at estimating depth for outdoor scenes but is less accurate for indoor scenes.

Qualitatively, we compare DRGS and our method in Figure 1. For the 2-view experiments, Depth-Anything estimates more accurate depth for the background scene and predicts a clearer distinction between the foreground and background. Consequently, the novel view images from our method better reconstruct the background and provide a relatively clearer view, even with only 2 training images.

In the 5-view Trex experiment, Depth-Anything again demonstrates superior and more nuanced depth estimation for the background stairs and the trex. As a result, our novel view rendering is much sharper than DRGS—the Trex's head is more defined, and the background stairs are clearer. However, Depth-Anything performs worse than ZoeDepth

Dataset	Method	2-view			5-view		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fern	3DGS	14.247	0.360	0.447	18.847	**0.611**	**0.214**
	DRGS	17.404	0.494	**0.324**	**19.494**	0.577	0.254
	Ours	**17.859**	**0.519**	**0.321**	19.227	0.570	0.237
Flower	3DGS	15.953	0.380	**0.346**	**20.674**	**0.643**	**0.157**
	DRGS	16.408	0.402	0.413	19.273	0.532	0.319
	Ours	**16.483**	**0.417**	0.396	20.317	0.576	0.280
Fortress	3DGS	12.831	0.364	0.410	18.005	0.595	0.252
	DRGS	**19.340**	**0.504**	**0.273**	**22.791**	**0.728**	**0.174**
	Ours	17.481	0.497	0.302	20.834	0.689	0.247
Horns	3DGS	13.541	**0.407**	0.511	11.702	0.324	0.489
	DRGS	13.793	0.389	**0.411**	15.741	**0.473**	**0.380**
	Ours	**14.021**	0.359	**0.416**	**16.406**	0.458	**0.383**
Leaves	3DGS	12.501	**0.259**	0.443	11.346	0.254	0.427
	DRGS	12.295	0.181	**0.542**	13.725	0.250	**0.493**
	Ours	**13.263**	**0.263**	0.498	**14.218**	**0.308**	0.472
Orchids	3DGS	11.841	0.180	**0.412**	**16.570**	**0.495**	**0.189**
	DRGS	12.081	0.181	0.491	16.142	0.395	0.348
	Ours	**12.406**	**0.201**	0.504	16.450	0.416	0.324
Room	3DGS	11.026	0.478	0.585	12.279	0.567	0.511
	DRGS	17.596	0.658	**0.364**	19.958	**0.771**	0.321
	Ours	**17.979**	**0.684**	**0.357**	**20.465**	**0.777**	**0.308**
Trex	3DGS	10.489	0.341	0.563	13.752	0.484	0.387
	DRGS	**14.510**	**0.459**	0.406	16.091	0.554	0.385
	Ours	14.286	0.417	**0.374**	**17.168**	**0.561**	**0.343**
Mean	3DGS	12.804	0.346	0.465	15.773	0.497	**0.329**
	DRGS	15.928	0.409	0.403	17.902	**0.535**	0.334
	Ours	**15.972**	**0.420**	**0.396**	**18.387**	**0.544**	**0.324**

Table 2. **Quantitative comparison between our model and 3DGS and DRGS baselines.** On average, our model outperforms both baselines across 8 datasets.

on the Fortress dataset. When the input image primarily features an object without a complex background, Depth-Anything struggles to estimate the subtle nuances of the object, resulting in a less accurate reconstruction.

4.3. Novel-view Guidance from Diffusion

To effectively utilize the ZeroNVS model, in our pipeline, for each original input view, we generate four novel views from nearby camera positions, and use these synthesized views as input into the DRGS method. For example, in the 5-view case, with 5 original views, we would generate an additional 20 novel views, so that the final input size to DRGS would contain 25 views in total. Additionally, to distinguish whether the depth loss from the novel views is effective, we experimented with both including and excluding depth loss on the synthesized views. However, due to the limited amount of resources we have and the resource-intensiveness of ZeroNVS, we are only able to run our experiment on two datasets from Nerf-LLFF, namely fern and fortress.

Quantitatively, as shown in Table 3, overall, training with the novel views improves scores in all three metrics across both datasets. However, the inclusion or exclusion of depth proves beneficial depending on the specific context, with no definitive approach emerging as superior. Another interesting phenomenon we observed is that, in the five-view case when training on the fern dataset, the original DRGS without synthesized novel views has the best performance. We speculate that the inferior performance of training with novel views is due to the possible inconsistencies between

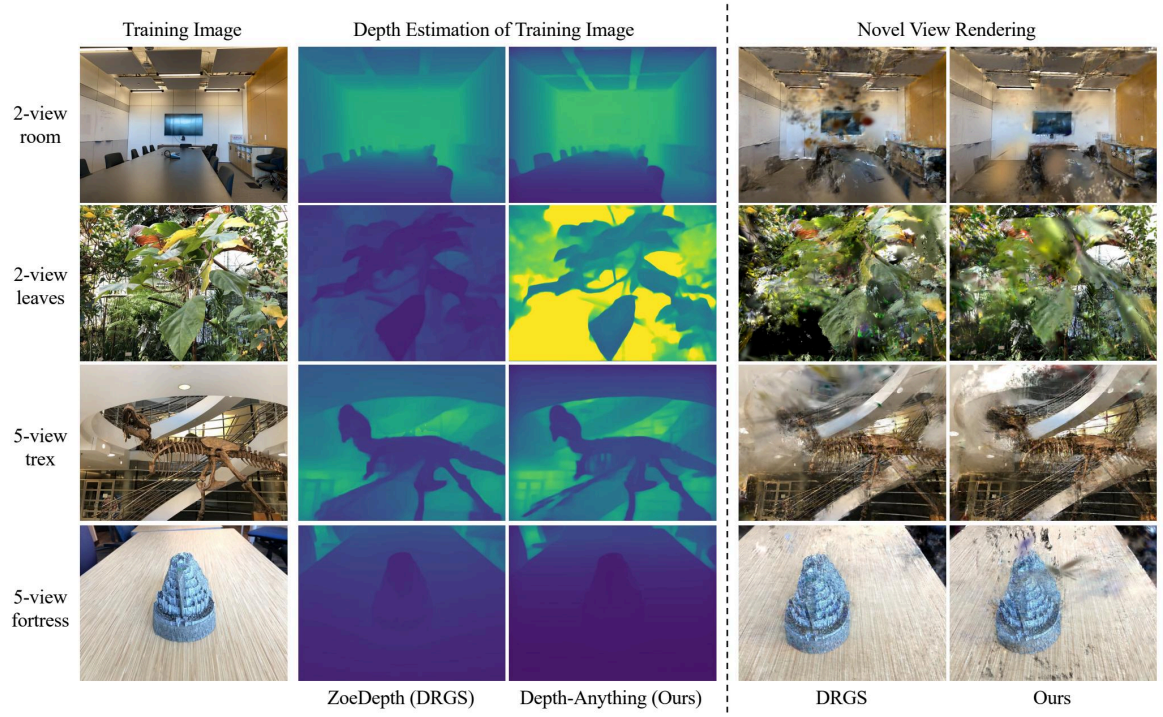

Figure 1. **Qualitative comparison between our model and DRGS on depth estimation and novel view synthesis.** Generally, Depth-Anything is able to estimate more nuanced and accurate depth maps, especially for the background, which results in better GS reconstruction. However, Depth-Anything is worse than ZoeDepth on Fortress, when the input image is mainly an object without complex background.

a large number of synthesized views, especially since the novel view synthesis from each original input view is independent of each other.

Qualitatively, as shown in Figure 2, training with novel views from ZeroNVS has significantly enhanced overall consistency and reduced artifacts, particularly in regions that were unobserved or less observed in the original input images (e.g. edges).

4.4. Virtual Room Tour Applications

4.4.1 3D Reconstruction Quality

We compare the reconstruction quality of our method with the original 3DGS. We also conducted experiments to apply the early stop strategy to 3DGS to make a comparison when two methods are trained at approximately the same time.

Qualitatively, our method produces robust reconstructions that closely match the true geometry of the rooms within a mere 8 minutes, as illustrated in Figure 3. In contrast, the original 3DGS method, which requires up to 1.5 hours, results in reconstructions with significant floating artifacts that seriously affect room views. When constrained to the same 8-minute timeframe, our method again consis-

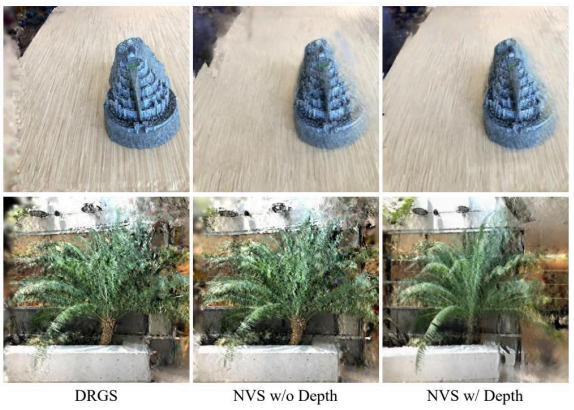

Figure 2. **Qualitative comparison between DRGS and our approach augmented by novel views with/without depth losses.** The augmented novel views significantly enhance the consistency and reduce the artifacts in reconstructions.

Dataset	Method	1-view			2-view			3-view			5-view		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fern	DRGS	15.426	0.357	0.415	14.581	0.330	0.477	15.890	0.383	0.469	**18.793**	**0.544**	**0.306**
	NVS w/o Depth	15.459	0.355	0.418	14.584	0.331	0.477	15.888	0.388	0.463	18.553	0.538	**0.306**
	NVS w/ Depth	**16.645**	**0.426**	**0.377**	**16.551**	**0.449**	**0.437**	**17.830**	**0.529**	**0.423**	16.483	0.462	0.550
Fortress	DRGS	16.115	0.517	0.3736	**20.333**	0.603	0.259	17.045	0.539	0.333	18.024	0.580	0.312
	NVS w/o Depth	**16.328**	**0.524**	**0.368**	19.602	0.625	0.295	16.921	0.621	0.283	18.249	0.682	0.259
	NVS w/ Depth	16.256	0.520	0.373	19.828	**0.645**	**0.242**	**17.679**	**0.631**	**0.267**	**18.590**	**0.693**	**0.251**

Table 3. **Quantitative comparison between DRGS baseline method and our approach with diffusion-synthesized novel views.** Overall, training with synthesized novel views from diffusion model outperforms the baseline.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓
bedroom	3DGS (1.5 hr)	24.417	**0.932**	0.141
	3DGS (8 min)	23.644	0.916	0.167
	Ours (8 min)	**24.921**	**0.931**	**0.111**
livingroom	3DGS (1.5 hr)	24.539	0.924	**0.116**
	3DGS (8 min)	25.408	**0.940**	0.125
	Ours (8 min)	**25.661**	**0.935**	0.135
apartment	3DGS (1.5 hr)	22.827	0.911	**0.121**
	3DGS (8 min)	23.318	**0.925**	0.128
	Ours (8 min)	**23.454**	**0.920**	0.150

Table 4. **Quantitative comparison between our method and 3DGS.** Our method is able to achieve better reconstruction similarity within a short amount of time.

tently delivers better visual quality and fewer artifacts compared to 3DGS.

Quantitatively, as shown in Table 2, our method outperforms the 3DGS baselines in terms of SSIM and PSNR across different test datasets. Both metrics measure more pixel-level accuracy, so it's indicative that our reconstructions align more closely with the ground truths in terms of the fine-grained structure and details. However, we noticed that our method often scores lower in LPIPS compared to 3DGS by LPIPS. This discrepancy suggests that while our method excels in detailed fidelity, 3DGS is slightly better at capturing the overall conceptual essence of the input scenes, especially after extended 1.5 hours of training durations.

4.4.2 Ablation Studies

We performed ablation studies to quantitatively measure the importance and influence of each technical approach we designed in Section 3.3. As shown in Table 5, sample selection is effective in reducing the training time with very slight degradation in reconstruction quality. Early stopping significantly reduces the training time while training extension boosts the reconstruction quality. Finally, the depth estimation model enhancement introduces a small increase in the reconstruction quality, especially for SSIM. Overall, our algorithm designs enable our method to effectively minimize training time while maximizing the quality of the 3D reconstruction.

Experiment	Time ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o sample selection	12 min	25.770	0.937	0.132
w/o early stop	1.75 hrs	26.333	0.930	0.118
w/o training extension	4 min	24.602	0.930	0.157
w/o depth enhancement	8 min	25.636	0.936	0.134
Ours	8 min	25.661	0.935	0.135

Table 5. **Ablation studies.** For each metric, we color the best result as green and the worst result as red. Our method strikes a balance between the training time and reconstruction quality.

5. Conclusion

In this project, we successfully enhanced 3D Gaussian Splatting (3DGS) for few-shot 3D scene reconstruction and applied it to create virtual room tours from single video clips. By integrating depth information and leveraging diffusion-based novel view synthesis, we achieved more accurate and visually coherent reconstructions with limited training data.

We have learned that incorporating depth information and diffusion models can significantly improve the stability and accuracy of 3DGS, particularly when the input images are limited. Our system for creating virtual room tours demonstrates the practical utility of enhanced 3DGS, capable of producing high-quality 3D reconstructions efficiently from casually shot videos.

There are a few potential future work ideas.

1. Besides depth, incorporating other geometric priors from the input images, such as optical flow and stereo matching data from Unimatch [11], and dense correspondence matching from DKM [4], may offer further improvements.

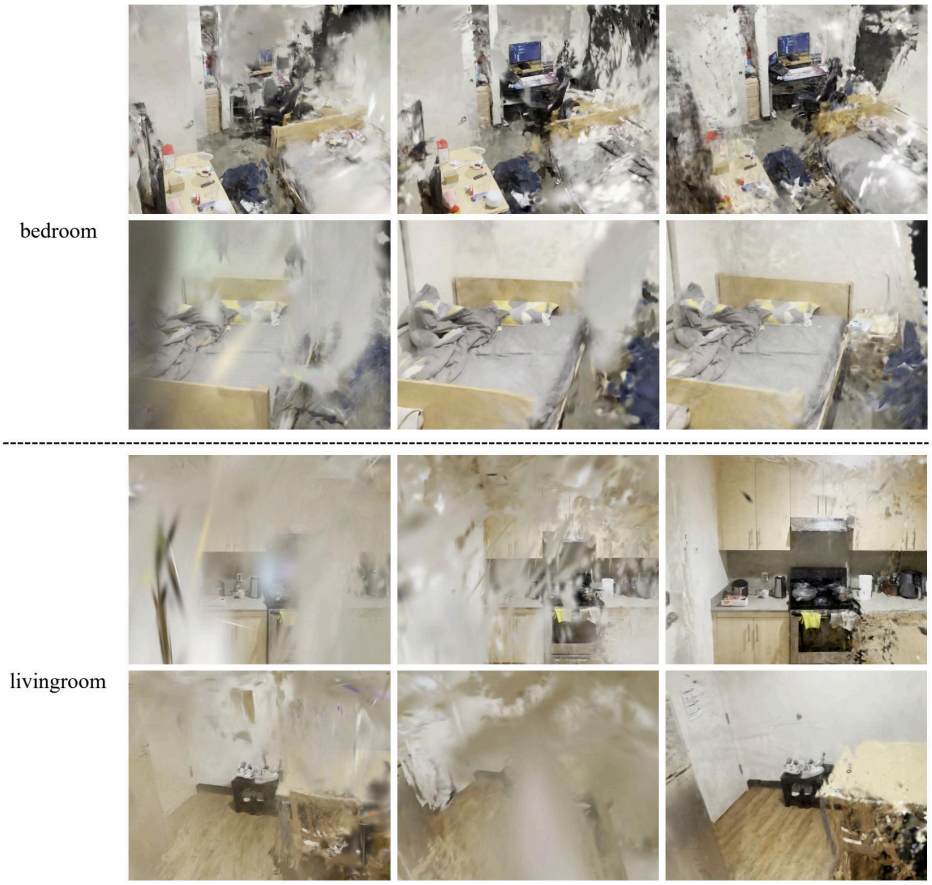

Figure 3. **Qualitative comparison between our model and 3DGS.** Our method better aligns with the true geometry of the room and generates fewer artifacts.

2. Similar to the approach from ReConfusion [10], which uses a multi-view conditioned diffusion model to produce plausible images for novel camera poses and regularize radiance field reconstruction, future works can leverage a similar diffusion prior to generate plausible novel-view images during training for 3DGS. This process involves rendering an image from the 3D Gaussian-Splatting field, perturbing it to a noisy latent, and generating a target image to supervise the rendering.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[2] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. *arXiv preprint arXiv:2311.13398*, 2023. 1, 2, 3

[3] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 1

[4] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 6

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1

[6] Yixing Lao, Xiaogang Xu, Xihui Liu, Hengshuang Zhao, et al. Corresnerf: Image correspondence priors for neural radiance fields. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[7] 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. 1, 3

[8] 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 real image. *arXiv preprint arXiv:2310.17994*, 2023. 2

[9] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[10] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. *arXiv*, 2023. 1, 7

[11] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[12] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2