Splatter Scene: Single/Few Images to 3D Gaussians on Indoor Scenes

Qianru Li, Yunfei Deng
Technical University of Munich
qianru.li@tum.de, philips.deng@tum.de

Abstract

One-view reconstruction of indoor scenes is highly under-constrained, leading to challenges in accurately predicting 3D Gaussians. Splatter Image [7] can reconstruct a 3D object represented by 3D Gaussians from one single image using a straightforward U-Net [5] architecture. However, this method struggles with indoor scene data due to the scaling ambiguity and a lack of constraints during training. Through this work, we propose Splatter Scene, which extends the ability of Splatter Image from single-object reconstruction to indoor-scene reconstruction. We incorporate depth information into the training process and use two views to further improve reconstruction quality, which helps significantly in rendering consistent novel views of an indoor scene from single/few images.

1. Introduction

As a newly proposed representation method, for its high rendering speed, 3D Gaussian Splatting [3] has been used in many reconstruction works. One-view reconstruction of indoor scenes poses challenges in accurately predicting 3D Gaussians due to highly under-constrained problem setting. This results in sub-optimal quality of novel view renderings and depth estimates, particularly in complex and occluded indoor environments. We want to propose a method that can render consistent novel views of an indoor scene from single/few images. To do this, we start with Splatter Image [7], which can reconstruct a 3D object represented by 3D Gaussians from a single image using a U-Net architecture. However, this method struggles with indoor scene data due to the scaling ambiguity and a lack of constraints during training. We incorporate depth supervision into the training process and use two input views to improve reconstruction quality, which helps producing consistent novel views of an indoor scene from single/few images significantly.

In summary, the main contributions of this work are as follows:

1. Incorporate depth supervision into monocular indoor scenes prediction using 3D Gaussians.
2. Use a 2D network for 3D scene reconstruction from single/few views.

2. Related Work

2.1. Single Image to Indoor Scenes

Splatter Image [7] proposes to regress pixel-aligned Gaussian parameters from a single view with a U-Net. However, it focuses on single object reconstruction, while we target a more general setting and larger scenes. PixelNeRF [9] presents a learning framework that predicts a continuous neural scene representation conditioned on one or few input images. However, due to costly inference time of NeRF [4]-based methods, to train a network based on pixelNeRF is rather time-consuming. In contrast, 3DGS (3D Gaussian Splatting) [3] avoids NeRF's expensive volume sampling via a rasterization-based splatting approach, where novel views can be rendered very efficiently from a set of 3D Gaussian primitives. Our 3DGS-based model's rendering speed is much faster, which is suitable for efficient training and evaluation.

2.2. Two Images to Indoor Scenes

Using two images to reconstruct indoor scenes allows for leveraging stereoscopic cues, which can significantly enhance the depth estimation and spatial understanding of a scene. PixelSplat [1] proposes to regress Gaussian parameters from two input views, where the epipolar geometry is leveraged to learn cross-view aware features. However, it estimates depth by sampling from a predicted distribution, which is ambiguous and can lead to poor geometry reconstruction. MVSplat [2] learns to predict depth directly from the feature matching information encoded within a cost volume, which makes it more geometry-aware and leads to a more lightweight model and better geometries. Our task differentiates from MVSplat by not requiring the construction of a cost volume. This approach not only spares the need for feature matching and cost volume creation but also enhances the reconstruction speed.

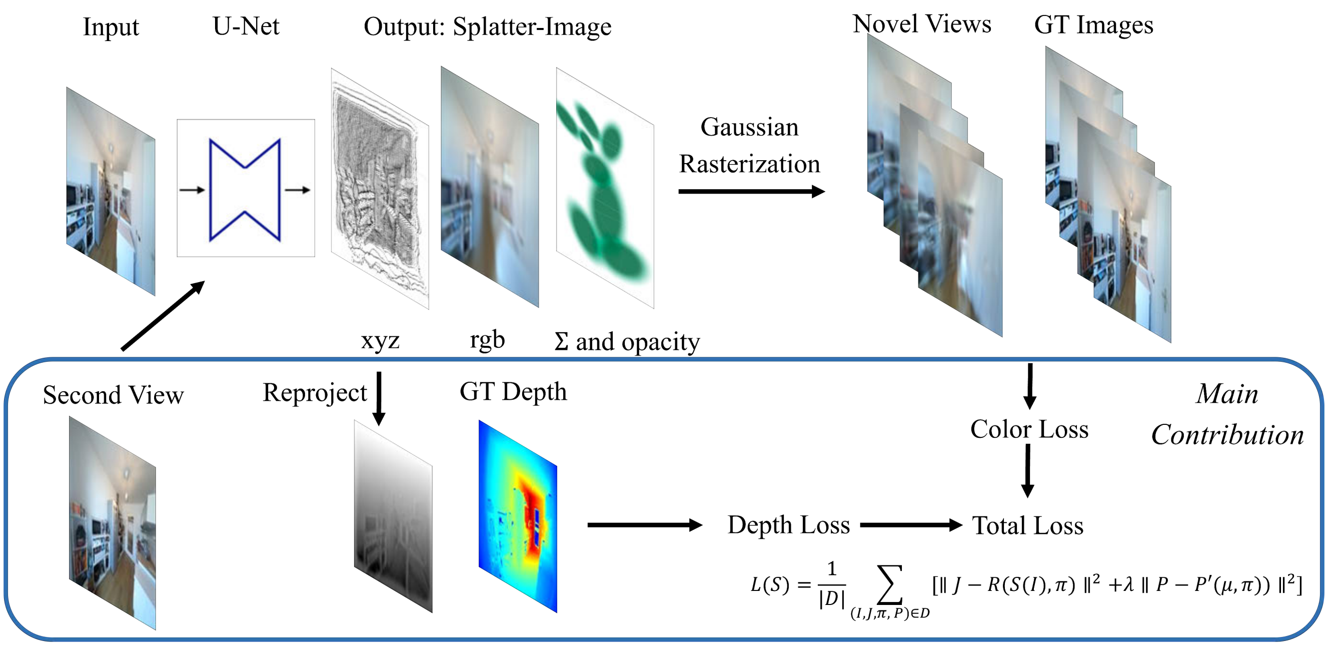

Figure 1. Pipeline of Splatter Scene. Given an $H \times W \times 3$ input scene image, we use a single U-Net to predict a $H \times W \times K$ tensor, consisting of Gaussian's parameters for each pixel, which is also called a 'Splatter Image' [7]. After that, novel views of the scene are rendered using Gaussian rasterization, and we compare them with ground truth images. The predicted depth is acquired by the reprojection of the predicted xyz position of each Gaussian, and we compare it with the ground truth depth. Through this comparison, the network learns the typical scaling of indoor scenes which solves the scale ambiguity problem when predicting real-world scenes. Moreover, a second view of the scene is used as input image to further improve the reconstruction result.

3. Method

Our proposed pipeline is shown in Figure 1.

3.1. Gaussian Rasterization

Gaussian Rasterization is a method to distribute points from a 3D space onto a 2D plane based on a Gaussian function. Specifically, each 3D point is projected onto the 2D image plane not as a single pixel, but as a 'splatter' area determined by its distance and a Gaussian distribution function. We formulate the single-image-to-3D reconstruction task as an 'inverse' of the 3D Gaussian renderer $\mathcal{R}$ [3]:

$$\theta = S(I) \tag{1}$$

in which $\theta = \{(\sigma_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, c_i), i = 1, \ldots, G\}$ is the Gaussian mixtures, I is the source image, and S is a modified SongUNet [6] network.

3.2. Depth Supervision

Based on the Splatter Image pipeline, we have integrated depth supervision. Specifically, we take the xyz positions output from the Gaussian splatter and project them onto 2D depth images given the corresponding novel view camera pose. These projected depth images are then compared with the ground truth depth images to calculate the depth loss. Therefore, the network is able to automatically rescale the predicted xyz positions and learn to predict the actual depth of indoor scenes, which in turn improves the quality of the rendered novel views.

3.3. Loss Function

Our loss function consists of three parts: the L2 RGB loss, the depth loss, and the mask loss. The RGB loss and depth loss are computed per pixel using the L2 loss. The mask loss serves as a regularization method, penalizing the network when it generates depth images with zero values. The loss function is a linear combination of average color loss and depth loss with a weight $\lambda = 0.5$ plus a regularization term:

$$\begin{aligned}\mathcal{L}(S) = \frac{1}{|D|} \sum_{(I,J,\pi,P)\in D} \Big[&\|J - R(S(I),\pi)\|^2 \\ &+ \lambda \|P - P'(\boldsymbol{\mu},\pi)\|^2 \Big] + \mathcal{L}_{mask}\end{aligned} \tag{2}$$

in which D is the dataset, J is the target image from target view, π is the view point, P is the ground truth depth, P' is the depth projector, $\boldsymbol{\mu}$ is the xyz positions of predicted Gaussians, and $\mathcal{L}_{mask}$ is the ratio of zero values in the whole projected depth map.

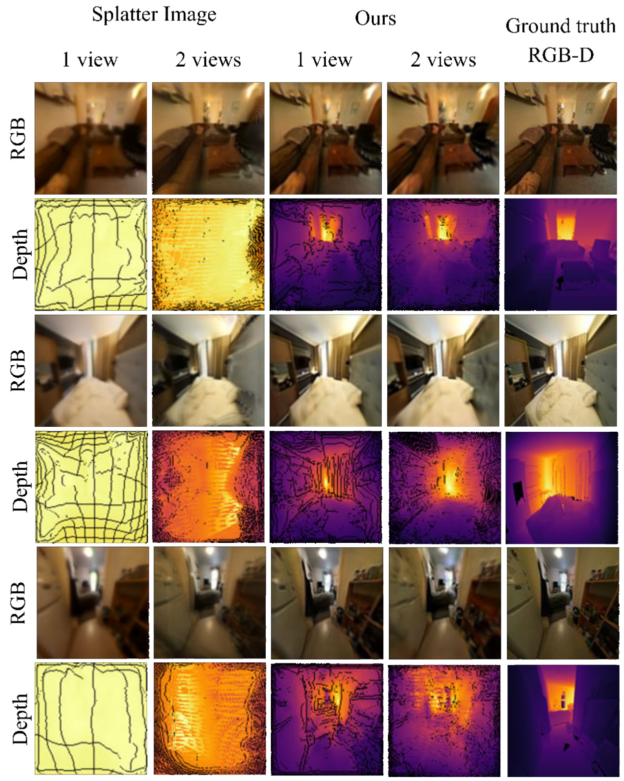

Figure 2. Qulitative results of Splatter Scene. This figure shows the original Splatter Image's rendering and predicted depth of 1 view and the first of 2 views input images, compared with our results.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓ [m]
1 view Splatter Image	17.75	0.450	0.501	1.864
1 view (ours)	**21.71**	**0.697**	**0.327**	**0.516**
2 views Splatter Image	19.77	0.657	0.389	1.702
2 views (ours)	**22.37**	**0.736**	**0.316**	**0.560**

Table 1. Quantitative results. The best results are highlighted in bold, which shows that our method outperforms Splatter Image on indoor scene novel view rendering and depth prediction

3.4. Two Images Input Prediction

For better visualization effects, we also use two image inputs for scene prediction. Given the relative camera pose between the two input views, two U-Nets each predict one set of Gaussians simultaneously and combine them for novel view rendering using the relative pose. An attention layer connects these two U-Net networks to collectively update the weights. We can also use this way of adding more input views for more than two images input.

4. Experiments

4.1. Dataset

We use ScanNet++ [8] dataset for training and evaluation. ScanNet++ contains meshes, images and poses captured from 380 indoor scenes by a laser scanner, a DSLR camera and an iPhone. We use the high-quality DSLR images, for which pixel-aligned depth maps can be rendered from the provided scene reconstruction using a laser scanner. The camera convention follows COLMAP, which means x points right, y down, and z forward (away from the camera). We also use camera-to-world poses as input during training. The original RGB images in ScanNet++ are preprocessed for training, namely undistorted from Fisheye to Pinhole camera model and resized to 128 × 128 dimensions. The rendered depth images are regularized in the unit of meter. We select the views with close camera poses as a scene for training and testing, and form a new dataset with 35 training scenes and 7 test scenes.

4.2. Implementation Details

The metrics for evaluation on image quality are pixel-level PSNR, feature-level LPIPS and patch-level SSIM. The depth precision is measured by RMSE (Root Mean Square Error) in the unit of meter. For training and testing, all experiments are conducted on an NVIDIA RTX4090 GPU with a 24GB memory size. The training iterations are all set to be 16000, which took 2.5 hours and 3 hours for single image input and two image input, respectively. For single image input, the batch size is set to 8, and for two images input the batch size is 6 because of limited GPU memory size. We use the Adam optimizer and set the learning rate to be 0.00005 for all experiments.

4.3. Quantitative Results

Since we resolve the limitation of Splatter Image using depth supervision, we compare our results with the original Splatter Image on indoor scene reconstruction effects. In Table 1, we show that for one image input and two images input, all the metrics are better than the original Splatter Image, which manifests our proposed method's effectiveness on real-world scene reconstruction ability. One thing that needs to be mentioned is that for the original Splatter Image, the generalization on unseen scenes fails quite often (4 in 7 test scenes).

4.4. Qualitative Results

In Figure 1, we show that compared with our results, the original Splatter Image's rendering and predicted depth of 1 view and the first of 2 views input images are poorer. For one thing, Splatter Image barely captures high-frequency details of the scene, whose renderings are more blurry. For

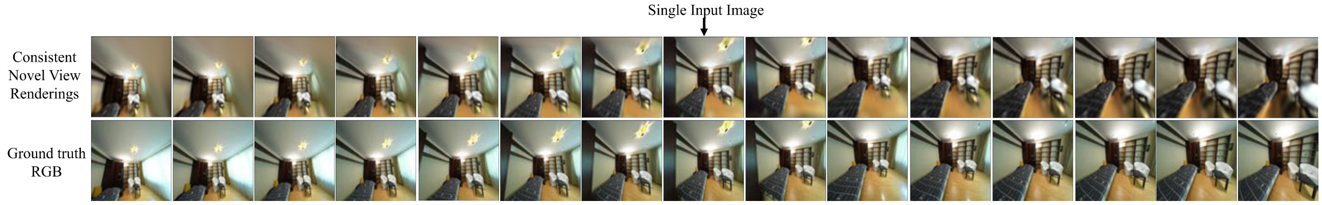

Figure 3. A demo for showing the consistent novel view rendering using Splatter Scene

another, the depth predicted by Splatter Image hardly reflects the layout of the room, which looks like just the projection of an image plain into the 3D space. However, due to a limited number of scenes in the training set, our method predicts the kitchen's layout with higher error (the sixth row) compared with a hotel room and living room, where similar layouts appear more often in the training set. For two images input, the predicted depth is more complete compared with only one view input. In Figure 3, we demonstrate that Splatter Scene can render geometrically and visually consistent novel views using only single input image.

5. Conclusion

In this work, we propose a pipeline that can reconstruct indoor scenes with only one or few input images, which resolves the scale ambiguity and inaccurate depth prediction problems with Splatter Image. We incorporate depth supervision into the training process and use two input views to improve reconstruction quality. Experimental results showcase that our method can generalize well on unseen indoor scenes with only a small number of training scenes (35), which achieves near 0.5m depth error and 22.37 PSNR novel views rendering quality.

6. Discussion

Although we have shown the effectiveness of our method on indoor scene's novel view rendering and depth prediction with single/few images, this method still has its limitations. First, our pipeline relies highly on high-quality camera pose and depth information during training, which may cause problems with robustness using lower-quality datasets. Secondly, due to resource restrictions, the number of scenes currently used in the training set is relatively small. With more scenes, the model could generalize better in highly diverse scenarios. In reality, much more data is available, but we are limited to using fewer scenes for training due to these resource constraints. Moreover, due to limited computing power and costly rendering time, the resolution for training is set to be 128 × 128, which could not meet the requirements of high-quality scenarios like industry-level applications. For two images input, the correspondence between two views is not fully exploited, which would help a lot in improving depth prediction. We are expecting future research to resolve these limitations.

References

[1] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. *arXiv preprint arXiv:2312.12337*, 2023. 1

[2] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 1

[3] 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, 2

[4] 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

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[6] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[7] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *arXiv preprint arXiv:2312.13150*, 2023. 1, 2

[8] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 3

[9] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1