SGD: Street View Synthesis with Gaussian Splatting and Diffusion Prior

Appendix

In this appendix, we provide additional details omitted from the main manuscript due to the limited space. First, we present additional figures to underscore the motivation behind our proposed method (Appendix A). Then we present more implementation details on fine-tuning Diffusion Model [6] and training 3DGS [3] (Appendix B). We also explore the influence of the Diffusion Model's prior on the generated results through dedicated experiment (Appendix C). Finally, we showcase more rendering results on the KITTI [2] and KITTI-360 [4] datasets (Appendix D).

A. Motivation

Novel View Synthesis (NVS) for autonomous driving scenarios is a challenging task. The ideal training images for both NeRF [5] and 3DGS [3] should encompass all possible perspectives of the scene, which exhibit considerable disparities with the data collected by moving vehicles. The viewpoints offered by a vehicle-mounted camera are quite constrained. Take the white car in Fig. 7 as an example, it is only observed from its side rear in the training view, causing the rendering model to overfit these viewpoints. While the current approach, such as Zip-NeRF [1], is able to render the vehicle clearly from a test view close to the training view, it produces unsatisfactory artifacts and deformation when the rendering viewpoint is shifted by a certain distance and rotated by a certain angle.

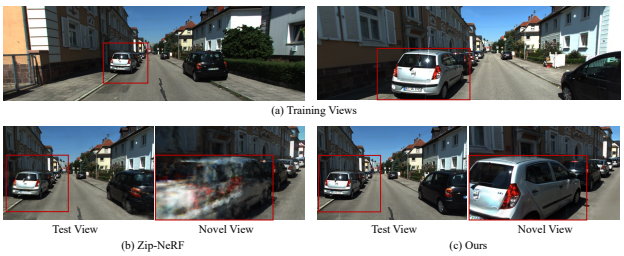


Figure 7. An example of how the current method [1] overfits the training views, while our method overcomes this problem.

B. More Implementation Details

Diffusion Model. Our Diffusion Model is adapted from Stable Diffusion 1.5 [6] and is fine-tuned on about 12,000 images with 512×512 resolution from the KITTI-360 [4] dataset. Considering the original size of KITTI-360 images is 1408×376, a preliminary cropping step to 600×376 is performed before the resizing, to avoid over-distorting the images. We conduct *center-crop* on the training images. For the reference images, we use *random-crop* during the training process, which could ensure a certain perspective gap exists between the reference image and the training image, so as to enhance the robustness of the model. During inference, the reference images are pre-processed with *center-crop*.

When selecting the reference images, we randomly choose one image from the five frames preceding the training image and one from the five frames succeeding the training image separately. During inference for the novel viewpoint, we identify its closest training viewpoint and utilize its adjacent frames as reference images. Regarding the depth maps, due to the limitation of LiDAR point clouds in capturing the scene above a certain height, we apply a mask to the top 80 rows of pixels in the images. In practice, we found that the inpainting capability of the Stable Diffusion Model is effectively able to complete this portion of content. To enable classifier-free guidance in the first training stage, we set both text prompts and reference images to be empty with a 10% probability.

3D Gaussian Splatting. We only initialize the 3D Gaussian models with LiDAR point cloud. The detailed procedure involves first projecting the LiDAR frame onto its corresponding image frame to assign a color to each LiDAR point. Then these points are re-projected into 3D space, creating colored 3D point clouds. Finally, all frames of point clouds are accumulated and then voxel-downsampled with the voxel-size of 5. We train both our model and the baseline 3DGS model for 50,000 iterations. We first train the model for 500 iterations without sampling pseudo views for adequate warm-up. Subsequently, for every 10 iterations, 4 pseudo views are sampled for training.

Figure 8. The impact of the strength of the Diffusion Model's prior on the generated result. *(c) is a novel view, its original image is rendered by 3DGS.

C. Additional Experiment

Additional Ablation Study. As described in Sec. 3.2 of the main manuscript, during the training stage of 3DGS, we render some randomly sampled pseudo views, and utilize a fine-tuned Diffusion Model to generate guidance images for these views to regularize the training. Specifically, the pseudo view rendered by 3DGS is passed through the VAE Encoder to obtain a latent feature map, to which noise at level t is added, where $t \sim [t_{\text{min}}, t_{\text{max}}]$. This noised latent feature is denoised by the Diffusion model from level t to t_{min}, and then it is decoded to obtain the generated image. Specifically, we set $t_{\text{max}} = 10$, and employ a hyper-parameter s, which indicates strength, to control the noise level t, according to $t = s \times t_{\text{max}}$.

In Fig. 8, we show the results of ablation experiments on hyper-parameter s. The first column labeled with *original image* refers to the image being fed into the Diffusion Model, while the generated image with hyper-parameter s increasing from 0.2 to 0.8 is exhibited in the other columns. It can be observed that a smaller s makes generated images more similar to the original image, while a large s introduces higher diversity and deviation in details. For novel viewpoints in Fig. 8 (c), smaller s makes the generated image preserve noise rendered by 3DGS. As s increases, the image becomes cleaner but loses some details. In practice, we randomly select $s \sim [s_{\text{min}}, s_{\text{max}}]$ for each sampled pseudo view, where $s_{\text{min}} = 0.2$, s_{max} starts at 0.6 and decreases to 0.4 over the training process. This strategy guarantees when 3DGS-rendered images are of lower quality in the early stage of training, our model relies more on the guidance from the Diffusion Model's prior. Accompanied by the quality of 3DGS renderings improves with ongoing training, it is necessary to reduce the impact of the Diffusion Model-generated images on the details.

	KITTI_0009-10%		
	PSNR↑	SSIM↑	LPIPS↓
Street Gaussians [7]	17.99	0.700	0.195
Ours w/ Street Gaussians	19.01	0.754	0.174

Table 5. Quantitative results on KITTI dataset with sparse view input.

Figure 9. Qualitative results on KITTI_0009 sequence with sparse view input.

Additional Experiments with Sparse View Input. We have added more experiments to demonstrate the effectiveness of our method under sparse view inputs. Fig. 9 and Tab. 5 show the qualitative and quantitative results with only 10% input for one sequence in KITTI. It can be observed that when the input views are very sparse, our method produces fewer artifacts compared to Street Gaussians [7]. This is because our fine-tuned Diffusion Model, can generate images that are highly faithful to the original scene given the

reference image and LiDAR depth. When the number of input views decreases, increasing the number of sampled pseudo views allows the training process to rely more on the priors of the Diffusion model. However, this comes with certain trade-offs, for example, in the second row of Fig. 9, the billboard in the top right becomes blurry as the Diffusion Model is less sensitive to text.

D. More Rendering Results

We provide more novel view rendering results of our method and our competitors [1, 3] on the KITTI [2] and KITTI-360 [4] datasets in Fig. 10.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 1, 3

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. The kitti vision benchmark suite. *URL http://www. cvlibs. net/datasets/kitti*, 2(5), 2015. 1, 3, 5

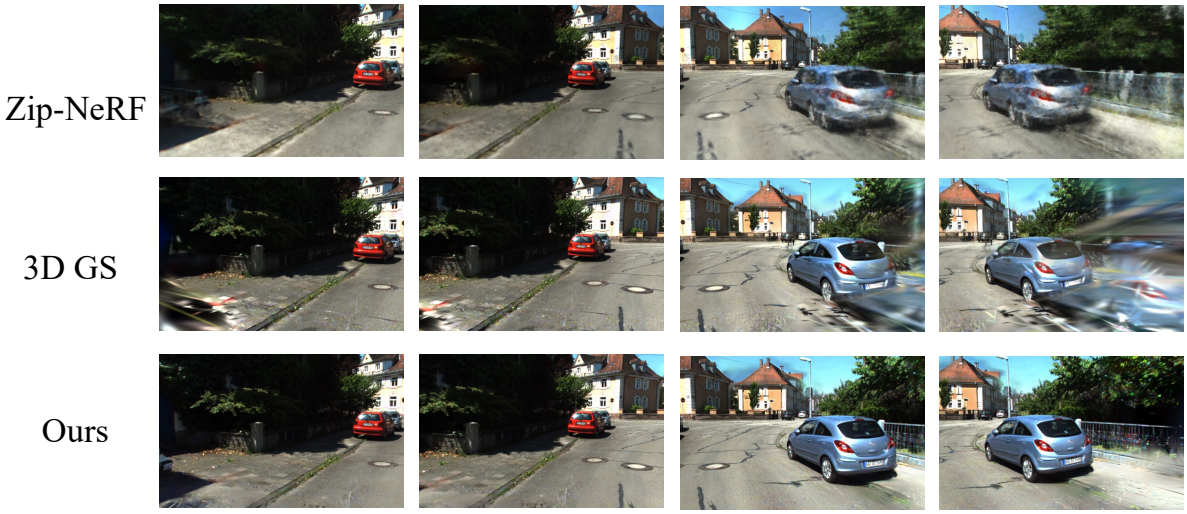
[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 3

[4] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 1, 3, 5

[5] 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

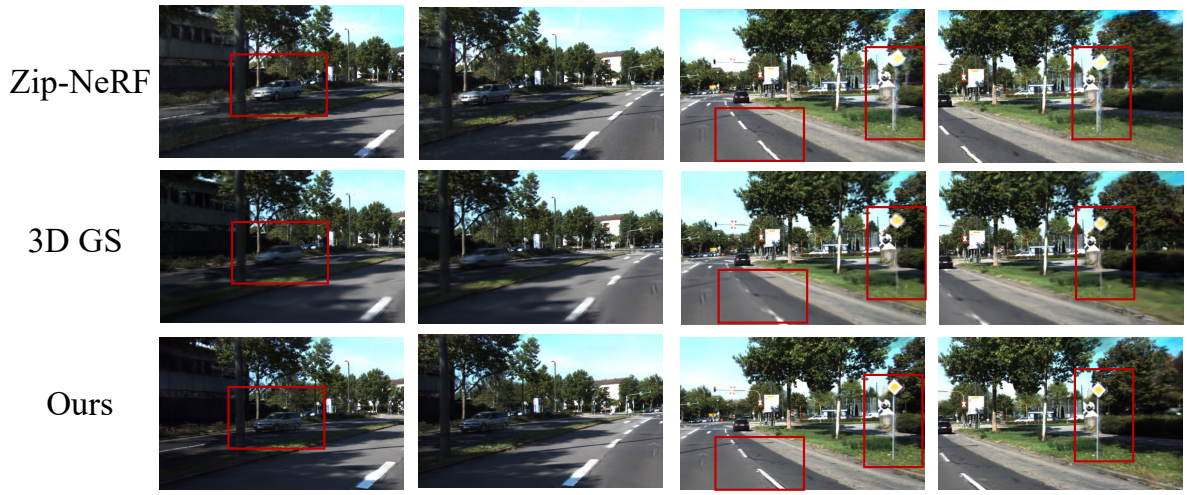
[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[7] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 2



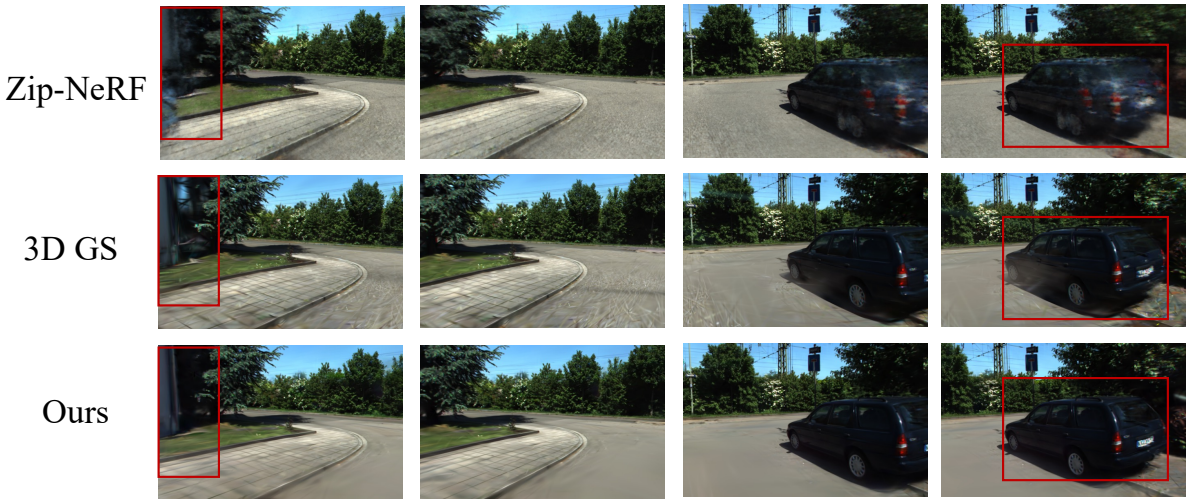
(a)

(b)

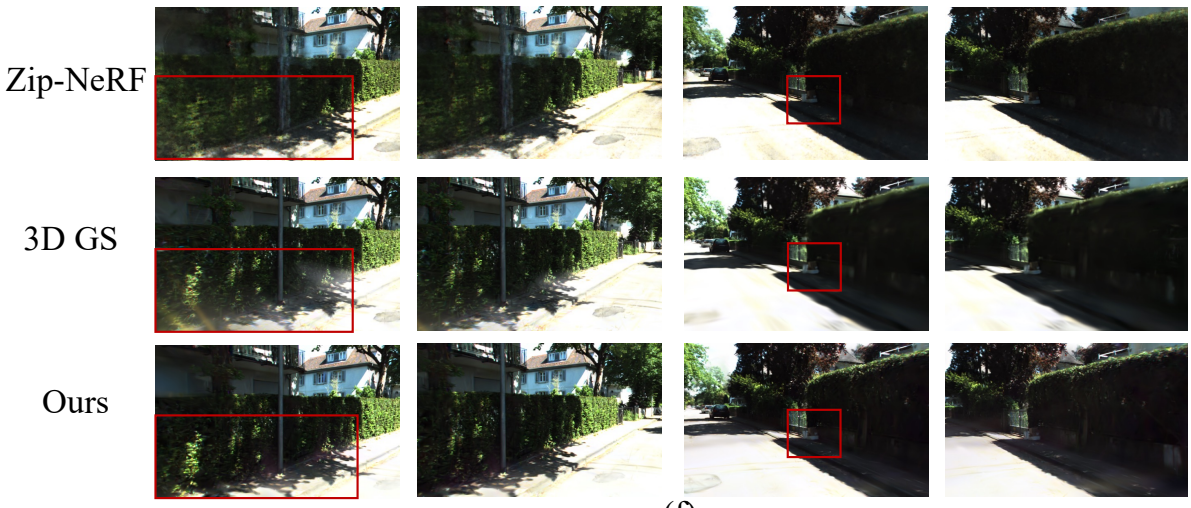


(c)

(d)



(e)



(f)

Figure 10. More qualitative results of novel views rendering on KITTI [2] and KITTI-360 [4].