Novel View Synthesis Under Large-Deviation Viewpoint for Autonomous Driving

Xin Ma[1*], Jiguang Zhang[2*], Peng Lu[1†], Shibiao Xu[1] , Chengwei Pan[3]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Beihang University, Beijing, China
{maxin034, lupeng, shibiaoxu}@bupt.edu.cn, jiguang.zhang@ia.ac.cn, pancw@buaa.edu.cn

Abstract

Novel view synthesis is a critical task in autonomous driving. Although 3D Gaussian Splatting (3D-GS) has shown success in generating novel views, it faces challenges in maintaining high-quality rendering when viewpoints deviate significantly from the training set. This difficulty primarily stems from complex lighting conditions and geometric inconsistencies in texture-less regions. To address these issues, we propose an attention-based illumination model that leverages light fields from neighboring views, enhancing the realism of synthesized images. Additionally, we propose a geometry optimization method using planar homography to improve geometric consistency in texture-less regions. Our experiments demonstrate substantial improvements in synthesis quality for large-deviation viewpoints, validating the effectiveness of our approach.

Introduction

Novel View Synthesis is an advanced computer vision technique that reconstructs scenes from multiple views. It holds particular significance in autonomous driving, enabling the generation of diverse driving scenarios and viewpoints from real-world data. This capability enhances the diversity and realism of training datasets for autonomous driving systems, reducing reliance on costly data collection. Additionally, novel view synthesis improves system adaptability to various road conditions and lighting environments, ultimately contributing to the safety and robustness of autonomous driving technology.

In recent years, 3D Gaussian Splatting (3D-GS) (Kerbl et al. 2023) has achieved significant advancements in novel view synthesis, particularly in enhancing image quality. This paper focuses on applying 3D-GS to synthesizing views in autonomous driving scenarios. Research in this field can generally be categorized into two areas: geometry-based methods leverage geometric priors to optimize the scene's structure; and implicit network-based methods, which integrate implicit representations into 3D-GS to learn continuous feature representations for improved view quality. In general, current novel view synthesis methods can produce high-quality images for viewpoints that are present in the training set.

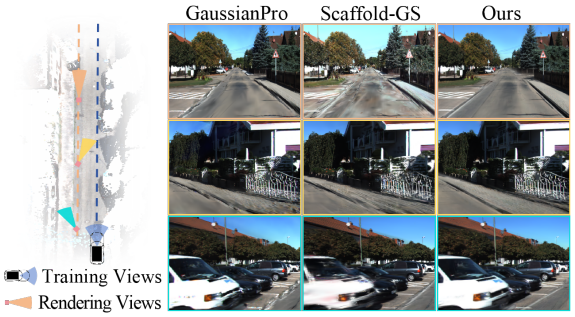

Figure 1: **Large-Deviation View Synthesis.** *(Left)* Given input images and camera poses for the right lane (blue), our method generates photorealistic, 3D-consistent images of the left lane (orange). *(Right)* Our approach preserves photorealistic rendering at viewpoints distant from the training views, while other methods produce severe artifacts.

However, current methods face challenges in maintaining high-quality rendering when viewpoints deviate significantly from the training data. As illustrated in Figure 1 (left), to train vehicles for overtaking maneuvers, our goal is to generate a series of images along an overtaking trajectory (orange lane), which are not included in the training data. We select three typical viewpoints along this trajectory: a translation, a leftward rotation, and a rightward rotation relative to the current trajectory. As shown in Figure 1, the quality of the synthesized images noticeably degrades, with road details becoming distorted and the front of the vehicle appearing blurry.

Several factors contribute to the observed issues, with two key factors being particularly significant.

First, most novel view synthesis models based on 3D-GS utilize spherical harmonics to model scene lighting. Due to the limited viewpoints in the training data, the spherical harmonics coefficients effectively capture the lighting field only near these viewpoints, often leading to overfitting. As a result, when synthesizing views that deviate from the training viewpoints, the colors generated by the spherical harmonics coefficients can diverge from the true values, reducing real-

*These authors contributed equally.
†Corresponding Author.

ism and quality in the synthesized views, as illustrated by the blurriness in Figure 1.

Second, street scenes often contain large, texture-less regions, such as wide roads, where the number of extractable feature points is typically limited. The point clouds generated through Structure from Motion (SfM) methods are also sparse, requiring the use of large-scale Gaussians to approximate the geometric structure in these regions. Due to limited viewpoint supervision, these large-scale Gaussians frequently overfit the available data, leading to significant discrepancies between the geometric representation and the actual structure. While this inconsistency has a minor impact on view quality near the training viewpoints, it greatly reduces the fidelity of synthesized images when the viewpoints deviate significantly from training data.

To address above challenges, this paper proposes two components to enhance the performance of existing 3D-GS in handling large-deviation viewpoints rendering.

To overcome the limitations of current lighting models, we propose an improved illumination model that uses attention mechanisms to aggregate light field information from neighboring views. This method synthesizes color values for Gaussians at specific viewpoints, replacing traditional spherical harmonics. Specifically, we first learn feature vectors for each Gaussians to capture its unique characteristics. Then, we select neighboring images that are spatially close to the target viewpoint as reference views. From these images, candidate pixels are chosen based on geometric consistency. Aggregation weights are computed using the feature vectors, allowing us to effectively combine the color values of these candidate pixels to produce the synthesized color.

For the geometric structures of texture-less regions, we propose a training sample augmentation method based on homography matrices. This method enhances the quality of synthesized views by incorporating diverse perspective images of texture-less regions. Specifically, given that texture-less regions are often planar, we first employ a deep learning model to detect planes in 3D space. Using homography matrices, we then calculate the normal and color information of these planar regions from novel viewpoints, based on existing training samples. This information is subsequently used to guide the network in optimizing the shape of Gaussians, thereby improving the final rendering quality.

The main contributions of this paper are as follows:

- We introduce a novel illumination model for 3D-GS that estimates lighting information from neighboring images, effectively alleviating the limitations of spherical harmonics when synthesizing views that significantly deviate from the training data.
- We propose a geometry optimization method based on planar homography transformations, which enhances geometric consistency in texture-less regions such as roads and walls.
- Extensive experiments on public datasets demonstrate that the proposed methods significantly improve the synthesis quality of existing 3D-GS models, particularly for viewpoints that deviate substantially from the training set.

Related Work

The rapid advancement of novel view synthesis techniques, including 3D Gaussian Splatting (3D-GS), has attracted significant attention. Current research in this field can be broadly divided into two categories: geometry-based methods, which utilize geometric priors to optimize scene structure, and implicit network-based methods, which integrate implicit representations within 3D Gaussian Splatting (3D-GS) to learn continuous features, thereby enhancing the quality of synthesized views.

Several methods have been proposed to leverage 3D scene geometry priors to guide novel view synthesis. (Du et al. 2023) learns powerful multi-view geometry priors to render images from stereo image pairs. Some approaches address the problem of geometry inconsistency by employing holistic geometric regularization techniques (Deng et al. 2022). Depth information reduces the effort of inferring geometry through color consistency across multiple images (Johari, Lepoittevin, and Fleuret 2022; Somraj and Soundararajan 2023; Wang et al. 2023a). The supervised depth can also be achieved by using dense depth data obtained from additional sensors (Azinovic et al. 2022; Dey, Ahmine, and Comport 2022) or by exploiting estimated dense depth from pre-trained networks (Neff et al. 2021; Prinzler, Hilliges, and Thies 2023; Roessle et al. 2022).

Above methods require processing a large number of depth information. Particularly in high-resolution scenes, storing depth data and the corresponding color information consumes a large amount of memory. 3D Gaussian splatting (Kerbl et al. 2023) can partially retain the ability to reconstruct detailed local features with only sparse inputs, due to a free depth guide that can be obtained from sparse point clouds. (Xiong et al. 2023) leverages approximate splat positions as guidance. This approach aids in stabilizing optimization in few-shot scenarios and helps eliminate floating artifacts that occur at random locations. (Li et al. 2024) utilizes pre-trained models and efficient inference processes to achieve 3D reconstruction and novel view synthesis.

Implicit network-based methods have significantly advanced novel view synthesis. Some studies achieve novel view synthesis based on neural network to encode local radiance values. (Chen et al. 2025) based on 3D-GS builds a cost volume representation via plane sweeping that provide valuable geometry cues to render novel views. (Wewer et al. 2025) integrates a regression-based approach with a generative model, that efficiently encodes varying uncertainty within a latent space consisting of 3D feature Gaussians. Some methods are capable of generating highly realistic unseen viewpoints by learning the distribution of existing views. (Tseng et al. 2023) proposes a posed-guided diffusion models to generate a consistent novel views from a single image. (Guo et al. 2024) builds a 3D semantic network that directly predicts the semantic component from raw 3D Gaussians for fast inference.

The aforementioned methods are capable of generating high-quality images for novel viewpoints that are close to those in the training set. However, these methods fail to effectively render views when the viewpoint significantly deviates from the training viewpoints. This paper primarily fo-

cuses on addressing the issue of synthesizing high-quality images for viewpoints that are significantly distant from the training perspectives, and the proposed method can be widely applied to the aforementioned techniques to enhance their synthesis capabilities for such viewpoints.

Method

To achieve high-fidelity scene rendering in outdoor environments with complex structures, even when viewpoints significantly deviate from the training views.

As shown in Figure 2(a), we propose a illumination model based on epipolar geometry. The ResNet-50 is used to encode multi-scale features from the training views. Then, the visible Gaussians are projected onto the target viewpoint to compute epipolar lines on the adjacent feature maps. Sampling points are uniformly selected along these epipolar lines across multi-scale feature maps. Finally, a light field decoder is employed to decode the attributes of the Gaussians.

As shown in Figure 2(b), for the geometry optimization of texture-less regions based on planar homography, a plane encoder is used to extract real plane regions and plane parameters. Then the computed homography matrix is utilized to transform the plane regions to target view and supervise the training process. The plane regions contains RGB maps, and normal maps that extracted by (Huang et al. 2024) to supervise Gaussians. Subsequently, we will provide a detailed explanation of these two models.

Illumination Model Based on Epipolar Attention

Most 3D-GS based view synthesis models employ spherical harmonics for scene illumination modeling. With limited training viewpoints, spherical harmonics can only effectively capture the light field near the trained viewpoints. However, synthesizing views far from the training viewpoints tends to nearly random spherical harmonic coefficients, resulting in inaccurate illumination and degraded realism and quality of the generated views.

One approach to build an illumination model involves interpolation from neighboring images. However, this method can result in the model becoming trapped in local optima. To overcome this limitation and establish a more generalizable light field across different views, we propose an epipolar geometry based illumination method. By incorporating multi-view encoding, the method uses attention mechanisms to select the most likely color values, ensuring robust color consistency across various viewpoints.

Instead of relying on spherical harmonics, our method synthesizes the color of a point using light field obtained from the surrounding views. During the projection of a point onto neighboring views to sample its color, inaccuracies in scene geometry often introduce errors. To mitigate this, we employ epipolar line sampling to learn the features of Gaussians within a specific range and aggregate values along these lines to determine the final color. Throughout the aggregation process, attention mechanisms are crucial for learning feature representations that accurately correlate Gaussians with image pixels. This process ultimately leads to the more accurate Gaussians color for the synthesized viewpoint.

Voxelize Point Cloud and Projection. As shown in Figure 2, we begin with the sparse point cloud generated by COLMAP as the initial input. Then we voxelize the scene derived from this point cloud P using the equation: $G = \left\lfloor \frac{P}{\varepsilon} \right\rfloor \cdot \varepsilon$, where G represents the voxel centers, and ε denotes the voxel size. The center of each voxel $g \in G$ is treated as an anchor point in world coordinates, which is projected to the target viewpoint to calculate the currently visible Gaussians $\hat{G}$. Each visible Gaussian then activates neural Gaussians for all anchors visible within the frustum (Lu et al. 2024).

Unlike previous scene representations that rely on spherical harmonics to encode Gaussians appearance, our approach captures the light field from nearby viewpoints by leveraging epipolar geometry. To obtain a set of effective context views, we first identify spatially closest N views as context views which are spacial closest to the target view.

Multi-View Encoder. We observe that directly embedding epipolar geometry into the context images leads to inconsistency in geometric reconstruction. To resolve this, we introduce a multi-view encoder that simultaneously processes context images along with their relative poses. We employ ResNet50 to extract the multi-scale feature maps $\mathbf{f} \in R^{H_j \times W_j \times d}$ from each context view.

Epipolar Line Sampling. Subsequently, given a large-deviation viewpoint at position x_c and the camera projection matrix $P_k \in R^{3\times4}$, we calculate each Gaussian projected coordinate $\mathbf{u}_i = (u, v)$ on image plane. The set of tuples $\{(g_i, \mathbf{u}_i)\}_{i=1}^{M}$ forms the visible Gaussians and corresponding projected coordinate for the target view. Given intrinsic parameters K_t and extrinsic parameters $W_t = [R_t, t_t]$ relative to the context camera I_t. The epipolar lines $\mathbf{l}_i$ in the context camera is expressed as:

$$\mathbf{l}_i = \mathbf{F}_i[u, v, 1]^\top$$

$$\mathbf{F}_i = K_i^{-\top}([t_t]_\times R_t)K_t^{-1}$$

where $\mathbf{F}_i$ refers to the fundamental matrix.

Based on the established epipolar relationship between target view and context views, we sample feature point located at the epipolar line $\mathbf{l}_i$ of context views for a given pixel in the target view. For the tuples $\{(g_i, \mathbf{u}_i)\}_{i=1}^{M}$ of visible Gaussians under target view, we uniformly sample D pixel coordinates $\mathbf{v}_s^k = (u_s^j, v_s^j)_{j=1}^{D}$ along the epipolar line of multi-scale context features $\mathbf{f} = \{f_i, f_{i\downarrow1}, f_{i\downarrow2}, f_{i\downarrow3}\}$, where $\downarrow n$ denotes f_i being down-sampled by 2^n factors. We now obtain sampled feature $F_\mu \in R^{4\times D\times M\times N}$ for visible Gaussians under target view, where N denotes the number of context views closest to the target view.

For clarity, we restrict epipolar line sampling to the multi-scale feature maps of the two views closest to the target view, sampling 32 points on each map. This approach is not only computationally efficient but also ensures that the sampled features are more relevant and accurate. The closest views provide the most similar perspectives, resulting in more consistent information and reducing potential errors from larger viewpoint deviations.

Light Field Decoder. We propose to augment association between the sampled features of context views and target

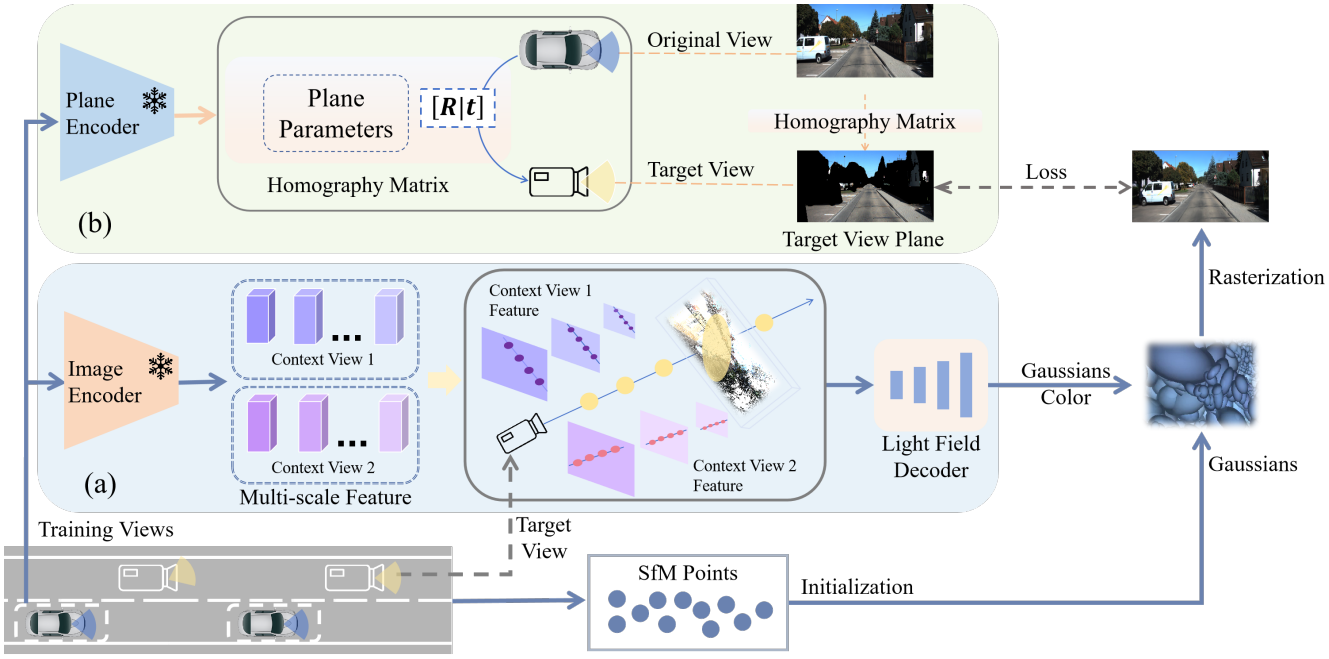

Figure 2: There are two models in our proposed large-deviation view synthesis method. (a) Firstly, the ResNet-50 is employed to encoder multi-scale features for training views. Then, the visible 3D point clouds are projected to target viewpoint and compute the epipolar line on the adjacent feature maps. Next, sampling points uniformly on these epipolar lines in multi-scale feature maps. Then, light field decoder is utilized to decode the attributes of Gaussians. (b) A geometry optimization method specifically designed for texture-less planar regions. The homography matrix is utilized to transform the plane regions to target view and supervise the training process.

view via a lightweight cross-attention decoder. Given a target camera at position $\mathbf{x}_t$, a sample coordinate of target view $\mathbf{u}_t$ and a context camera at position $\mathbf{x}_c$, we calculate their relative distance with $\delta_{ct} = \| \mathbf{x}_t - \mathbf{x}_c \|$ and the target ray viewing direction with:

$$\overrightarrow{d_{ct}} = \frac{\mathbf{x}_t - \mathbf{x}_c}{\| \mathbf{x}_t - \mathbf{x}_c \|_2}$$

The target camera position, target view, and context camera are embedded into a query token q using a shallow MLP, defined as $Q = \Phi([\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_c, \delta_{ct}, \overrightarrow{d_{ct}}])$. The context epipolar features F_μ are transformed into key tokens $K = \Theta(F_\mu)$ using an MLP, while the value terms $V = \Gamma(F_\mu)$ are computed with a separate MLP applied to the context epipolar samples. The weight matrix is estimated using $A'_{i,j} = \frac{QK^T}{\sqrt{d_2}}$. The output of the epipolar attention layer is decoded with a shallow MLP Ψ as:

$$\hat{f}^t = \Psi\left(\text{softmax}(A'_{i,j}) \cdot V\right) \in R^{3\times M}$$

Finally, the color of the visible Gaussians is assigned by the output of the epipolar attention feature $\hat{f}^t \in R^{3\times M}$, which denotes the RGB values of Gaussians. And these Gaussians are utilized to render the novel viewpoint deviated significantly from the training data.

Geometric Optimization of Texture-less Regions Based on Planar Homography

Street scenes often contain large, texture-less regions, such as wide roads, where the number of extractable feature points is typically sparse. As a result, the point clouds generated by Structure-from-Motion (SfM) methods are also sparse, leading to the larger scale of Gaussians to approximate the geometry structures in these regions. Due to insufficient viewpoint supervision, these large Gaussians often remain underfit, creating significant inconsistency between the geometric representations and the true structures. While these inconsistency have minimal impact on the quality of synthesized views when close to the training viewpoints, they significantly affect the fidelity of synthesized images when the viewpoint deviates substantially from those used during training.

To address this issue, we propose a geometry optimization method specifically for texture-less plane regions. Our approach integrates 3D Gaussians with 2D images and normal maps, using plane representations to guide the geometry of texture-less regions. Specifically, by utilizing homography matrices to establish correspondences between known and unknown trajectory views, this correspondence enables the construction of normal and view information for the unknown trajectory views, guiding the network to optimize the shape of Gaussians, thereby improving the final rendering quality.

Plane Encoder and Parameters Estimation. As depicted in Figure 2(b), to estimate plane regions given the camera pose, we decode plane parameters using a unified plane recovery model called PlaneRecTR (Shi, Zhi, and Xu 2023). It is an advanced framework for 3D plane recovery from a single view, specifically designed to accurately estimate plane parameters. The model operates by leveraging plane parameters, query vectors, and a global transformation relative to the camera coordinate system. For each input viewpoint, it outputs n real plane regions $\pi_i (i = 1, ..., n)$, along with the normal vectors and distances from the camera origin for each plane.

Homography Matrix. The 3D real plane π_i in the camera coordinate system is represented by the parameters $(d_i, \mathbf{n_i})$, where $\mathbf{n_i}$ denotes the normal vector of the plane, and d_i represents the distance from the camera origin to the plane. The homography matrix, which represents the transformation from the original camera pose to a novel large-deviation viewpoint, is expressed as:

$$H_{\text{o}\to\text{n}} = K(R + tn_d^T)K^{-1}$$
$$n_d^T = \frac{\mathbf{n_i}}{d_i}$$

where K is the camera's intrinsic matrix, R denotes the rotation matrix, and t represents the translation vector that maps the transformation between the two camera poses.

Normal and RGB Transformation. After obtaining the homography matrix, the plane content under the original pose P_{ori} is transferred to the large-deviation viewpoint P_{new}. To model the geometric consistency, the normal map corresponding to the plane coordinates is also transformed to the large-deviation viewpoint using the homography matrix. The transformation of the RGB values and normal vectors to the large-deviation viewpoint can be expressed as:

$$\mathbf{I}_{new} = H_{\text{o}\to\text{n}} \cdot \mathbf{I}_{ori}$$
$$\mathbf{N}_{new} = H_{\text{o}\to\text{n}} \cdot \mathbf{N}_{ori}$$

This transformation ensures that the normal map rendered in the novel pose is accurately constrained. Moreover, by applying the homography matrix, the normal and RGB values are consistently projected to the large-deviation viewpoints. This unified transformation guarantees consistency across a large region over different viewpoints. And these large-deviation views and corresponding normal maps are utilized to supervise the rendering process, which resulting in accurate and reliable guidance under the large-deviation camera pose.

Loss Function. We now have a rendered image from a novel and large-deviation camera viewpoint. Our loss function comprises three components: the image reconstruction loss L_1, the $L_{\text{D-SSIM}}$ loss within 3D-GS, and the normal constraint loss:

$$L = (1 - \lambda)L_1 + \lambda L_{\text{D-SSIM}} + L_{\text{normal}}$$

We explicitly enforce consistency between the rendered Gaussian normals and the plane normals using L_1 loss and angular loss, defined as L_{normal}:

$$L_{\text{normal}} = \sum_{p \in Q} \left(\|\hat{N}(p) - \bar{N}(p)\|_1 + \left(1 - \hat{N}(p)^\top \bar{N}(p)\right) \right),$$

Figure 3: Qualitative comparison of synthesized images focusing on the road region when translating viewpoints beyond the original trajectory on the KITTI dataset.

where $\hat{N}$ is the rendered normal map, $\bar{N}$ is the transferred plane normal map, and Q denotes the set of valid pixels after visible filtering.

Experiments

Experimental Setup

Datasets. We conduct experiments on two large-scale urban datasets: KITTI (Geiger, Lenz, and Urtasun 2012) and Waymo Open Datasets (Sun et al. 2020). We follow (Lu et al. 2023; Cheng et al. 2024) to select 6 scenes for each dataset, mainly containing static objects. As common practice for evaluation novel view synthesis performance, we select every 8th image in the sequences as the test set and the remaining images as the training set.

Metrics. For real street scene scenarios, the selected synthesis viewpoints along new trajectory do not have corresponding ground truth images. Therefore, we use the perceptual metrics: Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Sutherland, Arbel, and Gretton 2018) to evaluate the quality of image synthesis for viewpoints that deviate significantly from the training data. These metrics are well-established in evaluating the visual quality of generative models and provide insights into the fidelity of the synthesized images.

Compared methods. We compare our approach with recent methods. The 3D-GS (Kerbl et al. 2023) represents the scene using a set of anisotropic Gaussians that can be efficiently rendered via rasterization. Scaffold-GS (Lu et al. 2024) introduces anchor points to manage local 3D Gaussians and predicts their attributes to enhance scene coverage. To validate the advantages of our method in handling texture-less regions, we compare it with the GaussianPro (Cheng et al. 2024). GaussianPro leverages classical multi-view stereo (MVS) techniques to guide the densification of 3D Gaussians, particularly on texture-less surfaces. To distinguish our method from view interpolation method within the observed range, we compare it with the sparse-view-based FSGS (Zhu et al. 2025). It introduces a proximity-guided Gaussian unpooling mechanism, specifically designed for sparse-view settings to address the challenges of sparse initial point sets. In addition, we evaluate our method with approach F^2-NeRF (Wang et al. 2023b), based on neural radiance fields that support free camera trajectories.

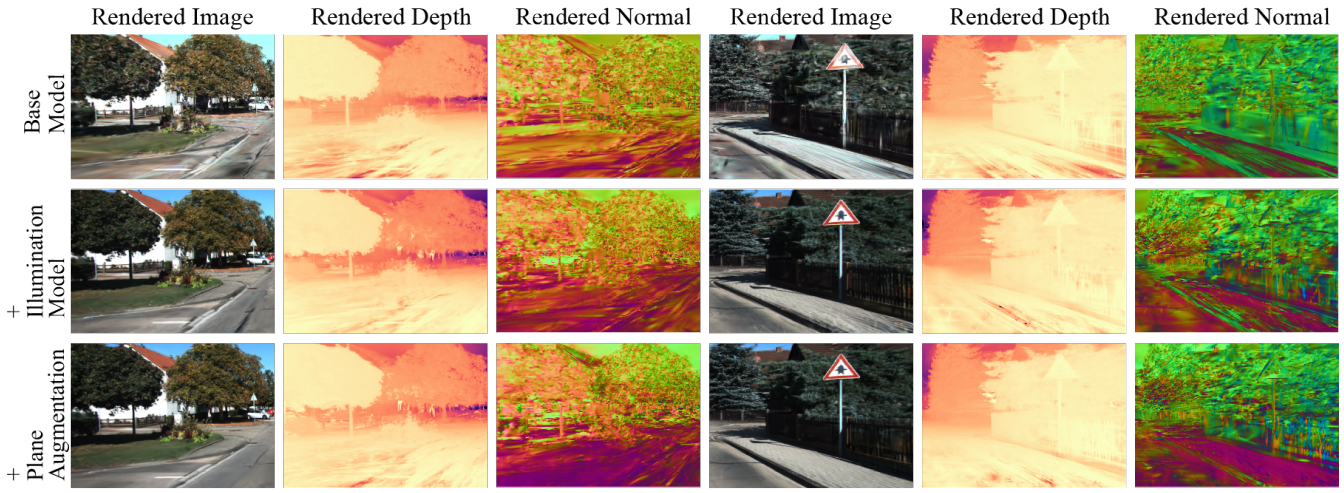

Figure 4: Visualization of rendered images, depth maps, and normal maps on the KITTI dataset. The visualization method is referenced by (Cheng et al. 2024)

Figure 5: Qualitative comparison of synthesized images focusing on the roadside when translating viewpoints beyond the original trajectory on the KITTI dataset.

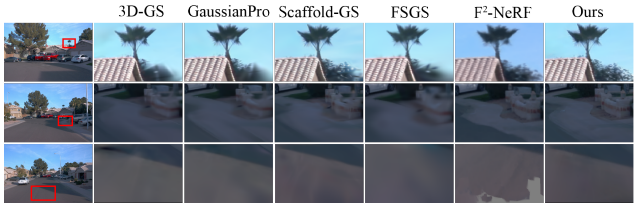

Figure 6: Qualitative comparison of synthesized images from viewpoints translated beyond the original trajectory on Waymo Open Dataset.

	KITTI		Waymo	
Method	KID↓	FID↓	KID↓	FID↓
3D GS	79.98	0.0104	76.73	0.0052
GaussianPro	90.04	0.1225	83.35	0.0030
Scaffold-GS	98.24	0.0251	73.45	0.0034
FSGS	101.26	0.1572	90.39	0.0083
F^2-NeRF	85.25	0.0785	79.28	0.0069
Ours	**71.58**	**0.0064**	**63.58**	**0.0015**

Table 1: Evaluation of KID and FID metrics on static scenes from the KITTI and Waymo Open datasets with viewpoints shifted beyond original trajectory.

Comparisons on Three Typical Viewpoints

In order to train vehicles to perform overtaking scenarios, our goal is to synthesize a series of images along an overtaking trajectory. We select three typical testing viewpoints, such as: a translation, a leftward rotation, and a rightward rotation relative to the current trajectory.

First, the camera is positioned with a large translation, the performance comparisons are presented in Table 1 and 2. Our method shows better reconstruction performance than other baseline methods. The superior performance validates the robustness of our method to handle complex lighting conditions and texture-less geometry. We visualize the qualitative comparison in Figure 3, GaussianPro, Scaffold-GS and FSGS exhibit noticeable road distortion, and F^2-NeRF lacks geometric details. It is also shown in Figure 5, baseline methods exhibit excessive blurriness and distort the edges of the road and car. In contrast, our method produces higher-quality results, such as rendering wide roads with fewer blurs and artifacts (last column). It can be attributed to that other methods focus on interpolating viewpoints within the observed range. In contrast, our task requires extrapolation, predicting viewpoints beyond this range, which further explains the underperformance of baseline methods.

As demonstrated in Figure 6, we also present results of reconstructing images for viewpoints translated beyond the original trajectory on Waymo Open Datasets. GaussianPro and Scaffold-GS fails to synthesize views with scene details, though they attempt to reconstruct the plane region. It is evident that our method consistently produces reasonable rendering results for large-deviation viewpoints (last column), while other methods resulting in severe blurring and artifacts. In addition, our method achieves a rendering speed of 83 FPS, compared to the baseline methods GaussianPro at

Method	KITTI		Waymo	
	FID↓	KID↓	FID↓	FID↓
3D GS	102.43	0.0195	91.79	0.0111
GaussianPro	102.56	0.0223	111.88	0.0204
Scaffold-GS	112.24	0.0313	83.65	0.0122
FSGS	119.90	0.0395	116.91	0.0231
F^2-NeRF	105.82	0.0326	93.91	0.0189
Ours	**87.78**	**0.0138**	**68.23**	**0.0050**

Table 2: Compared with the results in Table 1, a larger translation deviation was applied to both datasets. KID and FID were evaluated on static scenes from viewpoints shifted beyond the original trajectory.

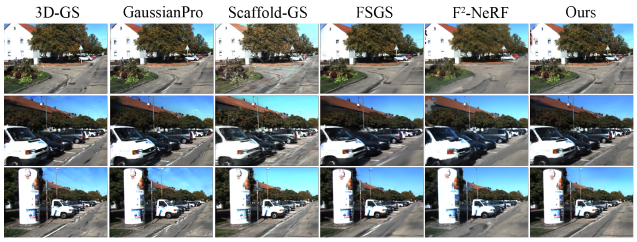

Figure 7: Qualitative results under $45°$ view rotation on KITTI Dataset.

Method	KITTI	
	KID↓	FID↓
3D GS	151.53	0.0599
GaussianPro	148.69	0.0566
Scaffold-GS	161.60	0.0829
FSGS	153.91	0.0821
F^2-NeRF	152.86	0.0793
Ours	**127.67**	**0.0530**

Table 3: Quantitative results of $45°$ view rotations on the KITTI dataset using KID and FID metrics, demonstrating our method achieves relatively better quality under large rotation variations.

Models	FID↓	KID↓
Base model	98.24	0.0251
+ Illumination Model	77.24	0.0115
+ Plane Augmentation	82.59	0.0128
Ours	**71.58**	**0.0064**

Table 4: Ablation studies. Quantitative comparisons of adding Illumination Model and Plane Augmentation components on novel deviation viewpoints.

109 FPS and Scaffold-GS at 103 FPS.

Second, the viewpoints are rotated $45°$ to the left or right for evaluation and the performance as shown in Table 3. Our method achieve highest metrics under novel viewpoints. The 3D-GS shows relative better performation than GaussianPro, Scaffold-GS and other methods. This indicates that GaussianPro and Scaffold-GS are slightly overfit to the training trajectory. As presented in the Figure 7, our method outperforms high-quality views under novel viewpoints.

Ablation Study

We conduct ablation studies on KITTI dataset to analyze the effects of the proposed illumination model and plane augmentation components on the task of novel view synthesis in Table 4. The "Base Model" refers to a standard architecture that excludes contributions introduced in our proposed approach. We find that all components of our approach are essential for high-quality performance. As the visualization results shown in Figure 4, the illumination model effectively enhances the scene's appearance, leading to improved rendering quality. Additionally, plane augmentation further refines the geometry and rendering quality of planar regions.

Illumination Model. In order to understand the effectiveness of the proposed illumination model, we make a comparison on whether add this model. The results in Table 4 show that our proposed model improves multi-view consistency in the rendered results, which is demonstrated by KID and FID perceptual metrics. As the visualization results shown in Figure 4, visualization of rendered images, depth maps, and normal maps also demonstrate that our method provides more accurate geometry and a more realistic representation. Adding illumination model (second row) effectively inproves the scene's appearance, proven by the rendered images.

Plane Augmentation. We present the quantitative results in Table 4, the results of adding plane augmentation model outperforms the Base Model by a large margin in FID and KID. As the visualization depicted in Figure 4, adding plane augmentation model achieves a better visual quality on the unsupervised texture-less regions (third row), the ability of generating a realistic novel view in this case highlights the advantage of our proposed plane augmentation model. It reveals that our approach generalizes better to unobserved regions of the scene, as shown by a performance improvements compared to the Base Model.

Conclusion

This paper advances novel view synthesis with 3D Gaussian Splatting (3D-GS) by overcoming the critical challenges posed under large viewpoint deviations. We introduce a novel attention-based illumination model that overcomes the limitations of spherical harmonics by leveraging neighboring view information, thus enhancing accuracy of light field and reducing rendering artifacts. Additionally, we introduce a geometry optimization method based on planar homography transformations, which effectively enhances geometric consistency in sparse, texture-less regions. Our extensive experimental results on publicly available datasets confirm the effectiveness of our method, showing substantial improvements in view quality, especially for viewpoints that deviate significantly from those in the training data.

Acknowledgments

This work is supported by the Beijing Natural Science Foundation (No.JQ23014), Beijing Natural Science Foundation Joint Fund (No.L241056), National Natural Science Foundation of China (Nos.62271074, 62071157, 62302052, 62171321, 62162044, 32271983, 62365014, U21A20515).

References

Azinovic, D.; Martin-Brualla, R.; Goldman, D. B.; Niebner, M.; and Thies, J. 2022. Neural RGB-D Surface Reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6280–6291. New Orleans, LA, USA.

Chen, Y.; Xu, H.; Zheng, C.; Zhuang, B.; Pollefeys, M.; Geiger, A.; Cham, T.-J.; and Cai, J. 2025. MVSplat: Efficient 3d Gaussian Splatting from Sparse Multi-View Images. In *European Conference on Computer Vision*, 370–386. Springer.

Cheng, K.; Long, X.; Yang, K.; Yao, Y.; Yin, W.; Ma, Y.; Wang, W.; and Chen, X. 2024. GaussianPro: 3d Gaussian Splatting with Progressive Propagation. In *Forty-first International Conference on Machine Learning*.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-Supervised NeRF: Fewer Views and Faster Training for Free. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12872–12881. New Orleans, LA, USA.

Dey, A.; Ahmine, Y.; and Comport, A. I. 2022. Mip-NeRF RGB-D: Depth Assisted Fast Neural Radiance Fields. *Journal of WSCG*, 30(1-2): 34–43.

Du, Y.; Smith, C.; Tewari, A.; and Sitzmann, V. 2023. Learning to Render Novel Views from Wide-Baseline Stereo Pairs. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4970–4980. Vancouver, BC, Canada.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361. Providence, RI.

Guo, J.; Ma, X.; Fan, Y.; Liu, H.; and Li, Q. 2024. Semantic Gaussians: Open-Vocabulary Scene Understanding with 3D Gaussian Splatting. *arXiv:2403.15624*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH '24: Special Interest Group on Computer Graphics and Interactive Techniques Conference*, 1–11. Denver CO USA: ACM.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. GeoNeRF: Generalizing NeRF with Geometry Priors. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18344–18347. New Orleans, LA, USA.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. DNgaussian: Optimizing Sparse-View 3d Gaussian Radiance Fields with Global-local Depth Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20775–20785.

Lu, F.; Xu, Y.; Chen, G.; Li, H.; Lin, K.-Y.; and Jiang, C. 2023. Urban Radiance Field Representation with Deformable Nneural Mesh Primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 465–476.

Lu, T.; Yu, M.; Xu, L.; Xiangli, Y.; Wang, L.; Lin, D.; and Dai, B. 2024. Scaffold-GS: Structured 3d Gaussians for View-Adaptive Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20654–20664.

Neff, T.; Stadlbauer, P.; Parger, M.; Kurz, A.; Mueller, J. H.; Chaitanya, C. R. A.; Kaplanyan, A.; and Steinberger, M. 2021. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks. In *Computer Graphics Forum*, volume 40, 45–59. Wiley Online Library.

Prinzler, M.; Hilliges, O.; and Thies, J. 2023. DINER: Depth-aware Image-based Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12449–12459.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Niebner, M. 2022. Dense Depth Priors for Neural Radiance Fields from Sparse Input Views. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12882–12891. New Orleans, LA, USA.

Shi, J.; Zhi, S.; and Xu, K. 2023. PlaneRecTR: Unified Query Learning for 3D Plane Recovery from a Single View. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9343–9352. Paris, France.

Somraj, N.; and Soundararajan, R. 2023. ViP-NeRF: Visibility Prior for Sparse Input Neural Radiance Fields. In *SIGGRAPH '23: Special Interest Group on Computer Graphics and Interactive Techniques Conference*, 1–11. Los Angeles CA USA: ACM.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S.; Krivokon, M.; Gao, A.; Joshi, A.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2443–2451. Seattle, WA, USA.

Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying Mmd Gans. *stat*, 1050: 12.

Tseng, H.-Y.; Li, Q.; Kim, C.; Alsisan, S.; Huang, J.-B.; and Kopf, J. 2023. Consistent View Synthesis with Pose-Guided Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16773–16783. Vancouver, BC, Canada.

Wang, G.; Chen, Z.; Loy, C. C.; and Liu, Z. 2023a. SparseNeRF: Distilling Depth Ranking for Few-Shot Novel View Synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9031–9042. Paris, France.

Wang, P.; Liu, Y.; Chen, Z.; Liu, L.; Liu, Z.; Komura, T.; Theobalt, C.; and Wang, W. 2023b. F^2 -NeRF: Fast Neural Radiance Field Training with Free Camera Trajectories. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4150–4159. Vancouver, BC, Canada.

Wewer, C.; Raj, K.; Ilg, E.; Schiele, B.; and Lenssen, J. E. 2025. Latentsplat: Autoencoding Variational Gaussians for Fast Generalizable 3d Reconstruction. In *European Conference on Computer Vision*, 456–473. Springer.

Xiong, H.; Muttukuru, S.; Upadhyay, R.; Chari, P.; and Kadambi, A. 2023. SparseGS: Real-time 360° Sparse View Synthesis Using Gaussian Splatting. *arXiv:2312.00206*.

Zhu, Z.; Fan, Z.; Jiang, Y.; and Wang, Z. 2025. FSGS: Real-time Few-Shot View Synthesis Using Gaussian Splatting. In *European Conference on Computer Vision*, 145–163. Springer.