2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

LSADS-Gaussian: Gaussian Splatting for Large-Scale Autonomous Driving Scene Reconstruction

Ping Wang[1,*], Ben Li[1,*], Bo Qian[2], Chuan Jin[1], Can Tian[3], and Yusheng Ji[4]

***Abstract*—The rapid advancement of 3D scene understanding techniques presents a significant opportunity for enhancing autonomous driving simulation systems. As these systems are increasingly required to operate in complex, large-scale, and unbounded real-world environments, efficient and high-fidelity 3D reconstruction of common outdoor scenes has become a critical prerequisite for realistic and extensible autonomous driving simulation. 3D Gaussian Splatting has achieved state-of-the-art performance in novel view synthesis, coupled with real-time rendering efficiency. However, large-scale reconstruction for autonomous driving scenarios faces several challenges as scenes grow in complexity: (1) limited views with insufficient pose diversity, (2) inadequate representation of geometric structural details, and (3) complex lighting conditions involving saturation and shadow variations. To cope with these challenges, we propose LSADS-Gaussian, a novel model for large-scale autonomous driving scene reconstruction. The model consists of a Multimodal Gaussian Network (MGN) module composed of two Gaussian sub-networks, designed to perform Gaussian aggregation and optimization from multi-sensor data, a Geometric Representation Guidance (GRG) module refines and enhances geometric consistency, and a Lighting Enhancement (LE) module introduces learnable illumination coefficients to maintain illumination consistency. Extensive experiments show that LSADS-Gaussian outperforms the state-of-the-art methods.**

I. Introduction

Autonomous driving has made unprecedented progress in recent years [1], particularly in perception [2, 3], HD map construction [4, 5], and planning [6]. With the rise of end-to-end systems that directly generate driving control signals from raw sensor inputs, traditional open-loop evaluation protocols are no longer sufficient. A promising alternative lies in real-world closed-loop evaluation, where the key challenge is achieving efficient and high-fidelity scene reconstruction. Currently, there are two main 3D reconstruction frameworks, Neural Radiance Fields (NeRF) [7] and 3D Gaussian Splatting (3DGS) [8]. NeRF has made significant advances through detailed modeling techniques and architectural innovations. However, it remains limited by its implicit geometry representation and high computational demands. In contrast, 3DGS employs an explicit geometric representation in the form of Gaussian point clouds, which encode scene information through attributes such as position, color, rotation, scale, and opacity. These advantages effectively address the major limitations inherent in current 3D scene reconstruction methods.

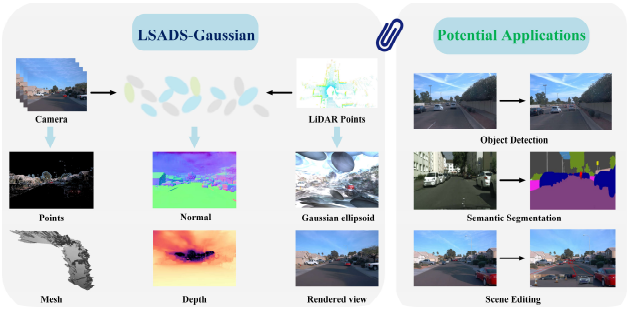

Fig. 1. **LSADS-Gaussian.** An efficient and extensible framework for large-scale autonomous driving scene reconstruction with 3D Gaussian splitting representation. The framework is applicable to potential work and can be effectively extended to downstream tasks, including object detection, semantic segmentation, and scene editing.

Extensive efforts have been devoted to the application domains of 3D scene reconstruction [9-11]. However, large-scale outdoor scenes captured in autonomous driving sensor datasets remain challenges for 3DGS. Such scenes are typically acquired from sparse viewpoints with limited view diversity, as cameras are mounted on forward-moving vehicles. Moreover, autonomous driving scenes tend to be large-scale and unbounded, with much of their content located at infinity from the camera. Optimizing solely with image reconstruction loss often leads to local optima. Furthermore, outdoor images often display variations in exposure and brightness, influenced by complex lighting and weather conditions. Cameras may adopt different exposure times at different frames, resulting in overall brightness fluctuations in the images. Consequently, how to improve geometric reconstruction accuracy is still a challenge.

To address these issues, we propose LSADS-Gaussian, a novel framework for reconstructing large-scale outdoor scenes in autonomous driving applications. Unlike previous 3DGS-based methods [12, 13], our method achieves accurate 3D scene reconstruction from sparse image views and LiDAR scans. Specifically, the model employs a MGN to fuse LiDAR and image features to refine and aggregate Gaussian point clouds. And then, a GRG module to better capture the geometry of large-scale outdoor scenes by regularizing Gaussian positions with directed-depth estimation under depth and normal constraints. Finally, the LE module is used during rendering to ensure illumination consistency across different viewpoints and environmental conditions.

[1] Ben Li, Ping Wang, and Chuan Jin are with the Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 200092, China. Ping Wang is also with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (*Corresponding author: Ben Li)* (e-mail: liben0210@outlook.com).
[2] Bo Qian is with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8657, Japan.
[3] Can Tian is with the Geely Automotive Research Institute, Ningbo 315300, China.
[4] Yusheng Ji is with the Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo 101-8430, Japan.
* Equal Contribution.

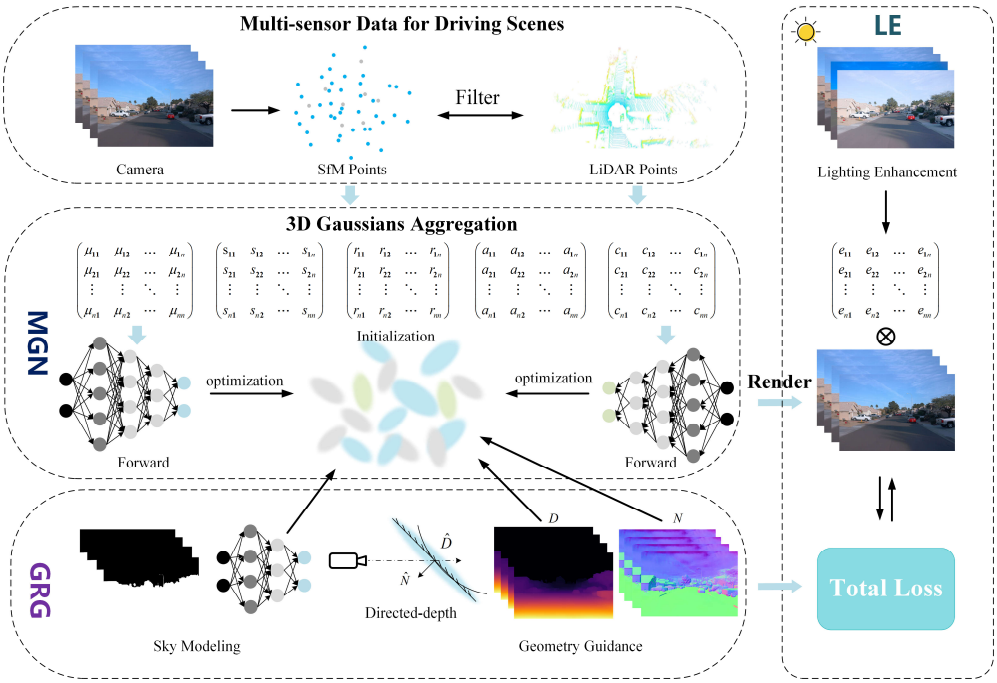

Fig. 2. **Overview.** LSADS-Gaussian mainly comprises three sub-modules: a MGN module for fusing camera and LiDAR data to aggregate 3D Gaussians, a GRG module for refining geometry with directed-depth estimation and sky masking, and a LE module for stabilizing illumination during rendering.

The main contributions of this work are summarized as follows:

- To overcome the sparsity of viewpoints with restricted pose diversity, a MGN module is proposed to effectively integrate LiDAR and camera images features to refine the 3D scene geometry.
- A GRG module is designed to infer the directed-depth of Gaussians under depth and normal constraints to optimize geometric details of large-scale outdoor scenes.
- A LE module introduces the illumination variations coefficients to effectively enhance illumination consistency.
- Integrating the above modules, we propose a novel LSADS-Gaussian for large-scale autonomous driving scene reconstruction. Extensive experiments show that LSADS-Gaussian model outperforms the state-of-the-art models.

II. Related Work

A. Novel view synthesis

Novel view synthesis is a critical task in 3D scene reconstruction. Numerous approaches have been proposed to represent 3D geometry and render novel views [14, 15]. Efficient representations including meshes, voxels, and multi-plane images can produce high-fidelity renderings given sufficient supervision. Recently, NeRF-based methods [16, 17] show that implicit radiance fields are capable of learning detailed scene representations and generating high-quality novel views. However, these methods frequently suffer from slow convergence and limited model capacity. Following this, 3DGS [8], characterized by explicit representation and differentiable point-based splatting, supports real-time novel view rendering and has attracted widespread attention in the field.

B. Bounded Scenes Reconstruction

Bounded scene reconstruction has been widely studied in computer vision and graphics. With clearly defined spatial boundaries and relatively constrained environments, these scenes (such as indoor environments, tabletop objects, or small-scale outdoor settings) provide favorable conditions for 3D reconstruction. Numerous methods have been proposed, leveraging representations such as meshes, voxels, and multi-plane images. Specifically, for object-centric 360° multi-view datasets, Mip-NeRF 360 [9] can effectively learn continuous scenes, integrating image and point cloud data to optimize scene geometric accuracy. For indoor environments, DN-Splatter [18] employs globally optimized 3D Gaussian positions and RGB-D-based rendering loss to enhance geometric accuracy. Unfortunately, these methods perform poorly in large-scale outdoor scenes for autonomous driving.

C. Unbounded Scenes Reconstruction

Unlike indoor or object-centric settings with dense view coverage, urban outdoor scenes are often collected from sparse and forward-facing viewpoints, resulting in low view diversity and incomplete scene coverage. To overcome these limitations in large urban environments, NeRF-based methods [19-21] have undergone extensive refinements. Block-NeRF [16] and Mega-NeRF [22] employ a block-based combination strategy to model large-scale scenes. Afterwards, MARS [23] designs an instance-aware simulation framework that adopts distinct networks for background and vehicle modeling, while additionally modeling the sky separately. Nevertheless, they fail to fully exploit multi-sensor geometric priors and geometric details at scene boundaries to enhance scene representation.

Recently, 3DGS-based methods [24-26] have achieved remarkable success. DrivingGaussian [26] uses a LiDAR prior for Gaussian Splatting to reconstruct scenes with greater details and maintain geometric consistency. Unfortunately, they lack the geometric structure of scenes beyond the LiDAR range, leading to blurry and missing details in distant objects. Meanwhile, DN-Splatter [18] and PGSR [12] consider depth and normal priors to constrain the geometric consistency of the entire 3D scene. However, they ignore the offset depth along the normal direction of Gaussian ellipsoids in 3D scene, leading to Gaussian aliasing. To overcome the impact of complex illumination in outdoor scenes, current methods [27, 28] consider angular illumination effects on objects but overlook inter-frame illumination consistency.

To address these limitations, our LSADS-Gaussian framework optimizes the integration of camera images and LiDAR point clouds to aggregate 3D Gaussians for accurate geometry learning. A geometric guidance strategy refines Gaussian positions through directed-depth updates under depth and normal priors, with sky masking to handle infinitely distant regions. Additionally, a Lighting Enhancement module stabilizes illumination during rendering and effectively enhances illumination consistency. The overall framework is illustrated in Fig. 2.

III. Preliminaries

A. Definition and Notation

Definition 1: Gaussian network $\mathcal{G}$. We define a multimodal Gaussian feedforward network as 3D scene geometric structure $\mathcal{G} = \{\boldsymbol{G} \in \mathbb{R}^n\}$. A 3D Gaussian ellipsoid is defined as $\boldsymbol{G} = \{\boldsymbol{\mu} \in \mathbb{R}^3, \boldsymbol{s} \in \mathbb{R}^3, \boldsymbol{q} \in \mathbb{R}^4, a \in \mathbb{R}^1, \boldsymbol{c} \in \mathbb{R}^3\}$, where each notation represents position, scale, rotation, opacity scalar, and RGB color vectors, respectively. Color attributes are modeled using spherical harmonics (SH). The 3D Gaussian ellipsoid is mathematically described as:

$$\boldsymbol{G}(\boldsymbol{x}) = \frac{1}{(2\pi)^{3/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (1)$$

where $\boldsymbol{\mu}$ denotes the Gaussian position center, and $\boldsymbol{\Sigma}$ denotes the covariance matrix.

$$\boldsymbol{\Sigma} = \boldsymbol{RSS}^T\boldsymbol{R}^T, \quad (2)$$

where $\boldsymbol{S}$ denotes a diagonal scaling matrix, $\boldsymbol{R}$ denotes a rotation matrix, parameterized as a scaling vector $\boldsymbol{s}$ and a quaternion $\boldsymbol{r}$, respectively.

For image generation from a specific viewpoint, 3D Gaussian ellipsoids $\boldsymbol{G}$ are projected onto the 2D image plane as ellipses $\boldsymbol{G^{2D}}$. To approximate the projection process, given a viewing transformation $\boldsymbol{W}$, the covariance matrix $\boldsymbol{\Sigma}'$ in camera coordinate is given as follows:

$$\boldsymbol{\Sigma}' = \boldsymbol{JW\Sigma W}^T\boldsymbol{J}^T, \quad (3)$$

where $\boldsymbol{J}$ is the Jacobian of the affine approximation. For each pixel, a sequence of Gaussians $\boldsymbol{G}^s$ are depth-sorted in ascending order, and the color C is rendered via alpha blending:

$$C = \sum_{i\in N} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \quad (4)$$

where α_i and c_i denote the blending coefficient and color of the *i-th* Gaussian $\boldsymbol{G_i}$, respectively, obtained from the learned opacity a_i and SH coefficients of the corresponding Gaussian.

Definition 2: Geometric representation guidance strategy. To more accurately capture the structural details of the scene, we employ depth and normal priors to jointly design a directed-depth estimation method that constrains the geometric consistency of the scene.

Definition 3: Lighting enhancement matrix $\boldsymbol{E^{light}}$. The illumination variations coefficient e_i, which enhance illumination consistency in the driving scenes, are denoted as $\boldsymbol{E^{light}} = \{e_1, e_2, \dots, e_t\}$.

B. Problem Modeling

The large-scale autonomous driving scene reconstruction problem can be regarded as learning a function f based on Gaussian network $\mathcal{G}$, camera images I, LiDAR points L, the lighting enhancement matrix $\boldsymbol{E^{light}}$, depth and normal priors. The function is employed to get Gaussians in the scene, defined as:

$$\{\boldsymbol{G} \in \mathbb{R}^n\} = f(\mathcal{G}, I, L|\boldsymbol{E^{light}}, \boldsymbol{D}, \boldsymbol{N}), \quad (5)$$

where $\boldsymbol{D}$, $\boldsymbol{N}$ denote depth and normal priors, respectively. $\boldsymbol{G}$ denotes the set of Gaussians used for scene representation.

IV. Methodology

A. Framework

As shown in Fig. 2, the LSADS-Gaussian mainly consists of three sub-modules. The Multimodal Gaussian Network (MGN) module composed of two Gaussian sub-networks, designed to perform Gaussian initialization, aggregation and optimization from multi-sensor data. The Geometric Representation Guidance (GRG) module employs a sky model to deal with appearances at infinite distance, and refines Gaussian positions through directed-depth updates under depth and normal priors to maintain geometric consistency of the scene. During rendering, the Lighting Enhancement (LE) module introduces illumination coefficients to ensure illumination consistency. To this end, we formulate a total loss function that iteratively optimizes the attributes of the entire Gaussian geometric scene. Through end-to-end training, all Gaussian parameters are jointly refined, enabling geometrically consistent and illumination-aware scene reconstruction.

B. Multimodal Gaussian Network Modeling

3DGS [8] demonstrates that 3D Gaussians can be effectively trained using Structure-from-Motion (SfM) [29] for point initialization. Nevertheless, it struggles to provide reliable initialization in large-scale autonomous driving scenes characterized by under-observed regions and complex unbounded environments. To address this limitation, we initialize and optimize 3D Gaussians using aggregated SfM and LiDAR point clouds captured by multi-sensor vehicles. Specifically, we introduce a filter that aggregates and samples SfM and LiDAR point clouds, with some drifting points from initialization being filtered out, achieving high-quality Gaussian ellipsoids.

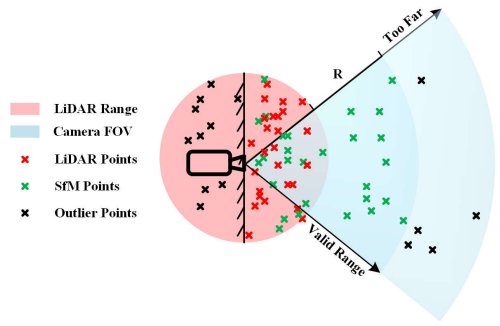

Fig. 3. The filtering strategy for SfM and LiDAR point clouds.

The filtering strategy is illustrated in Fig. 3. To eliminate drifting outliers and ensure consistent spatial coverage, we first map SfM and LiDAR point clouds into the world coordinate system, and then filter out SfM points located behind the camera view. Within the LiDAR range, SfM points are integrated to complement regions that are insufficiently captured by LiDAR. Furthermore, while LiDAR provides reliable measurements within its coverage, SfM points that lie outside the LiDAR field of view but remain within its maximum radial distance are also retained. To enrich this range structural details, we randomly sample 10k points to supplement the fused point cloud in large-scale environments and enhance its completeness and robustness.

Through the filtering process, we obtain a series of dense and reliable point clouds, which are used to optimize the 3D Gaussian ellipsoids positions $\boldsymbol{p}$. Subsequently, A small opacity a_i is used to derive the blending coefficient α_i by Equation (6). To update the covariance matrix $\boldsymbol{\Sigma_i}$, we introduce quaternion $\boldsymbol{q} = q_r + q_i \cdot \boldsymbol{i} + q_j \cdot \boldsymbol{j} + q_k \cdot \boldsymbol{k}$ to compute the rotation matrix $\boldsymbol{R}$, defined as:

$$\alpha_i = a_i \cdot \exp\left(-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i^{-1}} (\boldsymbol{p}-\boldsymbol{\mu_i})\right), \tag{6}$$

$$\boldsymbol{R} = \begin{pmatrix} 1-2({q_j}^2+{q_k}^2) & 2(q_i q_j - q_r q_k) & 2(q_i q_k + q_r q_j) \\ 2(q_i q_j + q_r q_k) & 1-2({q_i}^2+{q_k}^2) & 2(q_j q_k - q_r q_i) \\ 2(q_i q_k - q_r q_j) & 2(q_j q_k + q_r q_i) & 1-2({q_i}^2+{q_j}^2) \end{pmatrix}. \tag{7}$$

In addition, during the projection and rendering of Gaussian ellipsoids, the color attributes $c_i = \sum_{l=0}^{L}\sum_{m=-l}^{l} \beta_{lm} Y_l^m(\theta,\varphi)$ are modeled using spherical harmonics (SH). In this way, the color of each Gaussian ellipsoid is accurately represented under different camera viewpoints. In summary, our feed-forward network can be defined as:

$$\begin{gathered} \boldsymbol{\mu}, \boldsymbol{\Sigma}, a, \boldsymbol{c} = (x,y,z), (\boldsymbol{R}, \boldsymbol{S}), a, (f_0, f_{rest}) \\ \boldsymbol{\mu}, \boldsymbol{\Sigma} \xrightarrow{\boldsymbol{projection}} \boldsymbol{\mu^{2D}}, \boldsymbol{\Sigma^{2D}} \\ C = \sum_{i\in N} c_i \alpha_i \boldsymbol{G}(\boldsymbol{\mu^{2D}}, \boldsymbol{\Sigma^{2D}}) \prod_{j=1}^{i-1}\left(1-\alpha_j \boldsymbol{G}(\boldsymbol{\mu^{2D}}, \boldsymbol{\Sigma^{2D}})\right), \\ grad(C) \xrightarrow{\boldsymbol{gradient}} grad(\boldsymbol{\mu}, \boldsymbol{\Sigma}, a, \boldsymbol{c}) \end{gathered} \tag{8}$$

where $Y_l^m(\theta,\varphi)$ denotes SH basis functions, θ and φ denote the polar angle and azimuth angle of the direction from the Gaussian center to the camera. β_{lm} denotes the learnable SH coefficients. l denotes the degree and m denotes the order of the associated Legendre functions.

To this end, two identical forward networks are employed for Gaussian aggregation to maintain the consistency of scene Gaussian attributes, with the parameters are jointly optimized through a differentiable rasterization-based rendering pipeline. Furthermore, the Gaussian geometry scene is optimized using the Adaptive Density Control (ADC) algorithm [8], which continuously culls, duplicates, and splits Gaussians based on their position centers, scales, rotations, opacity scalars, colors, and scene-space sizes, respectively.

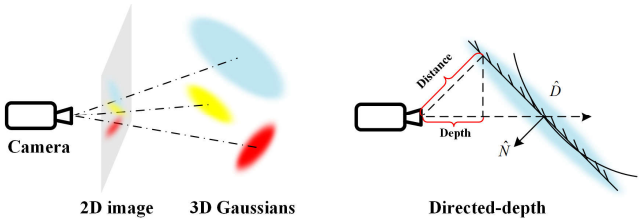

Fig. 4. Directed-depth estimation.

C. Geometric Representation Guidance

Due to the challenges of modeling fine-grained geometric attributes in large-scale autonomous driving scenes using 3D Gaussians, it is essential to impose geometric consistency constraints on 3DGS and employ appropriate geometric representation guidance strategies to more accurately capture the scene geometry. Therefore, as illustrated in Fig. 4, we introduce depth and normal priors, and propose a directed-depth estimation method to more effectively refine geometric consistency, thereby avoiding Gaussian aliasing.

During optimization, specifically, we compress the Gaussian ellipsoid along its minimum scale direction, thereby obtaining a geometric shape that best approximates the original scene. This edge-aware geometric approximation enables the model to capture fine-grained structural details while maintaining geometric consistency across large-scale driving scenes. We directly determine the normal direction based on the rotation $\boldsymbol{R}$ and scaling $\boldsymbol{S}$ matrices, where the normal $\widehat{\boldsymbol{n}}$ is aligned with the minimum scale factor of each Gaussian and minimizes this factor, defined as:

$$\widehat{\boldsymbol{n}}_{\boldsymbol{i}} = \boldsymbol{R_i} \cdot OneHot(argmin\, \boldsymbol{S_i}(s_1,s_2,s_3)), \tag{9}$$

$$\mathcal{L}_s = \sum_{i=1}^{n} argmin\, \boldsymbol{S_i}(s_1,s_2,s_3), \tag{10}$$

where $OneHot(\cdot) \in \mathbb{R}^3$ returns a unit vector with zeros in all positions except at the index corresponding to the minimum scaling factor in $\boldsymbol{S_i}(s_1,s_2,s_3)$.

To ensure consistent orientation, when the dot product between the camera viewing direction and the Gaussian normal is negative, the normal is flipped to enforce a positive dot product. The normals are subsequently transformed into camera space using the current camera transform and alpha-composited via the rendering equation, resulting in a single per-pixel normal estimate $\widehat{\boldsymbol{N}}$.

$$\widehat{\boldsymbol{N}} = \sum_{i\in n} \widehat{\boldsymbol{n}}_{\boldsymbol{i}} \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j). \tag{11}$$

The distance between the Gaussian and the camera center is expressed as: $d_i = \frac{\boldsymbol{n_i}\cdot(\boldsymbol{p_c}-\boldsymbol{\mu_i})}{\|\boldsymbol{n_i}\|}$, where $\boldsymbol{p_c}$ is the camera center. Per-pixel distance map $\boldsymbol{D'}$ is obtained via a discrete volume rendering approximation, in the same manner as color values:

$$\boldsymbol{D'} = \sum_{i\in n} d_i\, \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j). \tag{12}$$

Referencing Fig. 4, after obtaining the distance and normal of

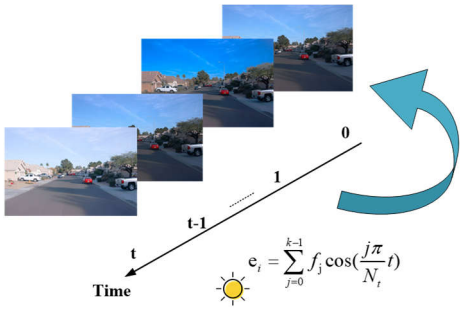

Fig. 5. Lighting enhancement during rendering for scene reconstruction.

the Gaussian through rendering, the corresponding directed-depth map can be determined $\widehat{\boldsymbol{D}} = \boldsymbol{D}'cos(\psi)$ based on the angle ψ between the camera ray direction and the Gaussian normal $\widehat{\boldsymbol{N}}$.

Therefore, our geometric representation guidance strategy provides two main advantages. It directly derives normals and depths from Gaussian geometry, allowing covariance matrices to be updated during backpropagation to automatically refine these attributes without introducing extra parameters. Moreover, unlike α-blending methods that assign Gaussian centers as depths and lead to inconsistent maps, our directed-depth maps are computed from Gaussian normals and viewing rays, ensuring guided Gaussian geometry is fully consistent with the actual geometric surface.

To alleviate scene boundary artifacts, we employ a gradient-based depth loss derived from the current camera image for adaptive depth regularization. The depth loss is reduced in regions with large image gradients (i.e., edge regions), ensuring that the regularization is primarily concentrated on smoother and textureless areas, which are typically more challenging for photometric regularization alone. The gradient-aware depth loss that adaptively balances regularization across image regions according to scene geometry and texture, defined as:

$$\mathcal{L}_D = \exp(\partial I)\frac{1}{n}\sum \log\left(1 + \left\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\right\|_1\right), \quad (13)$$

where ∂I denotes the gradient of the current camera image.

Due to the noise in rendered depth maps, particularly in complex large-scale autonomous driving scenes, previous methods [8, 26] often result in Gaussian aliasing. We instead first apply a L1 loss for regularization, and further impose a prior on the total variation of inferred normals to enhance spatial smoothness across neighboring pixels, defined as:

$$\mathcal{L}_N = \frac{1}{n}\sum\left\|\widehat{\boldsymbol{N}} - \boldsymbol{N}\right\|_1 + \sum_{i,j}\left(\left|\partial_i \widehat{\boldsymbol{N}}_{\boldsymbol{i,j}}\right| + \left|\partial_j \widehat{\boldsymbol{N}}_{\boldsymbol{i,j}}\right|\right), \quad (14)$$

where $\partial_i \widehat{\boldsymbol{N}}_{\boldsymbol{i,j}}$ and $\partial_j \widehat{\boldsymbol{N}}_{\boldsymbol{i,j}}$ denote the gradients of the inferred normal along the i-axis and j-axis directions, respectively.

To address infinitely distant regions, we utilize sky mask and build a separate sky model, ensuring that the sky is modeled independently from the scene geometry. After optimizing the Gaussian features of the scene geometry, we combine them with the entire scene to obtain the complete Gaussian representation of the scene geometry. To better constrain the sky, we introduce a BCE-based semantic regularization to alleviate this issue, defined as:

$$\mathcal{L}_M = -\frac{1}{n}\sum_{i=1}^{n} M_i log(P_i), \quad (15)$$

where M_i denotes the sky mask region. Combining the above strategies, we achieve complete Gaussian guidance of the scene.

D. Lighting Enhancement

For large-scale outdoor environment, external illumination changes cause inconsistent exposure conditions among camera images, leading to global lighting shifts. As the original 3DGS ignores illumination and shadow variations, it leads to inconsistent appearance across views and temporal frames. To model the global lighting variations at different times, we introduce the illumination variations coefficients e_i for each image, with a set of Fourier transform coefficients $\boldsymbol{f} \in \mathbb{R}^k$, where k denotes the number of Fourier coefficients. Given a timestep t, the coefficient e_i is computed using the Inverse Discrete Fourier Transform (IDFT):

$$e_i = \sum_{j=0}^{k-1} \boldsymbol{f_j} \cos\left(\frac{j\pi}{N_t} t\right). \quad (16)$$

Thus, light-enhanced image $\boldsymbol{I_i^e}$ are generated by incorporating the illumination variations coefficients e_i into the rendering process, formulated as: $\boldsymbol{I_i^e} = e_i \cdot \boldsymbol{I_i^r}$, where $\boldsymbol{I_i^r}$ denotes the rendered image. We define the following image loss:

$$\mathcal{L}_c = (1-\lambda)L_1(\boldsymbol{I_i^e} - \boldsymbol{I_i}) + \lambda L_{SSIM}(\boldsymbol{I_i^r} - \boldsymbol{I_i}), \quad (17)$$

where $\boldsymbol{I_i}$ denotes the ground truth image. The L_1 loss ensures that the light-enhanced image is consistent with the ground truth image. The L_{SSIM} loss enforces the rendered image to maintain structural similarity with the ground-truth image.

In addition, to enhance illumination consistency across adjacent frames, we introduce an illumination consistency loss to constrain the lighting variations between neighboring frames, defined as:

$$\mathcal{L}_{nc} = \frac{1}{W}\sum_{\boldsymbol{p_r} \in W}\left(1 - \frac{cov(\boldsymbol{I_r}(\boldsymbol{p_r}), \boldsymbol{I_n}(\boldsymbol{H_{rn}}\boldsymbol{p_r}))}{\sqrt{var(\boldsymbol{I_r}(\boldsymbol{p_r}))\cdot var(\boldsymbol{I_n}(\boldsymbol{H_{rn}}\boldsymbol{p_r}))}}\right), \quad (18)$$

where W denotes the set of all pixels in the image. $\boldsymbol{p_r}$ denotes pixel block of image $\boldsymbol{I_r}$. $\boldsymbol{H_{rn}}$ denotes the Homography matrix that maps $\boldsymbol{p_r}$ to the neighboring frame block $\boldsymbol{p_n}$ of image $\boldsymbol{I_n}$. $cov(\cdot,\cdot)$ calculates the covariance between two vectors, and $var(\cdot)$ calculates the variance of a vector.

In summary, our final training total loss $\mathcal{L}_{total}$ consists of the minimum scale factor loss $\mathcal{L}_s$, the gradient-aware depth loss $\mathcal{L}_D$, the normal regularization loss $\mathcal{L}_N$, the sky BCE-loss $\mathcal{L}_M$, the image reconstruction loss $\mathcal{L}_c$, the illumination consistency loss $\mathcal{L}_{nc}$:

$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_D + \mathcal{L}_N + \mathcal{L}_M + \mathcal{L}_c + \mathcal{L}_{nc}. \quad (19)$$

V. Experiments

In this section, we conduct extensive evaluations of LSADS-Gaussian across large-scale autonomous driving scenes, providing comprehensive qualitative and quantitative comparisons with existing approaches. To enable a fair comparison with baseline methods designed for small-scale indoor and object-level scenarios, we remove the LiDAR component from our framework and conduct extensive experiments under equivalent conditions.

TABLE I. QUANTITATIVE COMPARISON WITH BASELINES ON THE WAYMO DATASET AT 7K AND 30K ITERATIONS.

Methods	Input	Waymo-7K			Waymo-30K			Average		
		PSNR ↑	*SSIM* ↑	*LPIPS* ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓
Instant-NGP [10]	Images	25.88	0.682	0.375	29.78	0.769	0.264	27.83	0.726	0.320
Mip-NeRF [31]	Images	27.63	0.702	0.357	32.08	0.802	0.221	29.86	0.752	0.289
Mip-NeRF360 [9]	Images	27.61	0.698	0.355	32.61	0.818	0.235	30.11	0.758	0.295
Urban-NeRF [32]	Images + LiDAR	28.15	0.709	0.332	34.75	0.857	0.203	31.45	0.783	0.268
S-NeRF [33]	Images + LiDAR	28.67	0.728	0.317	35.43	0.878	0.198	32.05	0.803	0.258
EmerNeRF [34]	Images + LiDAR	29.75	0.760	0.301	36.27	0.925	0.183	33.01	0.843	0.242
3DGS [8]	Images	30.38	0.787	0.258	35.94	0.918	0.191	33.16	0.853	0.225
DeSiRe-GS [25]	Images+ LiDAR	31.32	0.813	0.224	36.68	0.947	0.182	34.00	0.880	0.203
DrivingGaussian [26]	Images+ LiDAR	31.01	0.793	0.231	36.13	0.922	0.190	33.57	0.858	0.211
LSADS-Gaussian	Images+ LiDAR	**32.41**	**0.821**	**0.219**	**37.96**	**0.968**	**0.169**	**35.19**	**0.895**	**0.194**

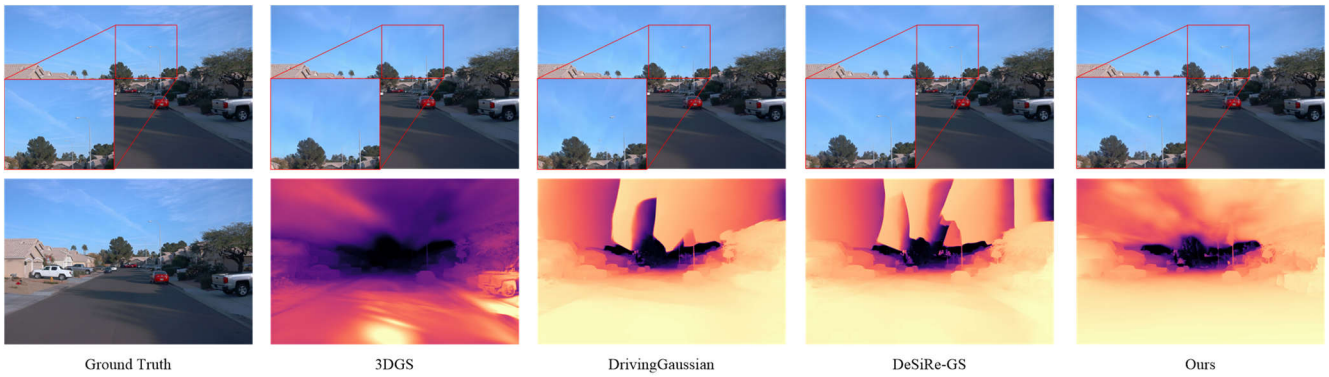

Fig. 6. Qualitative comparison on Waymo dataset. The first row shows rendered images and the second row shows depth maps.

A. Datasets

To comprehensively evaluate LSADS-Gaussian, we consider large-scale outdoor, indoor, and object-level scenes. For large-scale autonomous driving scenes, we select driving scenes from the Waymo Open Dataset [30], each with about 200 frames captured at 10 Hz and exhibiting significant ego-vehicle motion. We sample every 4th frame for testing and use the remaining frames for training. For small-scale indoor and object-level scenes, we employ the TnT and Mip-NeRF 360 datasets, and follow the experimental setup of PGSR [12] for reconstruction quality evaluation.

B. Implementation Details

Our model is primarily based on the 3DGS framework [8], with fine-tuned optimization parameters to fit the large-scale autonomous driving scenes. The training strategy and hyperparameter settings largely follow 3DGS. We train all scenes for 7k and 30k iterations. The learning rate for the illumination coefficient is set to 10^{-3}. The experimental server was equipped with an Intel i9-14900K CPU, NVIDIA RTX 4080 GPU. We initialize with a total of 1 million points, consisting of approximately 60% from the LiDAR point cloud, 30% SfM points generated by COLMAP, and the rest are randomly sampled. We adopt three widely used image quality metrics to evaluate image reconstruction: PSNR, SSIM, and LPIPS.

C. Baselines and Results Analysis

1) Performance Analysis

To evaluate performance, we compare LSADS-Gaussian with recent state-of-the-art methods, covering NeRF-based methods [9, 10, 31-34] and 3DGS-based methods [8, 25, 26]. As shown in Table I, LSADS-Gaussian far outperforms InstantNGP [10], which leverages multi-resolution grids and hash-based NeRF for accelerated rendering. Moreover, despite Mip-NeRF [31] and Mip-NeRF360 [9] being designed for multi-view unbounded outdoor scenes, their performance is limited by viewpoint constraints, and our method surpasses them across all evaluation metrics. To enrich the extraction of scene geometry, Urban-NeRF [32] introduces LiDAR into NeRF for large-scale autonomous driving scene reconstruction, but it merely uses LiDAR as depth supervision. In contrast, our method leverages LiDAR as a precise geometric prior and integrates it into Gaussian models, enabling the effective aggregation of 3D Gaussians for learning precise scene geometry. S-NeRF [33] is highly dependent on the effectiveness of scene decomposition, which limits its robustness in complex large-scale scenes. For EmerNeRF [34], it integrates multi-frame sequential data to optimize complex driving scenes and achieves significant improvements. However, insufficient constraints fail to ensure geometric consistency, resulting in noticeable blurring.

For 3DGS-based methods, 3DGS [8] represents the scene with a set of anisotropic Gaussians. Unfortunately, the absence of dense geometric details and effective geometric guidance leads to degraded performance. In large-scale sparse-data autonomous driving scenes, although DeSiRe-GS [25] and DrivingGaussian [26] introduce geometric regularizations, they overlook the complex illumination effects in outdoor environments and reliable geometric guidance strategy. Conversely, our method not only captures rich scene geometric features but also employs a geometric representation guidance module to maintain scene geometric consistency. Furthermore, we employ lighting enhancement during rendering to ensure illumination consistency.

TABLE II. QUANTITATIVE COMPARISON WITH DIFFERENT BASELINES ON MIP-NERF 360 DATASET.

Methods	Indoor scenes			Outdoor scenes			Average		
	PSNR ↑	*SSIM* ↑	*LPIPS* ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓
NeRF [7]	26.84	0.790	0.370	21.46	0.458	0.515	24.15	0.624	0.443
Instant-NGP [10]	29.15	0.880	0.216	22.90	0.566	0.371	26.03	0.723	0.294
Mip-NeRF360 [9]	31.72	0.917	0.180	24.47	0.691	0.283	28.10	0.804	0.232
NeuS [35]	25.10	0.789	0.319	21.93	0.629	0.600	23.52	0.709	0.460
3DGS [8]	**30.99**	0.926	0.199	24.24	0.705	0.283	27.62	0.816	0.241
2DGS [36]	30.39	0.923	0.183	24.33	0.709	0.284	27.36	0.816	0.234
PGSR [12]	30.41	**0.930**	**0.161**	24.45	0.730	0.224	27.43	0.830	0.193
LSADS-Gaussian	30.66	0.925	0.167	**25.08**	**0.747**	**0.211**	**27.87**	**0.836**	**0.189**

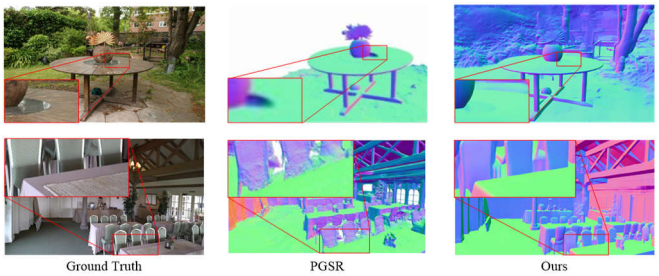

Fig. 7. Qualitative comparison of rendered normal maps with PGSR.

TABLE III. QUANTITATIVE RESULTS COMPARING RECONSTRUCTION PERFORMANCE WITH DIFFERENT COMPONENTS.

Components					Waymo-30K		
Images	*LiDAR*	*Random*	*GRG*	*LE*	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓
		✓	✓	✓	35.23	0.869	0.201
✓			✓	✓	36.12	0.911	0.196
	✓		✓	✓	36.33	0.917	0.189
✓	✓		✓	✓	37.35	0.945	0.173
✓	✓	✓		✓	36.61	0.923	0.185
✓	✓	✓	✓		37.15	0.948	0.181
✓	✓	✓	✓	✓	**37.96**	**0.968**	**0.169**

2) Reconstruction Results

To assess reconstruction quality, we conduct a detailed qualitative comparison between our rendered RGB images and depth maps from camera viewpoints and those rendered by state-of-the-art methods in Fig. 6. Specifically, 3DGS fails to accurately estimate depth details of near-field objects. While DeSiRe-GS and DrivingGaussian can recover the geometry of near-field ground objects, they miss important details of far-field objects located near the sky. By contrast, our rendered images and depth maps capture finer geometric details with smooth and consistent appearance, benefiting from aggregated Gaussians and geometric guidance strategy.

Moreover, to further assess the generalization capability of our model in reconstructing indoor and multi-view scenes, we reconstruct the scenes using only camera images and conduct quantitative comparisons on the TnT and MipNerf 360 datasets, evaluating image reconstruction and novel view synthesis quality against NeRF-based and GS-based baselines. In Table II, our model attains reconstruction quality comparable to current state-of-the-art methods on indoor scenes and achieves excellent performance on outdoor scenes, indicating strong generalization to novel viewpoints.

We further conduct a qualitative comparison of the rendered normal maps against PGSR to evaluate the smoothness and consistency of the geometric structures. As shown in Fig. 7, when reflective surfaces are present, PGSR often produces geometric holes, whereas our model can estimate smooth and complete geometric surfaces. Moreover, for complex shadowed planar regions, our method can clearly infer sharp and consistent geometric boundaries, resulting in smoother and more precise structures.

D. Ablation Studies

In this section, the following six composite models with different components are implemented to evaluate their effectiveness. We test each model and compare the quantitative results as shown in Table III.

Specifically, to analyze the impact of point cloud quantity, we conduct experimental comparisons using three different point cloud initialization and aggregation strategies. The results show that randomly generated points result in the poorest performance because of the absence of geometric priors. Similarly, initialization with SfM points inferred from images cannot sufficiently reconstruct precise scene geometry, due to their sparsity and the presence of significant structural errors. For the model initialized with LiDAR prior, it preserves relatively accurate structural priors and thus outperforms SfM initialization. However, it remains limited by the lack of geometric features in far-field regions. In contrast, our MGN module integrates features from LiDAR and SfM point clouds to learn accurate 3D scene geometry. In addition, a small number of randomly sampled points are introduced to enhance the generalization capability of model.

Furthermore, the geometric guidance strategy enables smoother and more consistent geometric boundaries in reconstruction. Without sky masking and directed-depth constraints, distant details are lost and Gaussian aliasing occurs. Due to the complex illumination conditions in large-scale outdoor scenes, the absence of lighting enhancement leads to unstable quantitative rendering results. Ultimately, the results show that each component plays a corresponding role in the proposed model, further improving reconstruction and rendering quality.

VI. CONCLUSION

In this work, we propose a novel LSADS-Gaussian model for large-scale autonomous driving scene reconstruction based on 3DGS. The LSADS-Gaussian mainly consists of three sub-modules. The Multimodal Gaussian Network (MGN) module optimizes the integration of camera images and LiDAR point clouds, enabling effective aggregation of 3D Gaussians for learning precise scene geometry. We further employ a Geometric Representation Guidance (GRG) module that refines Gaussian positions and geometric

features by updating their directed-depth under the constraints of depth and normal priors, while applying sky masking to separately model infinitely distant regions and avoid their disruption of geometric consistency. Finally, we introduce a Lighting Enhancement (LE) module during rendering to alleviate the negative impact of illumination instability on scene reconstruction. Extensive experiments demonstrate that our method outperforms current state-of-the-art techniques for large-scale autonomous driving scene reconstruction. In future work, we will further study the limitations posed by transparent and dynamic objects in scene reconstruction.

ACKNOWLEDGMENTS

This work was supported in part by International Strategic Innovative Project of National Key R&D Program of China (2023YFE0112500), International Internship Program of National Institute of Informatics, and the JSPS Grant-in-Aid for Early-Career Scientists under Grant JP25K21195.

REFERENCES

[1] L. Chen *et al.*, "Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046-1056, 2023.

[2] Y. Li *et al.*, "BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1477-1485, 2023.

[3] Z. Li *et al.*, "BEVFormer: Learning Bird's-Eye-View Representation from LiDAR-Camera via Spatiotemporal Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2020-2036, 2025.

[4] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMapNet: An Online HD Map Construction and Evaluation Framework," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 4628-4634, 2022.

[5] Y. Liu *et al.*, "VectorMapNet: End-to-end Vectorized HD Map Learning," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pp. 22352-22369, 2023.

[6] L. Jing *et al.*, "STT: Stateful Tracking with Transformers for Autonomous Driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4442-4449, 2024.

[7] B. Mildenhall *et al.*, "NeRF: representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[9] J. T. Barron *et al.*, "Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields," in *CVPR*, pp. 5470-5479, 2022.

[10] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1-15, 2022.

[11] Z. Li *et al.*, "Neuralangelo: High-Fidelity Neural Surface Reconstruction," in *CVPR*, pp. 8456-8465, 2023.

[12] D. Chen *et al.*, "PGSR: Planar-Based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 9, pp. 6100-6111, 2025.

[13] T. Lu *et al.*, "Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering," in *CVPR*, pp. 20654-20664, 2024.

[14] Z. Yu *et al.*, "Mip-Splatting: Alias-free 3D Gaussian Splatting," in *CVPR*, pp. 19447-19456, 2024.

[15] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction," in *CVPR*, pp. 19457-19467, 2024.

[16] M. Tancik *et al.*, "Block-NeRF: Scalable Large Scene Neural View Synthesis," in *CVPR*, pp. 8248-8258, 2022.

[17] Z. Yang *et al.*, "UniSim: A Neural Closed-Loop Sensor Simulator," in *CVPR*, pp. 1389-1399, 2023.

[18] M. Turkulainen *et al.*, "DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing," in *WACV*, pp. 2421-2431, 2025.

[19] R. Martin-Brualla *et al.*, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *CVPR*, pp. 7210-7219, 2021.

[20] Z. Wang *et al.*, "Neural Fields Meet Explicit Geometric Representations for Inverse Rendering of Urban Scenes," in *CVPR*, pp. 8370-8380, 2023.

[21] Z. Mi and D. Xu, "Switch-NeRF: Learning Scene Decomposition with Mixture of Experts for Large-scale Neural Radiance Fields," in *ICLR*, 2023.

[22] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-NERF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs," in *CVPR*, pp. 12922-12931, 2022.

[23] Z. Wu *et al.*, "MARS: An Instance-Aware, Modular and Realistic Simulator for Autonomous Driving," in *CICAI*, pp. 3-15, 2024.

[24] Y. Yan *et al.*, "Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting," in *ECCV*, pp. 156-173, 2025.

[25] C. Peng *et al.*, "DeSiRe-GS: 4D Street Gaussians for Static-Dynamic Decomposition and Surface Reconstruction for Urban Driving Scenes," in *CVPR*, pp. 6782-6791, 2025.

[26] X. Zhou *et al.*, "DrivingGaussian: Composite Gaussian Splatting for Surrounding Dynamic Autonomous Driving Scenes," in *CVPR*, pp. 21634-21643, 2024.

[27] J. Cao, Z. Li, N. Wang, and C. Ma, "Lightning NeRF: Efficient Hybrid Scene Representation for Autonomous Driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16803-16809, 2024.

[28] H. Mu, G. Zhang, M. Zhou, and Z. Cao, "End-to-end Semantic Segmentation Network for Low-Light Scenes," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7725-7731, 2024.

[29] J. L. Schonberger and J.-M. Frahm, "Structure-From-Motion Revisited," in *CVPR*, 2016.

[30] P. Sun *et al.*, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *CVPR*, 2020.

[31] J. T. Barron *et al.*, "Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields," in *ICCV*, pp. 5855-5864, 2021.

[32] K. Rematas *et al*., "Urban Radiance Fields," in *CVPR*, pp. 12932-12942, 2022.

[33] Z. Xie *et al.*, "S-NeRF: Neural Radiance Fields for Street Views," *arXiv preprint arXiv:2303.00749*, 2023.

[34] J. Yang *et al.*, "EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Super vision," *arXiv preprint arXiv:2311.02077*, 2023.

[35] P. Wang *et al.*, "NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi- view Reconstruction," *arXiv preprint arXiv: 2106.10689*, 2021.

[36] B. Huang *et al.*, "2D Gaussian Splatting for Geometrically Accurate Radiance Fields," in *ACM SIGGRAPH*, 2024.