

Street Gaussians without 3D Object Tracker

Ruida Zhang^{1,2}, Chengxi Li¹, Chenyangguang Zhang¹, Xingyu Liu¹, Haili Yuan¹,
 Yanyan Li², Xiangyang Ji¹, Gim Hee Lee²

¹Tsinghua University, ²National University of Singapore

{zhangrd23@mails, xyji@}.tsinghua.edu.cn, gimhee.lee@nus.edu.sg

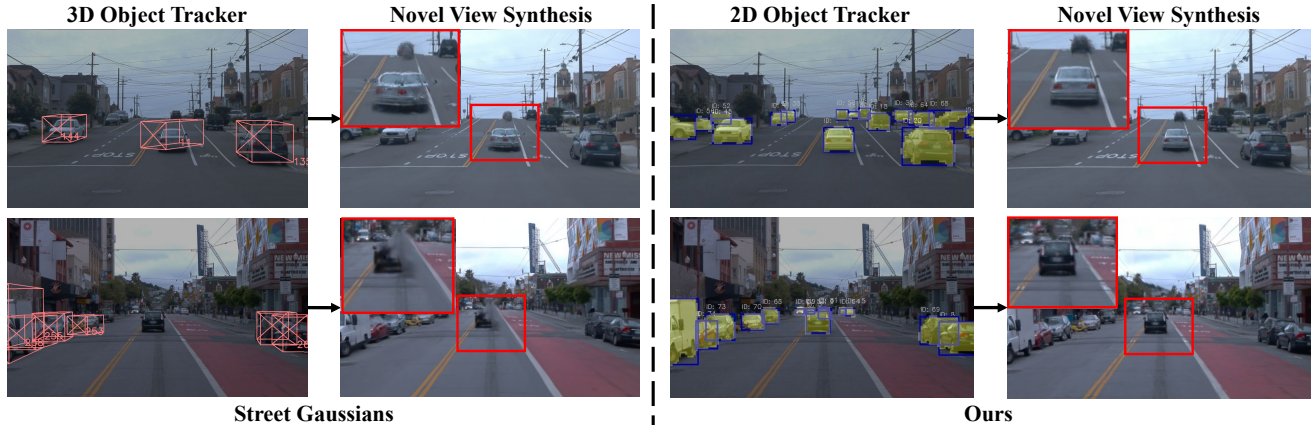


Figure 1. Comparison of 3D tracker-based Street Gaussians [89] (left) and our approach (right). Existing methods heavily rely on object poses, but 3D trackers struggle with limited generalization [21, 73, 97], leading to flaws in novel view synthesis. In contrast, 2D foundation models show better generalization [84, 88]. Our approach leverages a 2D foundation model [84] for object tracking and learns point motion within an implicit feature space to autonomously correct tracking errors, improving robustness across diverse scenes.

Abstract

Realistic scene reconstruction in driving scenarios poses significant challenges due to fast-moving objects. Most existing methods rely on labor-intensive manual labeling of object poses to reconstruct dynamic objects in canonical space and move them based on these poses during rendering. While some approaches attempt to use 3D object trackers to replace manual annotations, the limited generalization of 3D trackers – caused by the scarcity of large-scale 3D datasets – results in inferior reconstructions in real-world settings. In contrast, 2D foundation models demonstrate strong generalization capabilities. To eliminate the reliance on 3D trackers and enhance robustness across diverse environments, we propose a stable object tracking module by leveraging associations from 2D deep trackers within a 3D object fusion strategy. We address inevitable tracking errors by further introducing a motion learning strategy in an implicit feature space that autonomously corrects trajectory errors and recovers missed detections. Experimental results on Waymo-NOTR and KITTI show that our method outperforms existing approaches. Our code will be made publicly available.

1. Introduction

Modeling dynamic 3D street scenes underpins modern autonomous driving by enabling realistic, controllable simulations for tasks such as perception [11, 42, 56, 94, 105], prediction [25, 33, 47], and motion planning [13, 14, 16]. With the rise of end-to-end autonomous systems that require real-time sensor feedback [29, 31, 35], high-quality scene reconstructions have become essential for closed-loop evaluations [46, 96], particularly to simulate critical corner cases safely and cost-effectively.

Despite extensive efforts in achieving photo-realistic reconstruction of small-scale scenes, the unique challenges posed by the large-scale and highly dynamic nature of driving scenarios complicate effective 3D scene modeling. To address these challenges, most existing methods [22, 38, 65, 85, 106] rely on ground truth vehicle poses to differentiate between static background and moving vehicles. Typically, vehicles are reconstructed in canonical space and subsequently positioned based on known poses during rendering. However, collecting ground-truth poses is labor-intensive, limiting the applicability of these methods to scenes beyond the existing datasets.

To eliminate reliance of the systems on ground truth object poses, Street Gaussians [89] instead uses poses generated by 3D object trackers [82, 83]. However, by defining vehicle motion solely through these tracking poses, rendering quality becomes highly dependent on pose accuracy. While Street Gaussians optimizes object poses during training, it struggles with detection failures and large pose errors, as shown in Fig. 1. Unfortunately, 3D trackers often struggle to generalize effectively across different scenarios, primarily due to the scarcity of open-source large-scale 3D datasets [79, 103]. Collecting ground-truth poses for 3D tracking is time-consuming and costly, resulting in relatively few open-source datasets for autonomous driving. For example, even widely-used datasets such as Waymo [76], Nuscenes [6], KITTI [23, 49], Pandaset [86] collectively contain fewer than 4,000 annotated scenes, limiting the diversity and scale needed for robust 3D tracking. In contrast, 2D data is far easier to collect and annotate, leading to an abundance of 2D datasets. The BDD100K dataset [95] alone includes annotations for 100,000 driving scenes, and additional large-scale 2D perception datasets, such as COCO [50], LVIS [28], and OpenImages [41], bring the total number of annotated scenes into the millions. This wealth of 2D data has enabled 2D foundation models to achieve strong generalization across a variety of tasks, including embedding extraction [64, 68], detection [8, 62], segmentation [12, 40], and tracking [84]. Nonetheless, how to use these accurate and robust 2D trackers to effectively model dynamic objects in 3D is still an open challenge for street scene reconstruction.

In this paper, a novel architecture is designed to achieve high-fidelity novel view synthesis performance in street scenarios for autonomous driving applications. Unlike previous approaches such as Driving Gaussian [106] and Street Gaussian [89] that rely on ground truth vehicle poses or poses predicted by 3D trackers, the first contribution of this work is a stable object tracking module. This module leverages associations from 2D deep trackers within a 3D object fusion framework to enhance robustness and accuracy. Specifically, we integrate 2D tracking outputs with LiDAR data to trace the trajectory of each vehicle in 3D. We then incrementally reconstruct the point cloud of each vehicle frame-by-frame in the canonical space, estimating its pose by aligning successive frames with this canonical model. This approach eliminates the reliance on 3D trackers and enhances robustness across diverse environments.

Although 2D tracking models demonstrate stronger generalization ability, tracking errors still can exist, especially under adverse conditions such as severe occlusions or distant objects. Since moving the object points solely based on the tracked object pose directly exposes any tracking errors, we aim to go beyond a straightforward reliance on tracked trajectories. Instead, we propose to *learn* point mo-

tion from the predicted trajectory, equipping the model to autonomously identify and correct tracking inaccuracies, recover missed detections, and infer motion in new time steps. An implicit representation is essential for this purpose: 1) It enables the model to refine trajectories without being constrained by bounding box tracks, facilitating smoother and more continuous motion that can adaptively respond to changes. 2) An implicit feature space offers versatility to move each point of an object in a different way, recognizing that vehicles are not strictly rigid objects — for example, doors can open or close. This approach makes it possible to capture subtle, dynamic changes within objects, ultimately producing more accurate reconstructions and enhancing the robustness of novel view synthesis in challenging scenarios.

To this end, we leverage HexPlane representation [7] following 4DGS [81]. HexPlane stores motion-related features by decomposing the 4D spatial-temporal space into six 2D learnable feature planes. A decoder then utilizes these features to predict deformation offsets, dynamically adjusting 3D Gaussians over time. However, 4DGS relies solely on image reconstruction loss for supervision, which is insufficient for handling rapid object motion. When a Gaussian’s initial position is far from its current location, its projection on the image may fall outside the object’s actual area, preventing gradient propagation and hindering optimization. This issue is evident in S3Gaussian [32], which struggles with moving cars due to the lack of explicit motion supervision. To address this, we introduce a training strategy that supervises learned point motion using the predicted trajectory, providing explicit 3D supervision to guide HexPlane in capturing motion dynamics more accurately.

Our main contributions are summarized as follows:

- We eliminate the need of 3D tracker for street scene reconstruction by introducing a robust object tracking module which leverages 2D foundation model, achieving superior generalization ability across diverse scenarios.
- We introduce a motion learning framework to learn from the predicted trajectory in an implicit feature space, enabling it to automatically correct pose errors and infer motion for novel time steps.
- We outperform existing methods on Waymo-NOTR and KITTI datasets without relying on ground truth annotations.

2. Related Works

2.1. 3D Gaussian Splatting for Dynamic Scene

3D Gaussian Splatting (3DGS) [37, 44] has advanced scene reconstruction by enabling high-quality, real-time rendering with 3D Gaussians and efficient splat-based rasterization, reducing computation and parameters compared to NeRF-based methods [1–3, 61, 63] and other representa-

tions [15, 26, 27, 30, 45, 75, 102]. While originally designed for static scenes, 3DGS has been adapted for dynamic scenes [48, 53, 57, 74, 81, 92, 93]. Dynamic3DGS [57] directly stores information for each 3D Gaussian at every timestamp, and Yang *et al.* [92] approximate the spatiotemporal 4D volume by optimizing 4D Gaussian primitives. Deformable-3DGS [93] and GauFRe [48] employ deformation fields to model motions, while 4DGS [81] introduces HexPlane representation [7] to store spatial-temporal features efficiently. Gaussian Marbles [74] leverages isotropic “marbles” and a divide-and-conquer trajectory learning algorithm. Despite these advances, challenges persist in high-speed driving scenarios with rapid object motion. To tackle this challenge, we leverage a robust object tracking module to guide the motion learning process.

2.2. Street Scene Reconstruction

Existing autonomous driving simulation engines [20, 43, 52, 71] face high manual effort in creating virtual environments and a lack of realism in the generated data. The creation of high-fidelity simulations from driving logs is therefore essential for advancing closed-loop training and testing. Recent works [65, 72, 77, 78, 85, 87, 90, 91] have continuously made improvements to NeRF [61] to model street scenes dynamically. Despite the progress, NeRF-based methods remain computationally expensive and require densely overlapping views. Building on the effective 3DGS approach [37] for scene reconstruction, several 3DGS-based methods [10, 67, 89, 106] have emerged. PVG [10] model dynamic scenarios by using periodic vibration-based temporal dynamics. Driving Gaussian [106] introduces incremental static Gaussians and composite dynamic Gaussian graph. Street Gaussians [89] equips Gaussians with semantic logits, and optimize dynamic parts using tracked poses from 3D tracker. AutoSplat [38] enforces geometric constraints in road and sky regions for multi-view consistency. Among these approaches, most of them requires ground truth object pose [10, 38, 106] or the 3D tracker [89]. However, manual annotation is laborious and the 3D tracker is lack in generalization ability, which limits their applications in diverse scenarios. In contrast, S3Gaussian [32] models dynamics in a self-supervised manner using dynamic Gaussians from 4DGS [81], despite struggling with dynamic object modeling due to the lack of explicit motion supervision. We address this challenge by introducing a robust object tracking strategy based on a 2D foundation model [84] and apply motion supervision from the predicted object trajectory.

2.3. 2D and 3D trackers

Multiple Object Tracking (MOT) aims at locating multiple objects in each frame, and establishing correspondences between them across frames within input videos

[4, 5, 36, 58, 98]. The reliance of most MOT methods on object detection [4, 19, 58, 66, 98, 104] makes detection accuracy crucial for MOT performance. Recent advances in 3D object detection are promising [11, 18, 34, 42, 51, 54, 56, 60, 80, 94, 99–101], but these models often generalize poorly due to limited and biased datasets [21, 55, 73, 97]. Existing open-source autonomous driving datasets lack scale and show regional or environmental biases, such as vehicle density and weather conditions [6, 23, 59, 76]. The main reason is that building large-scale, well-annotated multimodal datasets is costly [6, 49, 59, 69]. In contrast, 2D image data is easier to capture and annotate, enabling extensive dataset collections [9, 17, 41, 95]. Visual foundation models trained on large-scale 2D data have shown strong generalization [12, 40, 64, 68, 84]. We thus opt to use 2D object tracker [84] to locate the dynamic objects.

3. Our Method

3.1. Prerequisites: 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [37] represents a 3D scene explicitly as a collection of 3D Gaussian primitives. Each Gaussian is defined by its center position $\mathcal{X} \in \mathbb{R}^3$ and covariance $\Sigma \in \mathbb{R}^{3 \times 3}$. To enforce semi-positive definiteness, the covariance matrix is further factorized into a scaling vector $S \in \mathbb{R}^3$ and a rotation matrix $R \in SO(3)$ by:

$$\Sigma = RSS^\top R^\top \quad (1)$$

Additionally, each Gaussian is described with its opacity $o \in \mathbb{R}$ and view-dependent color defined by spherical harmonic (SH) coefficient $C \in \mathbb{R}^k$, where k represents numbers of SH functions.

During rendering of novel views, differential splatting is applied to 3D Gaussians within the camera planes. The blending of N ordered points that overlap a pixel is given by the formula:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

where α_i and c_i represents the opacity and color of the i th splatted Gaussian, which is computed from per-point opacity and SH coefficients (see [37] for details).

3.2. Overview

Our method takes as input multi-view images $\mathbf{I}_t^{(j)}$ from multiple cameras placed around the vehicle, each indexed by time step t and camera index j , along with intrinsic $\mathbf{K}^{(j)}$ and extrinsic $\mathbf{E}_t^{(j)}$ matrices for each view. Additionally, a top-mounted LiDAR provides 3D point clouds \mathbf{L}_t for each frame. Using this multi-sensor data, our approach reconstructs the 3D scene and synthesizes novel views from any

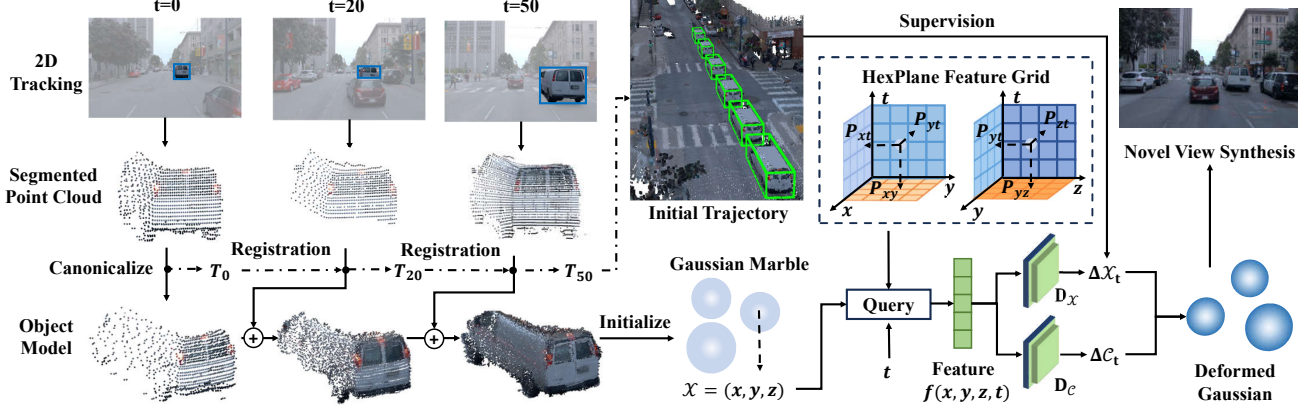


Figure 2. **Overview of our method.** To overcome the limited generalization of 3D object trackers, we introduce a robust object tracking module based on a 2D object tracker [84]. We integrate 2D tracking with LiDAR data to segment the object’s point cloud and incrementally reconstruct the object model in canonical space frame-by-frame. The canonical model is used to estimate object pose \mathbf{T}_t and serves as the initialization for Gaussian optimization. We model dynamic objects with isotropic Gaussian marble [74] to simplify optimization. Our approach learns point motion within the HexPlane [7] feature space, enabling tracking error correction and missed detection recovery. Decoders D_x, D_c predict point motion $\Delta\mathcal{X}_t$ and color change $\Delta\mathcal{C}_t$ based on the HexPlane features $f(x, y, z, t)$, with the predicted trajectory providing supervision. Finally, novel view synthesis is performed by splatting the deformed Gaussian onto the image plane.

desired camera pose and time frame, without reliance on 3D object trackers or ground truth object trajectories.

We present an overview of our method in Fig. 2. Visual images and point clouds are obtained from the sensor setup. For images, static and dynamic components are analyzed first. We employ Mask2Former [12] to segment the scene into static and dynamic parts. Specifically, the dynamic part includes humans (pedestrians, cyclists, *etc.*) and vehicles (cars, trucks, *etc.*), and all the other objects are categorized as static. We model the static part as in 3DGS [37] and focus on modeling dynamic objects in the following sections. To enhance robustness, we propose an object tracking strategy by leveraging a robust 2D tracker [84] and the LiDAR point cloud, providing accurate initial point clouds and object trajectories for subsequent Gaussian optimization (*cf.* Sec 3.3). We leverage an implicit representation HexPlane [7] following 4DGS [81] to learn a continuous and smooth per-point motion based on the object trajectory (*cf.* Sec 3.4). The optimization objective is described in Sec 3.5.

3.3. Tracking Cars in 3D with Robust 2D Tracker

To improve the robustness of novel view synthesis for diverse scenarios, the first challenge is how to track and build vehicles in 3D based on the 2D tracking results.

We employ GLEE [84] to track all vehicles within the 2D image plane, obtaining 2D object trajectories in each camera view. Each trajectory includes the 2D segmentation masks of the object over a time period. To lift 2D trajectories into 3D, we first re-project the LiDAR point clouds onto the image plane and assign points to objects based on their presence within the corresponding 2D segmentation masks \mathcal{M} . This yields a segmented object point clouds for each

camera view via the following function:

$$\mathcal{O}_t^{(j)} = \{\mathbf{P} | \Pi^{-1}(\mathbf{P}, \mathbf{K}^{(j)}, \mathbf{E}_t^{(j)}) \in \mathcal{M}, \mathbf{P} \in \mathbf{L}_t\}, \quad (3)$$

where j is the camera index, t is the time index, \mathbf{P} is a LiDAR point and $\Pi(\cdot)$ is the re-projection function. Given the slight misalignment between the LiDAR point cloud and RGB image, points near the edges of segmentation masks often fall outside the object boundaries. To address this issue, we perform outlier removal to improve stability in subsequent reconstruction based on the point distance to the point cloud center. We then associate the same objects across different camera views. Two point clouds in different viewpoints are considered belonging to the same object if they share more than 50 points in one time step. This association process results in a set of associated partial object point cloud in different time steps.

After obtaining the 3D partial objects in different time steps, the next step is to associate these partial point cloud into a unified and complete model. We initialize the 3D reconstruction of each object from the first frame which it appears. For simplicity, we assume an object is visible between frames 0 and T and denote the object point cloud at time t as \mathcal{O}_t . Our goal is to obtain a temporally consistent reconstruction \mathcal{O} and the object pose in each time frame. This reconstruction is done frame-by-frame incrementally.

Starting with the initial frame, we add \mathcal{O}_0 to \mathcal{O} . For each subsequent frame, we apply Iterative Closest Point (ICP) for the alignment of \mathcal{O}_t to \mathcal{O} by extracting the relative pose \mathbf{T}_t . The overlap between \mathcal{O}_t and \mathcal{O} is given by:

$$\tau_{\text{overlap}} = \mathcal{O} \cap \mathbf{T}_t^{-1} \mathcal{O}_t, \quad \mathbf{T}_t = \text{ICP}(\mathcal{O}_t, \mathcal{O}). \quad (4)$$

If ICP reveals an overlap τ_{overlap} of more than 30% with \mathcal{O} , we record the object pose \mathbf{T}_t and update \mathcal{O} by integrating $\mathbf{T}_t^{-1}\mathcal{O}_t$. If τ_{overlap} is between 10% and 30%, we only record the object pose P_t without updating \mathcal{O} . If τ_{overlap} is lower than 10%, we assume a 2D tracking failure for that frame and discard it to filter out tracking errors. This process is repeated for each frame using the last recorded pose as the initial guess for ICP to accelerate convergence. Through this approach, we reconstruct all objects and retrieve their poses in each frame, providing a robust initialization for subsequent 3D Gaussian optimization.

3.4. Learning Point Motion

A common approach to model object motion in prior works is directly transforming the object points using the given object pose. However, despite the strong generalization capabilities of 2D trackers, detection failures are inevitable in challenging scenarios, such as when objects are heavily occluded or located at great distances. Relying solely on object poses derived from these erroneous or missing detections can easily lead to rendering failures. Additionally, treating cars as rigid objects does not adequately address corner cases, such as when a car door is open or closed. Furthermore, these methods lack the ability to infer an object motion at arbitrary time stamps.

To address these limitations, we seek for a motion modeling approach to enhance robustness and flexibility. Instead of explicitly using object pose as rigid transformation, we learn the per-point motion in a pre-defined feature space. In this feature space, the object motion can be optimized through both explicit guidance and photometric loss. Moreover, the object motion can be interpolated through time and space in this feature space to compensate for missing detections. To this end, we leverage HexPlane representation [7] as in 4DGS [81] to efficiently capture spatial and temporal information by decomposing the 4D feature voxels into six learnable feature planes. This representation satisfies all our requirements and is also memory efficient.

We use three planes: P_{xy} , P_{yz} , and P_{xz} for the spatial dimensions, and another three planes P_{xt} , P_{yt} , and P_{zt} for the spatial-temporal features. Additionally, the Hexplane includes multiple resolution levels which is formulated as:

$$\{P_{ij}^\rho \in \mathbb{R}^{d \times \rho r_i \times \rho r_j} | (i, j) \in \mathcal{P}, \rho \in \{1, 2\}\}, \quad (5)$$

where $\mathcal{P} = \{(x, y), (x, z), (y, z), (x, t), (y, t), (z, t)\}$, d is the feature dimension, ρ denotes the upsampling scale, and r is the base resolution.

Given the center position of a Gaussian (x, y, z) and the time step t , features in all six planes are queried and combined via a small MLP ϕ_d as follows:

$$f(x, y, z, t) = \phi_d\left(\bigcup_{\rho} \prod_{(i, j) \in \mathcal{P}} \pi(P_{ij}^\rho, \psi_{ij}^\rho(x, y, z, t))\right), \quad (6)$$

where $\psi_{ij}^\rho(x, y, z, t)$ projects the 4D coordinate onto the respective plane, and π performs bilinear interpolation on the voxel features at each point. The features of all planes are combined using the Hadamard product.

Finally, two MLP decoders $D_{\mathcal{X}}, D_{\mathcal{C}}$ are utilized to predict the point motion $\Delta\mathcal{X}_t$ and color change $\Delta\mathcal{C}_t$ as $\Delta\mathcal{X}_t = D_{\mathcal{X}}(f(x, y, z, t))$ and $\Delta\mathcal{C}_t = D_{\mathcal{C}}(f(x, y, z, t))$, respectively. The deformed 4D Gaussians are formulated as: $\mathcal{G}_t = \{\mathcal{X} + \Delta\mathcal{X}_t, \mathcal{C} + \Delta\mathcal{C}_t, S, R, o\}$.

To allow the HexPlane feature to learn point motion from the predicted pose, we define the motion loss as:

$$\mathcal{L}_{\text{motion}} = \text{avg}_{\mathcal{X} \in \mathcal{O}} |\Delta\mathcal{X}_t - (\mathbf{T}_t\mathcal{X} - \mathcal{X})|, \quad (7)$$

where \mathcal{X} is the center position of a Gaussian in object \mathcal{O} and avg is the average operator. This loss encourages the predicted point motion $\Delta\mathcal{X}_t$ to align with the predicted pose. We apply this loss only for the first 40% iterations to provide a strong initial motion prior. The loss is subsequently removed, allowing the network to adjust and potentially compensate for pose errors and detection failures.

Following [74], we adopt isotropic Gaussian marbles for dynamic points to reduce degrees of freedom and simplify optimization. In this setup, the rotation of each Gaussian is represented by the identity matrix with identical scales across all three dimensions. The spherical harmonics coefficients are also limited to three dimensions, providing view-consistent color. This approach ensures that all deformations are captured purely by point motion and color changes, enforcing strong constraints and enhancing robustness in novel view synthesis.

Ours vs. S3Gaussian. Although S3Gaussian [32] also uses Hexplane representation [7] and the deformation network from 4DGS [81], our design serves a different purpose: S3Gaussian learns point motion without extra supervision, while we use 4DGS to refine motion and address detection failures. Consequently, S3Gaussian becomes susceptible to unsatisfactory results with rapid car motion due to the lack of explicit guidance. In contrast, our approach enables high-quality reconstruction of dynamic objects.

3.5. Optimization Objective

Besides the motion loss introduced above, the loss function comprises five components to collectively optimize the scene representation and the point motion. The primary loss \mathcal{L}_{rgb} is an L1 loss measuring the photometric difference between rendered and ground truth images while $\mathcal{L}_{\text{ssim}}$ evaluates their structural similarity. The L1 loss between the rendered depth map and the depth estimated from the LiDAR point cloud $\mathcal{L}_{\text{depth}}$ is used to supervise the Gaussian positions. Following K-Planes [70], a grid-based total variation loss \mathcal{L}_{tv} is introduced. Recognizing that the color of most dynamic points in the scene are unchanged, a L1 regulariza-

Task		Scene Reconstruction					Novel View Synthesis				
Method	Extra Input	PSNR↑	SSIM↑	LPIPS↓	DPSNR↑	DSSIM↑	PSNR↑	SSIM↑	LPIPS↓	DPSNR↑	DSSIM↑
3DGS [37]	N/A	25.77	0.833	0.160	20.26	0.604	24.46	0.802	0.170	18.12	0.521
EmerNeRF [90]	N/A	28.16	0.806	0.228	24.32	0.682	25.14	0.747	0.313	23.49	0.660
S3Gaussian [32]	N/A	31.35	0.911	0.106	26.02	0.783	27.44	0.857	0.137	22.92	0.680
MARS [85]	GT pose	28.24	0.866	0.252	23.37	0.701	26.61	0.796	0.305	22.21	0.697
StreetGS [89]	3D tracker	29.17	0.873	0.138	27.78	0.818	26.98	0.838	0.149	24.62	0.742
Ours	2D tracker	32.56	0.936	0.059	28.51	0.868	28.85	0.867	0.088	25.58	0.779

Table 1. Comparison with state-of-the-art methods on Waymo-NOTR dataset. StreetGS represents Street Gaussian [89]. The **best** and the **second best** results are denoted by pink and blue.

tion loss $\mathcal{L}_{\text{color-reg}}$ is applied to the deformation network by minimizing $\Delta\mathcal{C}$.

The total loss function is thus defined as:

$$\begin{aligned} \mathcal{L} = & \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} \\ & + \lambda_{\text{color-reg}}\mathcal{L}_{\text{color-reg}} + \lambda_{\text{motion}}\mathcal{L}_{\text{motion}}, \end{aligned} \quad (8)$$

where the weights are assigned as follows: $\lambda_{\text{rgb}} = 1.0$, $\lambda_{\text{depth}} = 1.0$, $\lambda_{\text{ssim}} = 0.1$, $\lambda_{\text{tv}} = 0.1$, $\lambda_{\text{color-reg}} = 0.01$ and $\lambda_{\text{motion}} = 1$.

The optimization process is divided into two stages. In the first stage, we render using only static Gaussians and supervise only the static regions of the images over 20,000 iterations. This phase focuses on reconstructing the static background and stabilizes the training process. In the second stage, we train all Gaussians on the whole images for 50,000 iterations.

4. Experiments

4.1. Experimental Setup

Datasets. We leverage Waymo-NOTR dataset [76, 90] and KITTI benchmark [23, 24]. The **NeRF On-The-Road (NOTR)** dataset, introduced by EmerNeRF [90], is a subset of the Waymo Open dataset [76] and includes diverse challenging driving scenarios, such as high-speed, exposure mismatch, and various weather conditions. We use the dynamic32 subset, consisting of 32 dynamic scenes, for evaluating dynamic reconstruction. Following EmerNeRF’s setup [90], we use three frontal camera images resized to 960×640 . For scene reconstruction, all image frames are used for training and evaluation; for novel view synthesis, every 10th time step is excluded for evaluation [90]. For the KITTI dataset, we follow the setup of MARS [85] to use 75% or 50% of the images for training with every 4th or every 2nd frame held out for testing, respectively.

Implementation Details. We use the Adam optimizer [39] with the same learning rate schedule as in 3DGS [37]. For long-sequence scene reconstruction, we follow S3Gaussian

[32] and segment the scene into multiple clips. The multi-resolution HexPlane encoder has a base resolution of 64, upsampled by factors of 2 and 4 as in [81], and other hyperparameters match those in 3DGS [37]. As LiDAR points lack data for the sky region, we add a plane of points above the maximum height of the scene to represent it following [38]. All experiments are run on a single NVIDIA RTX 3090 GPU, with training taking about 2 hours per video clip and inference speed at 100 FPS at 960×640 resolution.

Baseline Methods. We compare our approach against state-of-the-art methods, including NeRF-based methods (MARS [85], NSG [65], EmerNeRF [90]) and Gaussian-based methods (3DGS [37], StreetGaussian [89], S3Gaussian [32]). To ensure fair comparisons, we apply LiDAR point cloud initialization and add depth regularization to 3DGS. For the Waymo-NOTR dataset, we borrow results of MARS [85] and EmerNeRF [90] from S3Gaussian [32]. We evaluate the 3D-tracker-based method, Street Gaussians [89], using detection results from VoxelNext [11] and associate the detections with SimpleTrack [66]. Our goal is to assess the generalization ability of 3D-tracker-based methods in scenarios without ground truth tracking labels. To emulate a real-world scenario where a 3D tracker trained on a public benchmark dataset is applied to a novel environment, we use pretrained weights from nuScenes [6]—one of the largest and most diverse 3D datasets—and evaluate the model on the Waymo-NOTR dataset. Since NOTR is part of the training set for the Waymo perception task, using Waymo pretrained weights would introduce domain overlap and bias the evaluation, making the nuScenes weights a fairer choice. For a fair comparison, we use pretrained 2D tracker weights from [84] for our method, which have not been trained on the Waymo or KITTI datasets. Please refer to the supplementary material for discussion of the choice of 3D trackers. For the KITTI dataset, we borrow the results of all other methods from Street Gaussians. Please note that Street Gaussians leverages GT tracking annotations on KITTI dataset.



Figure 3. Qualitative comparison of novel view synthesis on NOTR dataset. Best viewed with zoom.

Metrics. We leverage peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and learned perceptual image patch similarity (LPIPS) to evaluate the rendering quality. Additionally, following EmerNeRF [90], we apply DPSNR and DSSIM metrics to dynamic objects by projecting their ground truth 3D bounding boxes onto the 2D image plane and calculating pixel loss within these regions.

4.2. Comparisons with the State-of-the-art

On the NOTR dataset, our method outperforms all competitors across every metric, as shown in Table 1. Our approach sets a new state-of-the-art in both the scene reconstruction and novel view synthesis tasks. Specifically, our method outperforms S3Gaussian [32] by 2.66 dB in DPSNR and 0.099 in DSSIM for novel view synthesis. It also surpasses Street Gaussians [89] by 1.87 dB in PSNR, mainly due

to the limited generalization of 3D trackers used in Street Gaussians. Despite optimizing object poses, Street Gaussians struggles with large pose errors and detection failures, leading to inferior performance. Additionally, we outperform NeRF-based methods like EmerNeRF and MARS.

We present qualitative comparison with S3Gaussian [32] and Street Gaussians [89] in Fig. 3. Results from S3Gaussian show that using 4DGS without explicit motion guidance results in weaker performance when handling moving vehicles (*e.g.* row 2, 3, 5). Street Gaussians suffers from tracking errors of 3D trackers, leading to erroneous reconstructions (*e.g.* row 1, 3) or entirely missed objects (*e.g.* rows 2, 4, and 5). In contrast, our approach performs robustly across diverse scenarios owing to our tracking strategy leveraging 2D foundation model and robust motion

		KITTI-75%			KITTI-50%		
Method	Extra Input	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS [37]	N/A	19.19	0.737	0.172	19.23	0.739	0.174
NSG [65]	GT pose	21.53	0.673	0.254	21.26	0.659	0.266
MARS [85]	GT pose	24.23	0.845	0.160	24.00	0.801	0.164
Street Gaussians [89]	GT pose	25.79	0.844	0.081	25.52	0.841	0.084
Ours	2D tracker [84]	25.49	0.889	0.063	25.11	0.877	0.067

Table 2. Comparison with state-of-the-art methods on KITTI dataset. The best and the second best results are denoted by pink and blue.

	Method	PSNR	SSIM	DPSNR	DSSIM
A	w/o track	25.25	0.761	21.71	0.662
B	w/o $\mathcal{L}_{\text{motion}}$	28.39	0.843	24.54	0.719
C	w/o iso.	28.58	0.847	25.35	0.765
D	3D tracker	28.02	0.835	24.61	0.746
E	D w/o $\mathcal{L}_{\text{motion}}$	27.54	0.821	23.53	0.703
F	GT pose	28.98	0.873	25.71	0.787
G	UNINEXT	28.76	0.864	25.43	0.772
H	Ours	28.85	0.867	25.58	0.779

Table 3. Ablation study of Novel View Synthesis on NOTR dataset.

learning module. Please refer to the supplementary material for **additional video results** and **editing examples**.

The results on the KITTI dataset are presented in Table 2. Our method outperforms the NeRF-based NSG and MARS across all metrics. In the KITTI-75% setting, our approach achieves a 0.045 higher SSIM and a 0.018 lower LPIPS compared to Street Gaussians, although its PSNR is 0.3 dB lower. A similar trend is observed in the KITTI-50% setting. This lower PSNR is primarily because Street Gaussians leverages labor-intensive ground truth tracking annotations, whereas our method uses a generalized 2D tracker—yet still attains comparable performance.

4.3. Ablation Studies

We conduct an ablation study on the NOTR dataset and the results are presented in Tab. 3.

Effect of object tracking module. In Tab. 3 (A), we model dynamic points using 4DGS [81] without performing any form of tracking, and observe significant drops across all metrics. This experiment highlight that the sole use of 4DGS does not provide sufficient accuracy for effective motion modeling in autonomous driving scenarios.

Effect of motion learning strategy. In Tab. 3 (B), we use the reconstruction of the tracking module for initialization

but remove the motion loss introduced in Eq. 7, resulting in a 1.04 dB drop in DPSNR. This result highlights the importance of our motion learning strategy. A comparison with rigid-transformation-based motion modeling [89] is provided in the supplementary material.

Effect of isotropic Gaussian marbles. In Tab. 3 (C), we substitute isotropic Gaussian marbles with the original anisotropic Gaussian ellipsoids for dynamic points, which leads to a decrease in DPSNR and DSSIM.

3D tracker / GT pose vs. Our object tracking module. In Tab. 3 (D), we use the 3D tracker [11, 66] to reconstruct dynamic objects and supervise motion learning, resulting in a 0.97 dB DPSNR decrease. Removing motion loss leads to a further 1.08 dB drop (Tab. 3 (E)), showing that our motion learning strategy compensates for tracking errors, though these errors still impact novel view synthesis, emphasizing the need for a more generalizable 2D tracker. Using ground-truth trajectories (Tab. 3 (F)) shows only a small gap, confirming the effectiveness of our 2D tracker in motion learning. See supplementary for tracking error analysis.

Choice of 2D tracker. To demonstrate the robustness of our method with respect to the choice of 2D tracker, we employ UNINEXT [88] for 2D tracking. As shown in Table 3 (G), our method maintains stable performance.

5. Conclusion

In this paper, we introduce a novel framework for robust dynamic 3D street scene reconstruction that eliminates the reliance on 3D object trackers. Addressing the generalization limitations of 3D trackers, we propose a robust object tracking strategy based on a 2D foundation model. Our framework also features a motion learning module within an implicit feature space to handle inevitable tracking errors by autonomously refining pose inaccuracies and recovering missed detections. Experiments on the Waymo-NOTR and KITTI datasets demonstrate its adaptability and superior performance. **Limitations** and **failure cases** are discussed in the supplementary material.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant 2018AAA0102801, PCL-CMCC Foundation for Science and Innovation Grant No. 2024ZY1C0040, and the National Research Foundation, Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008).

References

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 2
- [2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 2
- [4] Mk Bashar, Samia Islam, Kashifa Kawaakib Hussain, Md Bakhtiar Hasan, ABM Rahman, and Md Hasanul Kabir. Multiple object tracking in recent times: A literature review. *arXiv preprint arXiv:2209.04796*, 2022. 3
- [5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 3
- [6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3, 6
- [7] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2, 3, 4, 5
- [8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2
- [9] Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. D2-city: a large-scale dashcam video dataset of diverse traffic scenarios. *arXiv preprint arXiv:1904.01975*, 2019. 3
- [10] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint arXiv:2311.18561*, 2023. 3
- [11] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 1, 3, 6, 8
- [12] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 3, 4
- [13] Jie Cheng, Yingbing Chen, Qingwen Zhang, Lu Gan, Chengju Liu, and Ming Liu. Real-time trajectory planning for autonomous driving with gaussian process and incremental refinement. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8999–9005. IEEE, 2022. 1
- [14] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8679–8689, 2023. 1
- [15] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 3
- [16] Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Conference on Robot Learning*, pages 1268–1281. PMLR, 2023. 1
- [17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3
- [18] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 3
- [19] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmotformer: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9784–9794, 2023. 3
- [20] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3
- [21] George Eskandar. An empirical study of the generalization ability of lidar 3d object detectors to unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23815–23825, 2024. 1, 3
- [22] Tobias Fischer, Jonas Kulhanek, Samuel Rota Buló, Lorenzo Porzi, Marc Pollefeys, and Peter Kotschieder.

Dynamic 3d gaussian fields for urban areas. *arXiv preprint arXiv:2406.03175*, 2024. 1

- [23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 3, 6
- [24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6
- [25] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 1
- [26] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015. 3
- [27] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 3
- [28] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2
- [29] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 1
- [30] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *2022 International Conference on 3D Vision (3DV)*, pages 197–208. IEEE, 2022. 3
- [31] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1
- [32] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An, Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 2, 3, 5, 6, 7
- [33] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, 2022. 1
- [34] Ziqin Huang, Gu Wang, Chenyangguang Zhang, Ruida Zhang, Xiu Li, and Xiangyang Ji. Givepose: Gradual intra-class variation elimination for rgb-based category-level object pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22055–22066, 2025. 3
- [35] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 1
- [36] Lesole Kalake, Wanggen Wan, and Li Hou. Analysis based on recent deep learning approaches applied in real-time multi-object tracking: a review. *IEEE Access*, 9:32650–32671, 2021. 3
- [37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 4, 6, 8
- [38] Mustafa Khan, Hamidreza Fazlali, Dhruv Sharma, Tongtong Cao, Dongfeng Bai, Yuan Ren, and Bingbing Liu. Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction. *arXiv preprint arXiv:2407.02598*, 2024. 1, 3, 6
- [39] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3
- [41] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 2, 3
- [42] Jinyu Li, Chenxu Luo, and Xiaodong Yang. Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17567–17576, 2023. 1, 3
- [43] Wei Li, CW Pan, Rong Zhang, JP Ren, YX Ma, Jin Fang, FL Yan, QC Geng, XY Huang, HJ Gong, et al. Aads: Augmented autonomous driving simulation using data-driven algorithms. *Science robotics*, 4(28):eaaw0863, 2019. 3
- [44] Yanyan Li, Chenyu Lyu, Yan Di, Guangyao Zhai, Gim Hee Lee, and Federico Tombari. Geogaussian: Geometry-aware gaussian splatting for scene rendering. In *European Conference on Computer Vision*, pages 441–457. Springer, 2024. 2
- [45] Zhong Li, Yu Ji, Wei Yang, Jinwei Ye, and Jingyi Yu. Robust 3d human motion reconstruction via dynamic template construction. In *2017 International Conference on 3D Vision (3DV)*, pages 496–505. IEEE, 2017. 3
- [46] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of*

the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14864–14873, 2024. 1

- [47] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020. 1
- [48] Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. Gaufré: Gaussian deformation fields for real-time dynamic novel view synthesis. *arXiv preprint arXiv:2312.11458*, 2023. 3
- [49] 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. 2, 3
- [50] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2
- [51] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. CATRE: iterative point clouds alignment for category-level object pose refinement. In *European Conference on Computer Vision*, 2022. 3
- [52] Xingyu Liu, Chenyangguang Zhang, Gu Wang, Ruida Zhang, and Xiangyang Ji. Rasim: A range-aware high-fidelity rgb-d data simulation pipeline for real-world applications. In *2024 IEEE international conference on robotics and automation (ICRA)*, pages 17057–17064. IEEE, 2024. 3
- [53] Xingyu Liu, Yingyue Li, Chengxi Li, Gu Wang, Chenyangguang Zhang, Ziqin Huang, and Xiangyang Ji. Gfreedet: Exploiting gaussian splatting and foundation models for model-free unseen object detection in the bop challenge 2024. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2025. 3
- [54] Xingyu Liu, Gu Wang, Ruida Zhang, Chenyangguang Zhang, Federico Tombari, and Xiangyang Ji. UNOPose: Unseen object pose estimation with an unposed rgb-d reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3
- [55] Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Gu Wang, Jiwen Tang, Zhigang Li, and Xiangyang Ji. Gdrnp: A geometry-guided and fully learning-based object pose estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3
- [56] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 3
- [57] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 3
- [58] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial intelligence*, 293:103448, 2021. 3
- [59] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 3
- [60] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8):1909–1963, 2023. 3
- [61] 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. 2, 3
- [62] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv 2022. arXiv preprint arXiv:2205.06230*, 2, 2022. 2
- [63] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2
- [64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3
- [65] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 1, 3, 6, 8
- [66] Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3d multi-object tracking. In *European Conference on Computer Vision*, pages 680–696. Springer, 2022. 3, 6, 8
- [67] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6782–6791, 2025. 3
- [68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3
- [69] Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Domain generalization of 3d semantic segmentation in autonomous driving. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 18077–18087, 2023. [3](#)
- [70] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. [5](#)
- [71] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018. [3](#)
- [72] Yeji Song, Chaerin Kong, Seoyoung Lee, Nojun Kwak, and Joonseok Lee. Towards efficient neural scene graphs by learning consistency fields. *arXiv preprint arXiv:2210.04127*, 2022. [3](#)
- [73] Louis Soum-Fontez, Jean-Emmanuel Deschaud, and François Goulette. Mdt3d: Multi-dataset training for lidar 3d object detection generalization. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5765–5772. IEEE, 2023. [1, 3](#)
- [74] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. *arXiv preprint arXiv:2406.18717*, 2024. [3, 4, 5](#)
- [75] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 246–264. Springer, 2020. [3](#)
- [76] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. [2, 3, 6](#)
- [77] Shanlin Sun, Bingbing Zhuang, Ziyu Jiang, Buyu Liu, Xiaohui Xie, and Manmohan Chandraker. Lidarf: Delving into lidar for neural radiance field on street scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19563–19572, 2024. [3](#)
- [78] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12375–12385, 2023. [3](#)
- [79] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020. [2](#)
- [80] Yu Wang, Shaohua Wang, Yicheng Li, and Mingchun Liu. A comprehensive review of 3d object detection in autonomous driving: Technological advances and future directions. *arXiv preprint arXiv:2408.16530*, 2024. [3](#)
- [81] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. [2, 3, 4, 5, 6, 8](#)
- [82] Hai Wu, Wenkai Han, Chenglu Wen, Xin Li, and Cheng Wang. 3d multi-object tracking in point clouds based on prediction confidence-guided data association. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5668–5677, 2021. [2](#)
- [83] Hai Wu, Jinhao Deng, Chenglu Wen, Xin Li, Cheng Wang, and Jonathan Li. Casa: A cascade attention network for 3-d object detection from lidar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. [2](#)
- [84] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795, 2024. [1, 2, 3, 4, 6, 8](#)
- [85] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, pages 3–15. Springer, 2023. [1, 3, 6, 8](#)
- [86] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021. [2](#)
- [87] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023. [3](#)
- [88] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. [1, 8](#)
- [89] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *ECCV*, 2024. [1, 2, 3, 6, 7, 8](#)
- [90] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. [3, 6, 7](#)
- [91] Ze Yang, Yun Chen, Jingkan Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. [3](#)
- [92] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and

- rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 3
- [93] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 3
- [94] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 1, 3
- [95] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, Trevor Darrell, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5): 6, 2018. 2, 3
- [96] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 1
- [97] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9253–9262, 2023. 1, 3
- [98] Peng Zhang, Xin Li, Liang He, and Xin Lin. 3d multiple object tracking on autonomous driving: A literature review. *arXiv preprint arXiv:2309.15411*, 2023. 3
- [99] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European conference on computer vision*, pages 655–672. Springer, 2022. 3
- [100] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7452–7459. IEEE, 2022.
- [101] Ruida Zhang, Ziqin Huang, Gu Wang, Chenyangguang Zhang, Yan Di, Xingxing Zuo, Jiwen Tang, and Xiangyang Ji. Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation. In *European Conference on Computer Vision*, pages 467–484. Springer, 2024. 3
- [102] Ruida Zhang, Chenyangguang Zhang, Yan Di, Fabian Manhardt, Xingyu Liu, Federico Tombari, and Xiangyang Ji. Kp-red: Exploiting semantic keypoints for joint 3d shape retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20540–20550, 2024. 3
- [103] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387*, 2022. 2
- [104] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129:3069–3087, 2021. 3
- [105] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21336–21345, 2024. 1
- [106] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024. 1, 2, 3