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

DIAL-GS: Dynamic Instance Aware Reconstruction for Label-free Street Scenes with 4D Gaussian Splatting

Chenpeng Su[1*], Wenhua Wu[1*], Chensheng Peng[2], Tianchen Deng[1], Zhe Liu[1], Hesheng Wang[1†]

***Abstract*— Urban scene reconstruction is critical for autonomous driving, enabling structured 3D representations for data synthesis and closed-loop testing. Supervised approaches rely on costly human annotations and lack scalability, while current self-supervised methods often confuse static and dynamic elements and fail to distinguish individual dynamic objects, limiting fine-grained editing. We propose DIAL-GS, a novel dynamic instance-aware reconstruction method for label-free street scenes with 4D Gaussian Splatting. We first accurately identify dynamic instances by exploiting appearance–position inconsistency between warped rendering and actual observations. Guided by instance-level dynamic perception, we employ instance-aware 4D Gaussians as the unified volumetric representation, realizing dynamic-adaptive and instance-aware reconstruction. Furthermore, we introduce a reciprocal mechanism through which identity and dynamics reinforce each other, enhancing both integrity and consistency. Experiments on urban driving scenarios show that DIAL-GS surpasses existing self-supervised baselines in reconstruction quality and instance-level editing, offering a concise yet powerful solution for urban scene modeling. Our code and models are available at: `https://github.com/IRMVLab/DIAL-GS`.**

I. INTRODUCTION

Urban scene reconstruction has become a cornerstone technology in autonomous driving. By generating structured 3D representations of complex urban environments, it provides foundations for large-scale data synthesis, supporting both algorithm development and closed-loop testing in safety-critical scenarios [1].

To address the challenges of road scenes, many existing methods adopt supervised learning, which relies on labor-intensive manual annotations to accurately capture the spatial and semantic information of dynamic objects [2]–[10]. However, manual labeling is costly, and supervised models are inherently limited to the scope of annotated datasets, hindering their scalability. These drawbacks have motivated increasing interest in self-supervised reconstruction [11]–[14].

Without explicit supervisory signals, self-supervised methods are prone to dynamic-static confusion: static objects may be incorrectly modeled as dynamic due to data noise, while slowly moving objects may be mistakenly treated as static. To address this issue, DIAL-GS introduces an inconsistency-driven approach for precise instance-level dynamic perception. Specifically, when dynamic objects are forced to be represented by static Gaussians, the resulting static field merely records their instantaneous states in the past, which inevitably lag behind the current observations. This discrepancy manifests as the inconsistency between rendering and ground truth in both appearance and position, which DIAL-GS leverages as a reliable cue to distinguish dynamic instances from the static background.

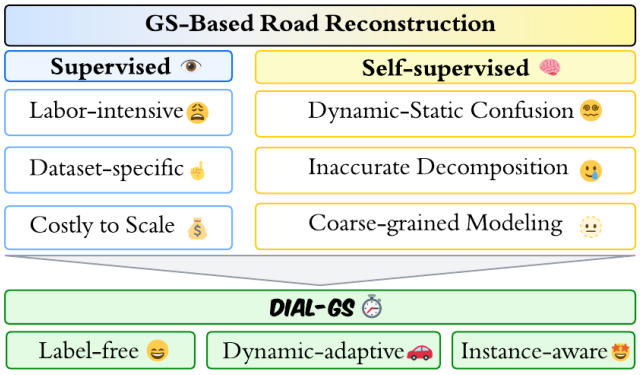

Fig. 1. Motivation. DIAL-GS overcomes the limits of supervised and self-supervised methods with label-free, dynamic-adaptive and instance-aware reconstruction.

Another challenge in self-supervised reconstruction lies in constructing a unified representation for the entire scene. Supervised paradigms often sidestep this difficulty by applying static Gaussians [15] to static background and employing time-varying Gaussians for dynamic objects [1], [2], [4]. In contrast, self-supervised frameworks cannot pre-identify elements as static or dynamic. This necessitates a representation that is simultaneously capable of preserving invariant spatial attributes of static background while modeling the spatiotemporal variations of dynamic objects.

Beyond unified representation, a further difficulty of self-supervised reconstruction lies in enabling scene editing, which is fundamental for data generation. Unfortunately, current self-supervised approaches lack instance awareness, reducing the edition to coarse static–dynamic decomposition. Without the ability to distinguish between individual dynamic objects, these methods cannot support per-instance modeling or fine-grained editing, which severely limits their applicability.

DIAL-GS adopts instance-aware 4DGS to handle these limitations. By jointly encoding identity and dynamic attributes, DIAL-GS provides a unified framework where both static and dynamic scene components are consistently modeled and each Gaussian primitive is enriched with ID features. Furthermore, we propose a reciprocal ID–dynamics training strategy. We

This work was supported by National Key R&D Program of China (Grant No.2024YFB4708900). It was also supported in part by the Natural Science Foundation of China under Grant 625B1026, 62225309, U24A20278, 62361166632.

[1] Shanghai Jiao Tong University, Shanghai 200240, China.
[2] University of California, Berkeley, Berkeley, CA 94720, USA.
* Equal contribution. † Corresponding author.

identify Gaussians belonging to the same instance via ID embeddings and enforce their dynamics consistency, while dynamic attributes are leveraged to select existing Gaussians and cluster their ID embeddings. In this way, the integrity of instance awareness and the consistency of dynamic modeling are jointly enhanced.

With these mechanisms, DIAL-GS realizes a dynamic-adaptive and instance-aware 4D reconstruction within the self-supervised regime. Our main contributions are summarized as follows:

1) We introduce an intuitive and accurate instance-level dynamic perception algorithm by exploiting the appearance and position inconsistency caused by motion.
2) We empower self-supervised reconstruction with instance awareness by proposing instance-aware 4DGS, and introduce a reciprocal mechanism in which instance awareness and dynamics mutually benefit.
3) Extensive experiments demonstrate that DIAL-GS surpasses prior methods in image reconstruction and novel view synthesis, while uniquely enabling instance-level editing – a capability absent from existing self-supervised approaches.

II. RELATED WORK

A. NeRF-based Driving Reconstruction

Numerous methods have investigated NeRF-based approaches for road-scene reconstruction. NSG [10] models dynamic multi-object scenes with a scene graph, enabling instance-level view synthesis and 3D detection. Block-NeRF [16] adopts block-wise representations with semantic masking and appearance codes to reconstruct large-scale scenes, while READ [17] introduces a real-time rendering engine with ω-net for photorealistic scene synthesis and editing. SUDS [18] factorizes scenes into static, dynamic, and far-field radiance fields with hash-grid acceleration, supporting scene flow estimation and semantic manipulation. EmerNeRF [19] advances self-supervised modeling by estimating flow-based correspondences and leveraging 2D foundation features for geometry and semantics.

Despite these advances, NeRF-based methods remain constrained by heavy computation, slow training and rendering, and limited scalability to large dynamic environments. Their reliance on dense sampling and implicit volumetric MLPs further hinders real-time applications and fine-grained editing [20]. By contrast, Gaussian Splatting achieves comparable visual quality with significantly faster performance and naturally accommodates instance-aware extensions.

B. GS-based Driving Reconstruction

3D Gaussian Splatting (3DGS) [15] has recently emerged as an efficient alternative to NeRFs, replacing implicit MLP-based volumetric fields with explicit Gaussian primitives. By rasterizing Gaussians, 3DGS enables fast rendering and, thanks to its explicit structure, naturally extends beyond view synthesis to tasks such as dynamic scene reconstruction, geometry editing, and physical simulation [21]. These properties make 3DGS particularly well-suited for large-scale driving-scene reconstruction, where efficiency is essential.

Supervised methods rely on labor-annotated datasets to guide geometry, semantics, and dynamics, achieving highly accurate and structured reconstructions. DrivingGaussian [3] combines incremental static reconstruction with a dynamic Gaussian graph for large-scale driving scenes. Street Gaussians [2] leverages a 4D spherical harmonics appearance model, tracked pose optimization, and point cloud initialization to improve rendering quality. AutoSplat [5] introduces geometry-constrained background modeling, template-based foreground initialization, and temporally adaptive appearance modeling. OmniRe [4] builds a holistic scene graph that unifies static backgrounds, vehicles, SMPL-modeled humans [22], and other non-rigid actors. While effective, supervised approaches are labor-intensive, expensive to scale, and inherently constrained by the distribution of annotated datasets, limiting their generalization to unseen scenarios.

Self-supervised approaches remove the need for annotations by exploiting temporal consistency, geometric cues, and multi-view signals in driving data. S^3 Gaussians [23] decomposes scenes with a spatio-temporal network that models Gaussian deformation through feature planes. PVG [12] introduces 4D Gaussians with periodic vibration and learnable lifespans, avoiding Hexplane-based deformation [24]. DeSiRe-GS [11] enhances separation with a motion-mask extraction mechanism, 3D regularization, and temporal cross-view consistency. Despite progress, these methods face two major limitations: (i) dynamic–static confusion, where noise, pose perturbations, or slow motion lead to misclassification of static versus dynamic objects, and (ii) lack of instance awareness, as they only coarsely separate static and dynamic components without distinguishing or editing individual objects.

Our work directly addresses both issues by enabling accurate instance-level dynamic perception and introducing instance-aware 4D Gaussians.

C. Semantic Scene Modeling with Gaussians

Previous semantic scene modeling approaches are usually NeRF-based [25]–[29]. However, with the rise of 3D Gaussian Splatting (GS), increasing efforts focus on injecting 2D semantic knowledge into 3D-GS. Gaussian Grouping [30] transfers SAM's segmentation capability [31] [32] into 3D, enabling zero-shot segmentation without 3D mask annotations. To resolve multi-granularity ambiguity, SAGA [33] introduces a promptable 3D segmentation framework with a scale-gated mechanism and contrastive distillation. Semantic Gaussians [34] further projects diverse pre-trained 2D features into 3D Gaussians and enriches them with semantic attributes.

Building on these approaches, a natural direction for self-supervised road-scene reconstruction is to achieve instance awareness by embedding trajectory-tracked instance IDs into dynamic Gaussians. However, existing works are mostly tailored for static or small-scale scenes, limiting their applicability to highly dynamic driving environments. To bridge this gap, DIAL-GS integrates ID-embedding with dynamic

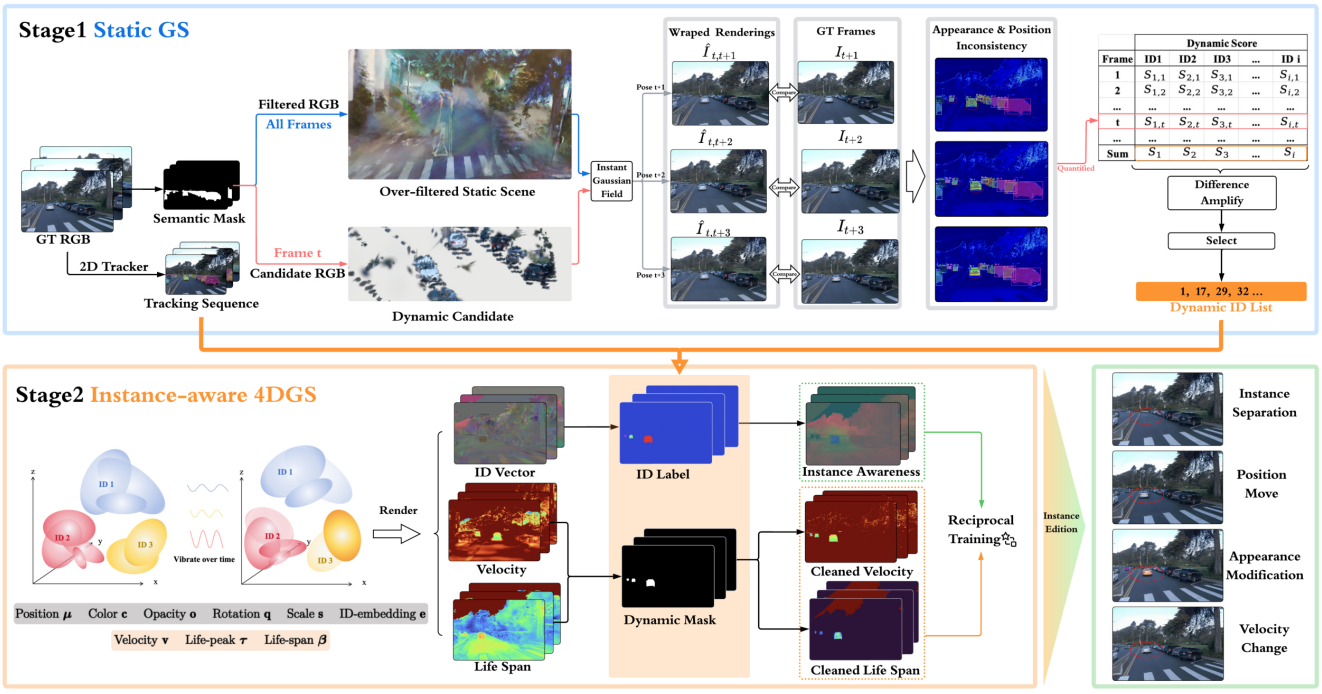

Fig. 2. Method overview. (i) Stage 1 conducts instance-level dynamic perception with static GS by exploiting inconsistency between warped renderings and ground-truth frames. Accumulated dynamic scores quantify inconsistency are used to obtain a dynamic ID list, according to which ID labels and dynamic masks are derived. (ii) Stage 2 reconstructs the scene with instance-aware 4DGS as the unified representation. Guided by the ID labels and dynamic masks, it achieves instance awareness and refines dynamic attributes. Then it performs reciprocal training to enhance both instance awareness integrity and dynamics consistency. (iii) With instance awareness, DIAL-GS further enables instance-level editing, a capability not supported by previous self-supervised approaches.

attributes and introduces a reciprocal training strategy that allow both to enhance each other.

III. METHOD

We address the problem of road-scene reconstruction under a self-supervised setting. Given a sequence of temporally aligned multi-view observations—including RGB images, LiDAR point clouds, camera intrinsics, and ego-poses—our objective is to recover a scene representation that captures both static structures and dynamic objects without any 3D annotation.

As shown in Fig. 2, we design a two-stage pipeline. In Stage 1, we first reconstruct an over-filtered static scene with all frames and build dynamic candidate Gaussians for each frame. By combining them, we establish an instantaneous Gaussian field, where dynamic objects lag behind ground-truth observations and exhibit inconsistency in both appearance and position when rendered from new views. We quantify the inconsistency as a dynamic score for each instance and select dynamic IDs based on the aggregated scores. With the dynamic ID list, we then derive ID labels and dynamic masks from tracking sequence. In Stage 2, the scene is subsequently reconstructed with instance-aware 4D Gaussians as a unified representation. Guided by the labels and masks from Stage 1, we realize instance awareness and clean dynamic attributes, and further propose the reciprocal identity-dynamics training to reinforce the integrity of ID-embedding and consistency of dynamic attributes. With these designs, DIAL-GS enables self-supervised reconstruction to support instance-level editing.

A. Instance-level Dynamic Perception

When represented with static GS, only instantaneous states of dynamic objects can be captured. As time progresses, the previously recorded state of a dynamic object inevitably diverges from the actual observation, much like how a frozen image of a moving object always differs from its true current state. Building on this intuition, we propose Instance-level Dynamic Perception driven by inconsistency.

Instantaneous Gaussian Field Establishment. We start by extracting the tracking sequence and semantic masks using BoT-SORT [35]. The semantic masks divide each frame into a filtered region (mainly static content, though some static objects may be excluded) and a candidate region (where potential dynamic objects remain). Using all filtered RGB frames together with all filtered LiDAR point clouds, we reconstruct an over-filtered static scene S_{over} with static GS, which provides a temporally consistent background.

At time t, the point cloud is back-projected to recover candidates' spatial structure and, with RGB supervision, dynamic candidate Gaussians C_t is obtained. Combined with the static field S_{over}, they form the instantaneous Gaussian field, which is then warped to frames $t+1, \ldots, t+k$ to simulate new-view observations:

$$\hat{I}_{t,t+k} = \mathcal{F}_{t \to t+k}(C_t, S_{over}), \tag{1}$$

where $\mathcal{F}$ stands for the warp process.

Dynamic Score and Instance Selection. To measure the inconsistency of each candidate, we perform instance segmentation on $\hat{I}_{t,t+k}$ and establish ID correspondences with the ground-truth frame I_{t+k}. The appearance inconsistency of instance i is defined as:

$$\mathcal{I}_{i,t}^{app} = \left\| I_{t+k} - \hat{I}_{t,t+k} \right\| \odot \frac{(\mathcal{M}_{i,t} \mid \hat{\mathcal{M}}_{i,t+k})}{\left\| (\mathcal{M}_{i,t} \mid \hat{\mathcal{M}}_{i,t+k}) \right\|}, \quad (2)$$

where $\mathcal{M}_{i,t}$ denotes the mask of instance i at frame t, and $(\cdot \mid \cdot)$ represents the union of two masks.

Meanwhile, the position inconsistency is defined as

$$\mathcal{I}_{i,t}^{pos} = \frac{|c_{i,t} - \hat{c}_{i,t+k}|}{\sqrt{A_{i,t}}} + \frac{|A_{i,t} - \hat{A}_{i,t+k}|}{A_{i,t}} + \frac{\sigma(E_{i,t,t+k})}{\sqrt{A_{i,t}}}, \quad (3)$$

where c denotes the bounding box center, A its area, $E_{i,t,t+k}$ the edge difference, and σ the standard deviation.

By combining these two inconsistency measures, we define the dynamic score of instance i at frame t as $S_{i,t} = \mathcal{I}_{i,t}^{app} + \mathcal{I}_{i,t}^{pos}$. We compute $S_{i,t}$ from the first frame and accumulate them to obtain the final score: $S_i = \frac{1}{n}\sum_t S_{i,t}$, where n denotes the total number of frames in which instance i appears. To enhance the separation between dynamic and static instances, we apply a cubic amplification and classify an instance as dynamic if $S_i^3 > \delta$. Finally, a dynamic ID list $\mathcal{D} = \{i | S_i^3 > \delta\}$ is obtained in stage one.

B. Self-supervised Reconstruction with instance-aware 4DGS

While existing self-supervised reconstruction methods have shown promise in capturing geometry and motion, they largely overlook instance awareness. Without distinguishing which Gaussian belongs to which object, these approaches often produce entangled representations that hinder reliable decomposition and fine-grained editing. This limitation motivates us to introduce instance-aware 4DGS as the unified representation.

Instance-aware 4D Gaussian. To consistently model both static and dynamic scene components and realize instance awareness, we embed ID vectors to PVG [12]. Thereby, the Gaussian of stage two is formulated with position $\boldsymbol{\mu}$, color $\mathbf{c}$, opacity o, rotation $\boldsymbol{q}$, scale $\boldsymbol{s}$, ID-embedding $\boldsymbol{e}$ as static attributes and velocity $\boldsymbol{v}$, life-peak τ, life-span β as dynamic attributes. The Gaussian vibrates around $\boldsymbol{\mu}$ and fades away according to τ and β:

$$\tilde{\boldsymbol{\mu}}(t) = \boldsymbol{\mu} + \frac{l}{2\pi} \cdot \sin\left(\frac{2\pi(t-\tau)}{l}\right) \cdot \boldsymbol{v}, \quad (4)$$

$$\tilde{o}(t) = o \cdot exp(-\frac{1}{2}(t-\tau)^2\beta^{-2}). \quad (5)$$

Identity Loss. We formulate the ID-embedding $\boldsymbol{e}$ as a static 8-bit vector inspired by [30] and render it with the original splatting pipeline.

$$\mathbf{E} = \sum_{i \in \mathcal{N}} \boldsymbol{e}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \quad (6)$$

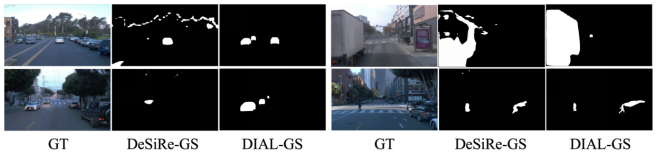

Fig. 3. Dynamic Mask Comparison. DeSiRe-GS misclassifies static region and incompletely captures dynamic parts. DIAL-GS obtains accurate and sharp dynamic masks instead.

where $\mathcal{N}$ is the total number of depth-sorted Gaussians, and $\boldsymbol{e}_i$ and α_i denote the ID embedding and density of the i-th Gaussian, respectively.

We then take a simple MLP l followed by a softmax as the classifier: $\hat{I}^{id} = argmax(softmax(l(\mathbf{E})))$, where $\hat{I}^{id}$ stands for ID rendering. By ignoring instances not in $\mathcal{D}$, ID label I^{id} can be derived from tracking sequence. The identity loss is then defined as the pixel-wise cross-entropy between the predicted ID renderings and the ID labels:

$$\mathcal{L}_{\text{id}} = -\frac{1}{P}\sum_{p=1}^{P}\sum_{c=1}^{C} I_{p,c}^{id} \log\left(\hat{I}_{p,c}^{id}\right), \quad (7)$$

where $P = h \cdot w$ is pixel number of the rendering, C is the max number of dynamic IDs, and $\hat{I}_{p,c}^{id} \in [0,1]$ is the predicted probability that pixel p belongs to class c and $I_{p,c}^{id} \in \{0,1\}$ is the one-hot ground-truth indicator.

Dynamic Attributes Regularization. PVG [12] tends to assign dynamics to static region inconsistent across time due to data noise, and fails to capture the dynamics of slowly moving objects. DeSiRe-GS [11] attempts to alleviate this issue by introducing a motion mask, yet it still inherits similar confusion since the mask itself is also learned in a fully self-supervised manner. In contrast, DIAL-GS leverages 2D trackers to directly generate instance mask and derive dynamic mask $\mathcal{M}$ by selecting IDs in $\mathcal{D}$. We then regulate dynamics with precise and instance-level mask:

$$\mathcal{L}_{\bar{v}} = \frac{1}{|\bar{\mathcal{M}}|} I^{\bar{v}} \odot \bar{\mathcal{M}}, \quad (8)$$

$$\mathcal{L}_{\beta} = -\frac{1}{|\bar{\mathcal{M}}|} I^{\beta} \odot \bar{\mathcal{M}}, \quad (9)$$

where $\bar{\boldsymbol{v}} = \boldsymbol{v} \cdot exp(-\frac{\beta}{2l})$ represents the instant velocity, and $I^{\bar{v}}$, I^{β} represent the rendering of $\bar{\boldsymbol{v}}$ and β respectively.

C. Reciprocal Identity-Dynamics Training

Another limitation of existing self-supervised methods is the inconsistency of dynamic attributes among Gaussians belonging to the same object. For example, different Gaussians of a single car may exhibit significantly different instant velocities or life-spans. Moreover, relying solely on $\mathcal{L}_{id}$ often results in incomplete instance awareness, as it only provides 2D supervision. To address this, DIAL-GS introduces a reciprocal training scheme, enabling both more complete ID-embedding and more coherent dynamic attributes.

3D Identity Loss. Due to the mechanism of PVG [12], Gaussians fade out by decreasing opacity while retaining

TABLE I

COMPARISON OF METHODS ON THE WAYMO OPEN DATASET AND KITTI DATASET.

Method	Waymo Open Dataset						KITTI					
	Image reconstruction			Novel view synthesis			Image reconstruction			Novel view synthesis		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
S-NeRF [36]	19.67	0.528	0.387	19.22	0.515	0.400	19.23	0.664	0.193	18.71	0.606	0.352
StreetSurf [37]	26.70	0.846	0.3717	23.78	0.822	0.401	24.14	0.819	0.257	22.48	0.763	0.304
3DGS [15]	27.99	0.866	0.293	25.08	0.822	0.319	21.02	0.811	0.202	19.54	0.776	0.224
NSG [10]	24.08	0.656	0.441	21.01	0.571	0.487	0.032	0.683	0.189	17.78	0.645	0.312
Mars [38]	21.81	0.681	0.430	20.69	0.636	0.453	27.96	0.900	0.185	24.31	0.845	0.160
SUDS [18]	28.83	0.805	0.317	25.36	0.783	0.384	28.83	0.917	0.147	26.07	0.797	0.131
EmerNeRF [19]	28.11	0.786	0.373	25.92	0.763	0.384	26.95	0.828	0.218	25.24	0.801	0.237
PVG [12]	32.46	0.910	0.229	28.11	0.849	0.279	32.83	0.937	0.070	27.43	0.896	0.114
DeSiRe-GS [11]	33.61	0.919	0.204	29.75	0.878	0.213	33.94	0.949	0.040	**28.87**	0.901	0.106
Ours	**36.88**	**0.948**	**0.113**	**30.14**	**0.880**	**0.183**	**34.23**	**0.954**	**0.039**	28.34	**0.911**	**0.084**

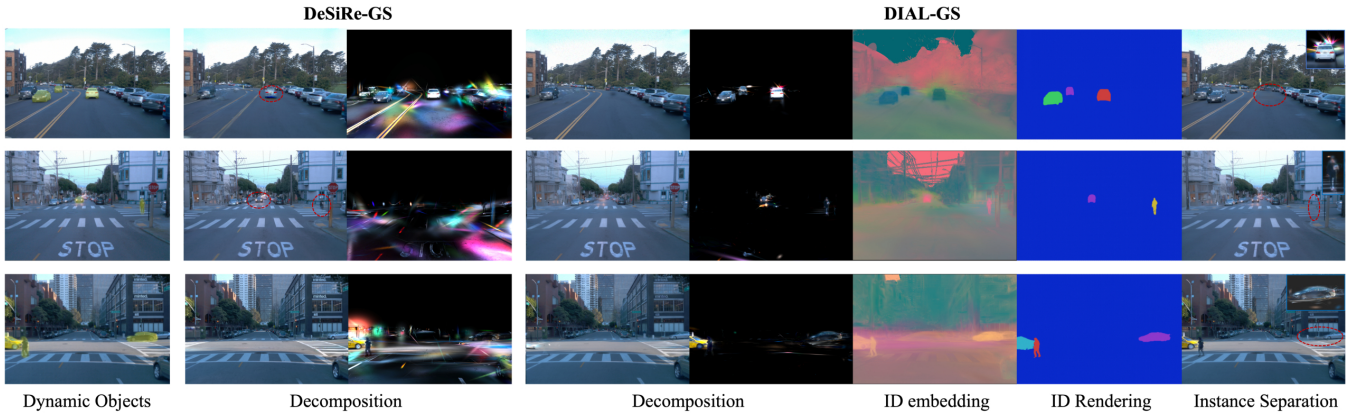

Fig. 4. Qualitative results. Decomposition with DeSiRe-GS [11] suffers from severe misclassification, whereas DIAL-GS achieves accurate decomposition and clear instance separation.

their static attributes such as position. Therefore, clustering all ID embeddings would incur prohibitive computation, so we leverage dynamic attributes to filter Gaussians that actually exist in the current frame and perform ID clustering only on them.

We select existing Gaussians of frame t as $\mathcal{G}_t^{exist} = \{j|\tilde{o}(t) > \epsilon\}$. For each Gaussian $j \in \mathcal{G}_t^{exist}$, we extract its predicted distribution as $P_j = softmax(l(\mathbf{e}_j))$ and predicted ID as $ID_j = argmax(P_j)$. We collect dynamic Gaussians as $\mathcal{G}_t^{dyn} = \{j|ID_j \in \mathcal{D}, j \in \mathcal{G}_t^{exist}\}$, and static Gaussians as: $\mathcal{G}_t^{static} = \mathcal{G}_t^{exist} - \mathcal{G}_t^{dyn}$. For each Gaussian in $\mathcal{G}_t^{dyn}$, we then search for its K nearest neighbors in $\mathcal{G}_t^{static}$ and obtain their predicted distributions Q_k. Finally, we adopt the KL divergence as the loss function to encourage nearby static Gaussians to align with the ID-embedding of dynamic Gaussians:

$$\mathcal{L}_{3d} = \frac{1}{K \cdot |\mathcal{G}_t^{dyn}|} \sum_{j=1}^{|\mathcal{G}_t^{dyn}|} \sum_{k=1}^{K} D_{KL}(P_j||Q_k). \tag{10}$$

Note that $\mathcal{L}_{3d}$ is activated in the later stage of training, when both the dynamic attributes and the ID-embedding have nearly converged. Such a scheduling strategy maximizes the reliability of $\mathcal{G}_t^{exist}$ while exploiting the stable ID-embedding optimized by $\mathcal{L}_{id}$. By enforcing 3D clustering, instance awareness is no longer restricted to 2D alignment but instead achieves a holistic embedding in the 3D space.

Dynamic Consistency Loss. After the ID-embedding has stabilized through $\mathcal{L}_{id}$ and $\mathcal{L}_{3d}$, we leverage reliable instance awareness to enforce consistency of dynamic attributes. Similar to the computation of $\mathcal{L}_{3d}$, we first extract $\mathcal{G}_t^{exist}$ and their predicted IDs. For each instance in $\mathcal{D}$, we gather its Gaussians by $\mathcal{G}_t^i = \{j \mid ID_j = i, i \in \mathcal{D}\}$. For every Gaussian in $\mathcal{G}_t^i$, we then search K nearest neighbors within $\mathcal{G}_t^i$ and obtain their instant velocities and life-spans. The consistency losses are defined as follows:

$$\mathcal{L}_{\text{mag}} = \frac{1}{K \cdot |\mathcal{G}_t^i|} \sum_{j=1}^{|\mathcal{G}_t^i|} \sum_{k=1}^{K} \frac{\|\bar{\boldsymbol{v}}_j - \bar{\boldsymbol{v}}_k\|_2}{\|\bar{\boldsymbol{v}}_{mean}\|_2}, \tag{11}$$

$$\mathcal{L}_{\text{dir}} = \frac{1}{2K \cdot |\mathcal{G}_t^i|} \sum_{j=1}^{|\mathcal{G}_t^i|} \sum_{k=1}^{K} \left(1 - \frac{\bar{\boldsymbol{v}}_j \cdot \bar{\boldsymbol{v}}_k}{\|\bar{\boldsymbol{v}}_j\|_2 \cdot \|\bar{\boldsymbol{v}}_k\|_2}\right), \tag{12}$$

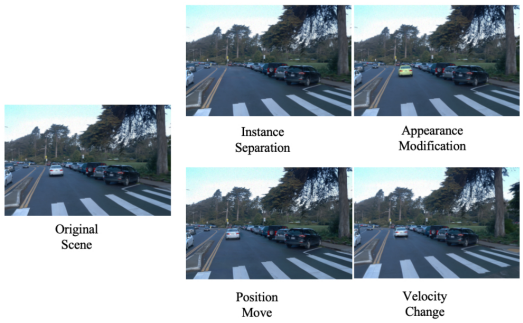

Fig. 5. Instance Edition. By realizing instance awareness, DIAL-GS supports instance edition within the self-supervised regime.

$$\mathcal{L}_{\text{beta}} = \frac{1}{K \cdot |\mathcal{G}_t^i|} \sum_{j=1}^{|\mathcal{G}_t^i|} \sum_{k=1}^{K} \frac{|\beta_j - \beta_k|}{\bar{\beta}}, \tag{13}$$

$$\mathcal{L}_{\text{consist}} = \lambda_{\text{mag}} \cdot \mathcal{L}_{\text{mag}} + \lambda_{\text{dir}} \cdot \mathcal{L}_{\text{dir}} + (1 - \lambda_{\text{mag}} - \lambda_{\text{dir}}) \cdot \mathcal{L}_{\text{beta}}. \tag{14}$$

The reinforcement of consistency encourages the 4DGS of the same dynamic object to form a coherent representation, thereby reducing artifacts in novel view synthesis (NVS) and allowing dynamic attributes to serve as reliable auxiliary cues besides ID-embedding for decomposition.

D. Optimization

In the first stage, all Gaussians are treated as static and the training loss consists:

$$\mathcal{L}_I = (1 - \lambda_{ssim}) \|I - \tilde{I}\| + \lambda_{ssim} \text{SSIM}(I, \tilde{I}), \tag{15}$$

$$\mathcal{L}_D = \|I^D - D_{gt}\|, \tag{16}$$

$$\mathcal{L}_o = -\frac{1}{hw} \sum O \cdot \log O - \frac{1}{hw} \sum \mathcal{M}_{sky} \cdot \log(1 - O). \tag{17}$$

The whole loss function of stage one is:

$$\mathcal{L}_{\text{stage1}} = \lambda_I \mathcal{L}_I + \lambda_D \mathcal{L}_D + \lambda_o \mathcal{L}_O. \tag{18}$$

In the second stage, we gradually introduce the proposed losses and the overall objective of stage 2 is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{stage2}} = \lambda_I \mathcal{L}_I + \lambda_D \mathcal{L}_D + \lambda_o \mathcal{L}_O + \lambda_{\bar{v}} \mathcal{L}_{\bar{v}} + \lambda_\beta \mathcal{L}_\beta + \\ \lambda_{id} \mathcal{L}_{id} + \lambda_{3d} \mathcal{L}_{3d} + \lambda_{\text{consist}} \mathcal{L}_{\text{consist}}. \end{aligned} \tag{19}$$

E. Instance-level Scene Edition

Without instance awareness, former self-supervised works can only perform coarse decomposition. On the contrary, DIAL-GS empowers self-supervised reconstruction with the ability to edit specific instances.

We jointly consider the $\mathbf{e}$, β, $\bar{\boldsymbol{v}}$ and $\tilde{o}$ to select the Gaussians belonging to instance i at frame t. By modifying $\tilde{\boldsymbol{\mu}}(t)$ and color $\mathbf{c}$, we change its position and appearance. We also use $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(t) + \Delta t \cdot \bar{\boldsymbol{v}}$ to change instance's velocity while keeping its original trajectory.

IV. Experiment

A. Experimental Setting

Dataset. We follow the experimental setup of PVG [12] and DeSiRe-GS [11], focusing on highly dynamic scenarios to enable comprehensive baseline comparisons.

Evaluation Metrics. We adopt PSNR, SSIM and LPIPS as metrics for the evaluation of image rendering quality.

Implementation Details. We trained our model on one NVIDIA A100 Tensor Core GPU. In the first stage, we use BoT-SORT [35], [39], [40] as the 2D tracker. We train S_{over} for 30,000 iterations and each C_t for 400 iterations. For efficiency and stability, we warp 2 frames to check inconsistency. We take δ as $1e-3$ to select dynamic IDs and ϵ as $5e-4$ to select existing Gaussians. During the 55,000 iterations of stage 2, we gradually introduce $\mathcal{L}_{2d}$, $\mathcal{L}_{\bar{v}}$, $\mathcal{L}_\beta$, $\mathcal{L}_{3d}$ and $\mathcal{L}_{consist}$ in sequence. We set K to 5 for KNN search. Weights mentioned are $\lambda_{ssim} = 1$, $\lambda_I = 1$, $\lambda_D = 1$, $\lambda_o = 0.05$, $\lambda_{\bar{v}} = 0.01$, $\lambda_\beta = 0.001$, $\lambda_{id} = 0.1$, $\lambda_{3d} = 1.5$, $\lambda_{\text{consist}} = 0.01$, $\lambda_{\text{mag}} = 0.4$ and $\lambda_{\text{dir}} = 0.2$.

B. Experimental Results

We report quantitative results on Waymo Open Dataset [41] and KITTI [42] in Tab. I in both image reconstruction and novel view synthesis. Ours achieves the best performance in most scenarios and demonstrates significant improvement especially in Waymo. In Fig. 4, we provide qualitative comparison against DeSiRe-GS [11]. DeSiRe-GS relies on staticness coefficient for decomposition. It can be observed that DeSiRe-GS [11] suffers severe dynamic misclassification: artifacts of dynamic objects remain in the static part, small or slow-moving dynamic objects are mistakenly treated as static, and the dynamic part includes clearly static elements such as road surfaces and parked cars. By contrast, DIAL-GS employs identity as the primary criterion, with the staticness coefficient as auxiliary support, thereby achieving precise decomposition and enabling instance-level separation — a capability that prior self-supervised methods could not realize.

C. Ablation Study

We conduct various ablation studies to verify the effectiveness of the instance-level dynamic perception and loss functions.

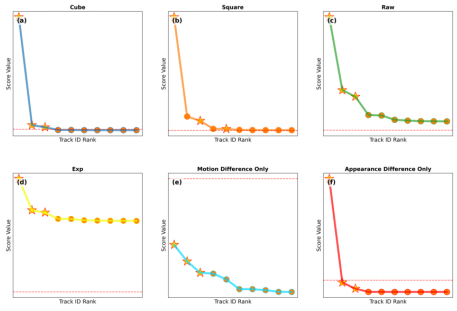

Fig. 6. Ablation studies on dynamic score. Axes denote IDs and their scores. Stars represent dynamic instances, while circles represent static ones. Red line represents the threshold.

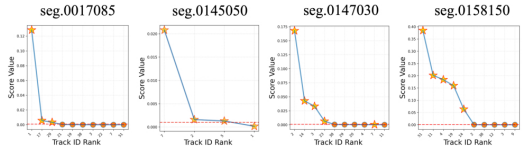

Fig. 7. Dynamic ID selection in different scenes. Axes and icons follow the same convention as in Fig. 6.

TABLE II

ABLATION STUDY WITH LOSS CONFIGURATIONS.

Exp.	$\mathcal{L}_{\bar{v}}$	$\mathcal{L}_{\beta}$	$\mathcal{L}_{id}$	$\mathcal{L}_{3d}$	$\mathcal{L}_{\text{consist}}$	PSNR↑	SSIM↑	LPIPS↓
(a)	–	✓	✓	✓	✓	29.9262	0.8792	0.1817
(b)	✓	–	✓	✓	✓	29.8243	0.8674	0.1989
(c)	✓	✓	–	–	–	29.8331	0.8788	0.1833
(d)	✓	✓	✓	–	✓	29.8405	0.8788	0.1823
(e)	✓	✓	✓	✓	–	29.9032	0.8792	**0.1813**
(f)	✓	✓	✓	✓	✓	**30.1368**	**0.8803**	0.1826

Dynamic score. In stage one, we test different methods to select dynamic IDs within the score ranking list. As shown in Fig. 6, subfigures (a)–(d) present the dynamic score ranking using S_i^3, S_i^2, S_i, and $\exp(S_i)$, respectively. Subfigures (e) and (f) further analyze S_i^3 when computed from only one type of inconsistency: $S_{i,t}^{(e)} = \mathcal{I}_{i,t}^{pos}$ and $S_{i,t}^{(f)} = \mathcal{I}_{i,t}^{app}$. The results demonstrate that stable separation between dynamic and static IDs is realized by jointly considering appearance and position inconsistency with the cubic transform. We evaluate the selection procedure across different scenes. As shown in Fig. 7, the cubic transformation suppresses static-instance scores toward zero while preserving relatively high scores for dynamic instances, enabling stable thresholding for dynamic ID separation. The few remaining misclassifications mainly arise from distant or small objects with limited observations.

Dynamic Attribute Regularization. We ablate $\mathcal{L}\bar{v}$ and $\mathcal{L}_{\beta}$ to examine the role of velocity and life-span regularization. As shown in Tab. II(a)(b), both losses contribute to improving novel-view rendering quality. Moreover, Fig. 8 illustrates that removing the regularization leads to inaccurate dynamic attributes in distant static regions and velocity grows excessively and life-span shrinks, ultimately causing overfitting to training views.

ID-embedding. The result in Tab. II(c) shows that ID-embedding not only enables instance awareness but also improves the reconstruction quality.

Reciprocal Identity–Dynamics Training. From Tab. II(d)(e),

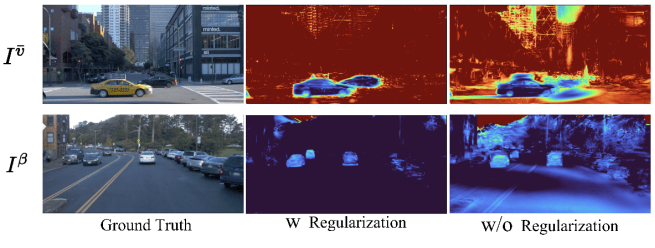

Fig. 8. Effect of Dynamic Attribute Regularization. The regularization ensures clean static regions while constraining velocity and life-span to avoid excessive growth or shrinkage.

Fig. 9. Effect of $\mathcal{L}_{3d}$. With $\mathcal{L}_{3d}$, the ID-embedding is more complete and the decomposition is more clean.

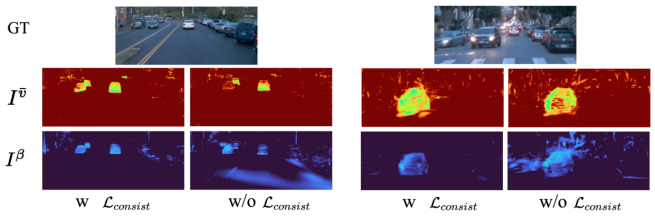

Fig. 10. Effect of $\mathcal{L}_{\text{consist}}$. The dynamic attributes are more consistent and reasonable with the guide of $\mathcal{L}_{\text{consist}}$.

we observe that the NVS quality drops when either $\mathcal{L}_{3d}$ or $\mathcal{L}_{\text{consist}}$ is removed. Fig. 9 shows that removing $\mathcal{L}_{3d}$ leads to incomplete ID embeddings, resulting in residuals along vehicle boundaries (first row) and artifacts within dynamic objects (second row). Fig. 10 shows that $\mathcal{L}_{\text{consist}}$ enhances the consistency of dynamic attributes within the same instance. As a result, the motion modeling adheres more closely to realistic physical behavior, which in turn improves both novel view synthesis and decomposition.

D. Discussion

While DIAL-GS achieves strong instance-aware reconstruction, it relies on external 2D tracker, which may introduce occasional errors. We nonetheless adopt this design because fully self-supervised dynamic perception remains unreliable for complex road scenes as shown by DeSiRe-GS [11] and discussed in Sec. III-B. Furthermore, generic self-supervised features [43] adapted by DeSiRe-GS [11] are not tailored to driving environments and lack awareness of object-motion patterns. In contrast, modern 2D trackers have already encoded rich semantic and physical priors specific to traffic scenes, producing more precise instance masks than those obtained from purely self-supervised approaches, making them a pragmatic and effective choice for our framework.

V. CONCLUSION

In this paper, we present DIAL-GS, a novel self-supervised framework with dynamic instance awareness. DIAL-GS achieves instance-level dynamic perception by leveraging inconsistency caused by motion. By proposing instance-aware 4DGS, DIAL-GS jointly encodes identity and dynamic attributes and further enables them to benefit each other in reciprocal identity–dynamics training strategy. Extensive experiments validate its effectiveness, demonstrating that DIAL-GS advances self-supervised reconstruction for real-world autonomous driving scenarios.

REFERENCES

[1] H. Zhu, Z. Zhang, J. Zhao, H. Duan, Y. Ding, X. Xiao, and J. Yuan, "Scene reconstruction techniques for autonomous driving: a review of 3d gaussian splatting," *Artificial Intelligence Review*, vol. 58, no. 1, p. 30, 2024.

[2] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians: Modeling dynamic urban scenes with gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 156–173.

[3] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 634–21 643.

[4] Z. Chen, J. Yang, J. Huang, R. de Lutio, J. M. Esturo, B. Ivanovic, O. Litany, Z. Gojcic, S. Fidler, M. Pavone, *et al.*, "Omnire: Omni urban scene reconstruction," *arXiv preprint arXiv:2408.16760*, 2024.

[5] M. Khan, H. Fazlali, D. Sharma, T. Cao, D. Bai, Y. Ren, and B. Liu, "Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction," *arXiv preprint arXiv:2407.02598*, 2024.

[6] T. Deng, S. Liu, X. Wang, Y. Liu, D. Wang, and W. Chen, "Prosgnerf: Progressive dynamic neural scene graph with frequency modulated auto-encoder in urban scenes," *arXiv preprint arXiv:2312.09076*, 2023.

[7] H. Zhou, J. Shao, L. Xu, D. Bai, W. Qiu, B. Liu, Y. Wang, A. Geiger, and Y. Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 336–21 345.

[8] S. Hwang, M.-J. Kim, T. Kang, J. Kang, and J. Choo, "Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–18.

[9] G. Hess, C. Lindström, M. Fatemi, C. Petersson, and L. Svensson, "Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 982–11 992.

[10] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2856–2865.

[11] C. Peng, C. Zhang, Y. Wang, C. Xu, Y. Xie, W. Zheng, K. Keutzer, M. Tomizuka, and W. 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*, 2025, pp. 6782–6791.

[12] Y. Chen, C. Gu, J. Jiang, X. Zhu, and L. Zhang, "Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering," *arXiv:2311.18561*, 2023.

[13] S. Sun, C. Zhao, Z. Sun, Y. V. Chen, and M. Chen, "Splatflow: Self-supervised dynamic gaussian splatting in neural motion flow field for autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 487–27 496.

[14] Y. Mao, R. Xiong, Y. Wang, and Y. Liao, "Unire: Unsupervised instance decomposition for dynamic urban scene reconstruction," *arXiv preprint arXiv:2504.00763*, 2025.

[15] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[16] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8248–8258.

[17] Z. Li, L. Li, and J. Zhu, "Read: Large-scale neural scene rendering for autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1522–1529.

[18] H. Turki, J. Y. Zhang, F. Ferroni, and D. Ramanan, "Suds: Scalable urban dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 375–12 385.

[19] J. Yang, B. Ivanovic, O. Litany, X. Weng, S. W. Kim, B. Li, T. Che, D. Xu, S. Fidler, M. Pavone, *et al.*, "Emernerf: Emergent spatial-temporal scene decomposition via self-supervision," *arXiv preprint arXiv:2311.02077*, 2023.

[20] A. Rabby and C. Zhang, "Beyondpixels: A comprehensive review of the evolution of neural radiance fields," *arXiv preprint arXiv:2306.03000*, 2023.

[21] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao, "Recent advances in 3d gaussian splatting," *Computational Visual Media*, vol. 10, no. 4, pp. 613–642, 2024.

[22] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[23] N. Huang, X. Wei, W. Zheng, P. An, M. Lu, W. Zhan, M. Tomizuka, K. Keutzer, and S. Zhang, "S3 gaussian: Self-supervised street gaussians for autonomous driving," *arXiv preprint arXiv:2405.20323*, 2024.

[24] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.

[25] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 838–15 847.

[26] X. Fu, S. Zhang, T. Chen, Y. Lu, L. Zhu, X. Zhou, A. Geiger, and Y. Liao, "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 1–11.

[27] K. Mazur, E. Sucar, and A. J. Davison, "Feature-realistic neural fusion for real-time, open set scene understanding," *arXiv preprint arXiv:2210.03043*, 2022.

[28] C. Peng, C. Xu, Y. Wang, M. Ding, H. Yang, M. Tomizuka, K. Keutzer, M. Pavone, and W. Zhan, "Q-slam: Quadric representations for monocular slam," in *Conference on Robot Learning*. PMLR, 2025, pp. 1763–1781.

[29] Y. Siddiqui, L. Porzi, S. R. Buló, N. Müller, M. Nießner, A. Dai, and P. Kontschieder, "Panoptic lifting for 3d scene understanding with neural fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9043–9052.

[30] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *European conference on computer vision*. Springer, 2024, pp. 162–179.

[31] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[32] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[33] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Segment any 3d gaussians," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1971–1979.

[34] J. Guo, X. Ma, Y. Fan, H. Liu, and Q. Li, "Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting," *arXiv preprint arXiv:2403.15624*, 2024.

[35] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[36] Z. Xie, J. Zhang, W. Li, F. Zhang, and L. Zhang, "S-nerf: Neural radiance fields for street views," *arXiv preprint arXiv:2303.00749*, 2023.

[37] J. Guo, N. Deng, X. Li, Y. Bai, B. Shi, C. Wang, C. Ding, D. Wang, and Y. Li, "Streetsurf: Extending multi-view implicit surface reconstruction to street views," *arXiv preprint arXiv:2306.04988*, 2023.

[38] Z. Wu, T. Liu, L. Luo, Z. Zhong, J. Chen, H. Xiao, C. Hou, H. Lou, Y. Chen, R. Yang, *et al.*, "Mars: An instance-aware, modular and realistic simulator for autonomous driving," in *CAAI International Conference on Artificial Intelligence*. Springer, 2023, pp. 3–15.

[39] M. Broström, "BoxMOT: pluggable SOTA tracking modules for object detection, segmentation and pose estimation models." [Online]. Available: https://github.com/mikel-brostrom/boxmot

[40] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," 2017.

[41] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020, pp. 2446–2454.

[42] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[43] Y. Yue, A. Das, F. Engelmann, S. Tang, and J. E. Lenssen, "Improving 2d feature representations by 3d-aware fine-tuning," in *European Conference on Computer Vision*. Springer, 2024, pp. 57–74.