CVPR
#461

EMD: Explicit Motion Modeling for High-Quality Street Gaussian Splatting

Anonymous CVPR submission

Paper ID 461

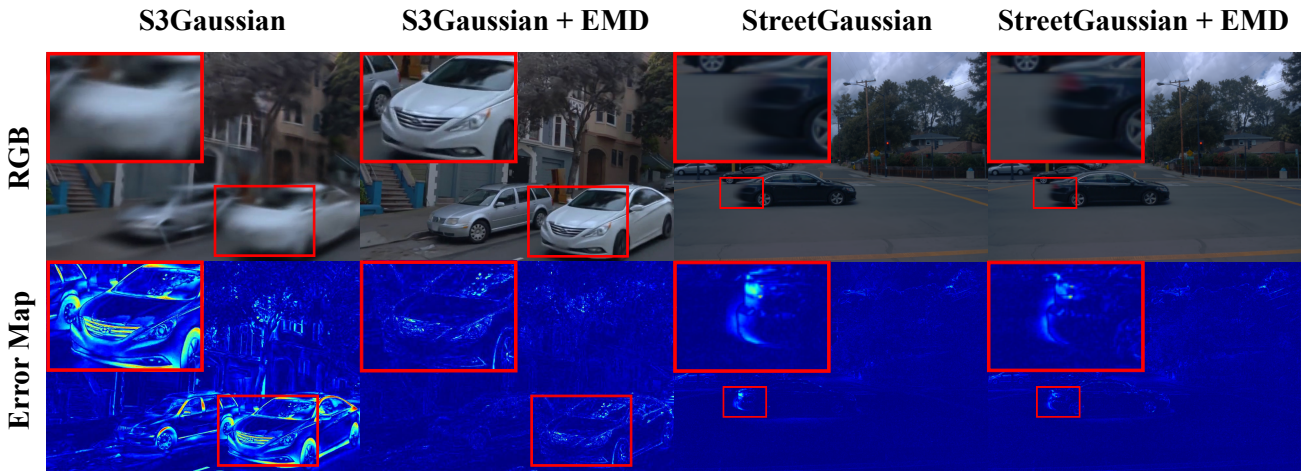

Figure 1. Previous street Gaussian splatting methods find it challenging to accurately model the motion of dynamic objects, which leads to blurry reconstructions. With the introduction of the proposed Explicit Motion Decomposition (EMD), which improves the decomposition of dynamic object motion, the current method achieves significantly better reconstruction quality.

Abstract

Photorealistic reconstruction of street scenes is essential for developing real-world simulators in autonomous driving. While recent methods based on 3D/4D Gaussian Splatting (GS) have demonstrated promising results, they still encounter challenges in complex street scenes due to the unpredictable motion of dynamic objects. Current methods typically decompose street scenes into static and dynamic objects, learning the Gaussians in either a supervised manner (e.g., w/ 3D bounding-box) or a self-supervised manner (e.g., w/o 3D bounding-box). However, these approaches do not effectively model the motions of dynamic objects (e.g., the motion speed of pedestrians is clearly different from that of vehicles), resulting in suboptimal scene decomposition. To address this, we propose Explicit Motion Decomposition (EMD), which models the motions of dynamic objects by introducing learnable motion embeddings to the Gaussians, enhancing the decomposition in street scenes. The proposed EMD is a plug-and-play approach applicable to various baseline methods. We also propose tailored training strategies to apply EMD to both supervised and self-supervised baselines. Through comprehensive experimentation, we illustrate the effectiveness of our approach with various established baselines. The code will be released.

1. Introduction

Novel view synthesis for dynamic scenes is essential in autonomous driving, enabling a variety of applications such as simulation, testing, and validation of perception systems. Traditional simulators like CARLA [10] and AirSim [36], while providing controlled environments, suffer from limited realism and require substantial manual effort in creating virtual scenes. Recent advances in neural rendering, particularly Neural Radiance Fields (NeRF) [26] and 3D Gaussian Splatting (3DGS) [16], have emerged as promising alternatives for photorealistic scene reconstruction. These pioneering approaches excel in capturing complex geometries and appearances through implicit or explicit neural representations, and have been extended to dynamic scenes by incorporating an additional time dimension into their representations and learning deformation networks to model non-rigid motions [4, 7, 20, 30–33, 42, 45, 49, 51]. Despite ad-

vancements, reconstructing autonomous driving scenes remains difficult due to complex multi-object dynamics, expansive environments, and varied motion patterns.

To tackle this challenge, existing works usually build upon dynamic NeRF and 3DGS frameworks, further separating autonomous driving scenes into static and dynamic components through two primary paradigms: Supervised methods utilize pre-trained models to acquire auxiliary conditions, such as segmentation masks from SAM [17], depth maps from DepthAnything [54], and both depth and optical flow from Dynamo [38], or 3D bounding boxes from various datasets. Representative works like StreetGaussian [52] demonstrate the effectiveness of using supervision signals for trajectory optimization and appearance modeling. In contrast, self-supervised methods achieve static-dynamic separation without explicit supervision, as shown by S3Gaussian [13], which leverages inherent motion cues. While both paradigms have shown promising results, their binary classification of scene elements as either static or dynamic overlooks the continuous spectrum of motion inherent in real-world street scenes.

To better address the different motion patterns in street scenes, we propose an Explicit Motion Decomposition (EMD) module that can be easily integrated into existing supervised and self-supervised frameworks (Fig. 2). EMD improves scene decomposition by incorporating motion-aware feature encoding and dual-scale deformation modeling. Specifically, we enhance each Gaussian primitive with learnable motion embeddings to capture its motion characteristics and design a hierarchical deformation framework that separately manages fast, global motions and slow, local deformations. This design allows for more efficient analysis of complex street scenes with varying motion speeds.

To showcase the versatility of our approach, we conduct extensive experiments on the Waymo-NOTR dataset by integrating EMD with representative methods from both paradigms: StreetGaussian for supervised and S3Gaussian for self-supervised settings. Our main contributions include:

- We propose EMD, a new plug-and-play module that effectively addresses varying motion speeds in street scenes through explicit motion modeling.
- We validate our method in both supervised and self-supervised settings, demonstrating consistent performance improvements across various evaluation protocols.
- We carried out a thorough evaluation of the Waymo-NOTR dataset, showcasing substantial improvements across various settings, achieving better reconstruction quality in both full scenes (+1.81 PSNR) and vehicle-specific regions (+2.81 PSNR) compared to respective baselines.

2. Related Work

Dynamic Scene Representation. Neural Radiance Fields (NeRF) [26] revolutionized novel view synthesis by introducing volumetric scene representation using MLPs[40], and has been significantly improved through various extensions [2, 3, 23, 27]. However, these methods face challenges in dynamic scene reconstruction due to the lack of temporal modeling capabilities. A common solution involves introducing time-dependent modeling through additional time conditions and deformation networks, as adopted by [30–32]. Additionally, several works also improve NeRF's reconstruction quality in both static and dynamic scenes by introducing additional supervision signals, such as semantic segmentation [15, 18, 57] , depth maps [8, 9, 12, 28, 50], and optical flow [5, 21]. Despite these advances, NeRF-based methods struggle with long training times and limited ability to handle complex motions. Recently, 3D Gaussian Splatting (3DGS) [16] has demonstrated remarkable potential by representing scenes with explicit 3D Gaussian primitives, achieving both high training efficiency and superior rendering quality. Inspired by NeRF-based approaches in handling dynamic scenes, several works have extended 3DGS to model temporal changes. Specifically, [25] proposes dynamic modeling by tracking Gaussian properties (position, orientation) while maintaining persistent appearance, while [56] extends the 3D representation to 4D by introducing temporal Gaussian distributions. Building upon these ideas, [55] designs dedicated deformation networks for motion modeling, [46] advances the efficiency by maintaining only canonical 3D Gaussians with HexPlane and [1] introduces embeddings to model fast and slow motions. These methods lay the groundwork for dynamic scene reconstruction in autonomous driving scenarios.

Autonomous Driving Simulation Traditional autonomous driving simulators like AirSim [36] and CARLA [10] require extensive manual effort for environment creation while struggling to achieve photorealistic rendering. To address this, Neural based approaches have emerged as promising solutions for simulating street scenes. Early NeRF-based methods [14, 29, 34, 43] introduced neural scene representation for large-scale urban environments, followed by improvements in efficiency and scalability [22, 24, 39, 44]. For dynamic urban scene reconstruction, NSG [29] introduced supervised scene decomposition using learned scene graphs and latent object representations. This supervised approach has since become a new paradigm, adopted by MARS [48], NeuRAD [41] and Multi-Level Neural Scene Graphs [11]. With the advent of 3D Gaussian Splatting, DrivingGaussian [59] also adopted this supervised paradigm and developed hierarchical scene representations combining dynamic

object graphs with incrementally updated static elements, while StreetGaussian [52] enhanced this framework with trajectory optimization and appearance modeling using 4D spherical harmonics. Recent work HUGS [58] further advances this line of research by incorporating physical constraints into the joint optimization of geometry, appearance, and semantics. To eliminate the need for expensive supervision, VDG [19] introduced pose-free reconstruction by integrating self-supervised visual odometry for pose estimation and depth initialization, while employing motion mask supervision for static-dynamic scene decomposition. Inspired by D2NeRF's [47] self-supervised decomposition in general dynamic scenes, SUDS [44] introduced this paradigm to autonomous driving by leveraging optical flow guidance, followed by EmerNeRF [53] proposing a flow-free approach with hash-grid based scene organization. Despite the dominance of NeRF-based methods, this self-supervised paradigm was later extended to 3D Gaussian Splatting by several works. PVG [6] pioneered this extension by introducing periodic vibration-based temporal dynamics for unified representation of both static and dynamic elements, followed by S3Gaussian [13].

However, current methods typically employ binary static-dynamic classification, which oversimplifies the complex motion patterns in street scenes. This limitation becomes particularly evident when reconstructing objects with varying velocities, as shown in our evaluations on the Waymo-NOTR dataset. To address this challenge, we propose Explicit Motion Decomposition (EMD), a plug-and-play approach that better captures the continuous spectrum of motion in street scenes. Before presenting our method, we first review the fundamentals of 3D and 4D Gaussian Splatting.

3. Preliminaries

3.1. 3D Gaussian Splatting

3D Gaussian Splatting (3D-GS) [16] proposes an explicit rendering approach to represent 3D scenes through a collection of 3D Gaussian primitives $\mathbb{G} = \{(\mu_k, \Sigma_k, \alpha_k, \mathbf{c}_k)\}_{k=1}^{K}$, where K is the total number of Gaussians. Each Gaussian primitive represents a probabilistic distribution of density in 3D space, defined by its probability density function:

$$G_k(x) = e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}, \tag{1}$$

where $x \in \mathbb{R}^3$ represents any point in the 3D world space, $\mu_k \in \mathbb{R}^3$ and $\Sigma_k \in \mathbb{R}^{3\times3}$ denote the mean position and covariance matrix in world space, respectively, where the covariance matrix determines the shape and orientation of the Gaussian, $\alpha_k \in [0, 1]$ is the opacity, and $\mathbf{c}_k$ encodes the view-dependent color information.

Rendering Process For rendering, each 3D Gaussian is projected onto the image plane, where the 3D mean μ_k is transformed to 2D mean μ_k^{2D}, and the world space covariance matrix Σ_k is transformed to screen space as $\Sigma_k' = JW\Sigma_k W^T J^T$, with W and J being the viewing transformation and projective transformation Jacobian matrices, respectively. The final pixel color at screen space position x is computed through front-to-back alpha compositing:

$$C(x) = \sum_{k\in\mathcal{N}(\mathbf{x})} \mathbf{c}_k \alpha_k(x) \prod_{j=1}^{k-1} (1 - \alpha_j(x)), \tag{2}$$

where $\alpha_k(x) = \alpha_k \exp\left(-\frac{1}{2}\left(x-\mu_k^{2D}\right)^T {\Sigma_k'}^{-1}\left(x-\mu_k^{2D}\right)\right)$ represents the opacity contribution of the k-th Gaussian at pixel x, computed as the product of the Gaussian's base opacity α_k and its projected 2D Gaussian evaluation. $\mathcal{N}(\mathbf{x})$ represents the set of indices of Gaussians intersecting pixel $x \in \mathbb{R}^2$.

In practice, for each Gaussian k, we parameterize its covariance matrix Σ_k using rotation quaternion $\mathbf{q}_k$ and scaling vector $\mathbf{s}_k$ as:

$$\Sigma_k = R(\mathbf{q}_k) S(\mathbf{s}_k) S(\mathbf{s}_k)^T R(\mathbf{q}_k)^T, \tag{3}$$

where $R(\mathbf{q}_k)$ is the rotation matrix defined by quaternion $\mathbf{q}_k$, and $S(\mathbf{s}_k)$ is the diagonal scaling matrix defined by scaling vector $\mathbf{s}_k$. The view-dependent color $\mathbf{c}_k$ is encoded using spherical harmonics (SH) coefficients:

$$\mathbf{c}_k(\mathbf{d}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} k_{l,m}^{(k)} Y_{l,m}(\mathbf{d}), \tag{4}$$

where $\mathbf{d}$ is the viewing direction, $Y_{l,m}$ are the spherical harmonic basis functions, and $k_{l,m}^{(k)}$ are the corresponding coefficients for the k-th Gaussian.

3.2. 4D Gaussian Splatting

4D Gaussian Splatting extends the static 3D-GS framework to handle dynamic scenes by incorporating temporal information. For a dynamic scene captured at different timestamps $t \in [0, T]$, each Gaussian primitive is now characterized by time-varying parameters: $\mathbb{G}(t) = \{(\mu_k(t), \Sigma_k(t), \alpha_k(t), \mathbf{c}_k(t))\}_{k=1}^{K}$. The probability density function of each Gaussian at time t becomes:

$$G_k(x, t) = e^{-\frac{1}{2}(x-\mu_k(t))^T \Sigma_k(t)^{-1}(x-\mu_k(t))}, \tag{5}$$

The temporal evolution of these parameters is typically modeled through one of two main approaches:

Deformation-based Approach This approach models temporal changes by applying a deformation field to the

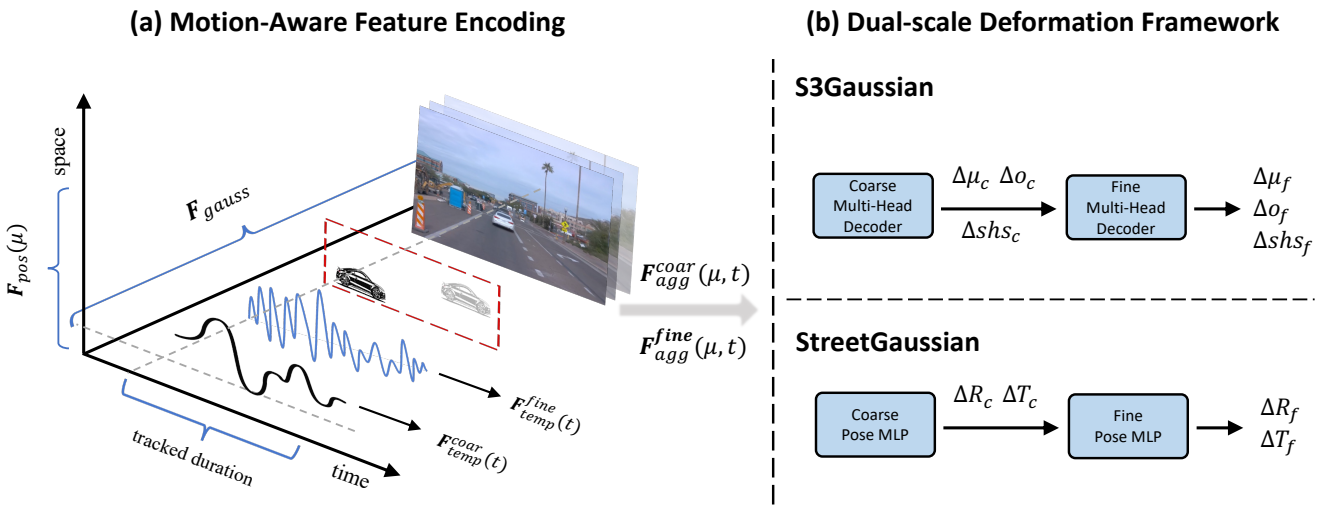

Figure 2. Overview of our Explicit Motion Decomposition (EMD) framework. Given input Gaussian primitives, our method processes them through two main components: (a) Motion-aware Feature Encoding, which combines spatial, temporal, and Gaussian-specific information to capture motion characteristics; and (b) Dual-scale Deformation Framework, which hierarchically models fast global motions and slow local deformations. The framework can be seamlessly integrated into both supervised (StreetGaussian) and self-supervised (S3Gaussian) approaches through our proposed integration strategies.

base Gaussians. For each timestamp t, the position of each Gaussian is updated as:

$$\mu_k(t) = \mu_k(0) + \Delta\mu_k(t), \tag{6}$$

where $\mu_k(0)$ is the initial position and $\Delta\mu_k(t)$ is the displacement predicted by a deformation network. Similarly, other parameters such as rotation, scaling, and color can be modeled as temporal offsets from their initial states.

Trajectory-based Approach Alternatively, the temporal evolution can be represented by explicitly modeling the continuous trajectory of each parameter. For instance, the position trajectory can be parameterized using a set of N control points $\{\mathbf{p}_i\}_{i=1}^{N}$ and basis functions $\{\phi_i(t)\}_{i=1}^{N}$:

$$\mu_k(t) = \sum_{i=1}^{N} \mathbf{p}_i \phi_i(t), \tag{7}$$

where the basis functions $\phi_i(t)$ can be B-splines or other temporal interpolation functions. The rendering process remains similar to static 3D-GS, but now Gaussian parameters are evaluated at the specific timestamp t before projection and compositing.

4. Methodology

4D Gaussian Splatting has shown promising results in dynamic scene reconstruction. However, modeling dynamic street scenes remains challenging due to diverse motion patterns. To better handle the varying motion patterns in street scenes, we propose Explicit Motion Decomposition (EMD), a plug-and-play module that can be seamlessly integrated into existing 4D Gaussian-based frameworks to enhance their capability in handling dynamic scenarios, as illustrated in Fig. 2.

4.1. Problem Formulation

Given a set of static 3D Gaussian primitives $\mathbb{G} = \{(\mu_k, \mathbf{s}_k, \mathbf{q}_k, \alpha_k, \mathbf{c}_k)\}_{k=1}^{K}$ and a timestamp t, our goal is to learn a deformation field $\mathcal{D}$ that maps each Gaussian's parameters from their canonical states to their corresponding deformed states at time t. For notational simplicity, we omit the Gaussian index k in the following formulation:

$$\{\mu_t, \mathbf{s}_t, \mathbf{q}_t, \alpha_t, \mathbf{c}_t\} = \mathcal{D}(\{\mu, \mathbf{s}, \mathbf{q}, \alpha, \mathbf{c}\}, t), \tag{8}$$

To effectively handle the diverse motion patterns in street scenes, especially the distinct movements between vehicles and pedestrians, we propose a motion-aware deformation module that processes input Gaussian parameters through two key components: motion-aware feature encoding and dual-scale deformation prediction.

4.1.1. Motion-aware Feature Encoding

Different types of objects in street scenes exhibit distinct motion characteristics. To capture these varied patterns, we first encode the input Gaussian parameters into a comprehensive feature space that combines spatial, temporal, and Gaussian-specific information:

$$\mathbf{F}_{aggr}(\mu, t) = [\mathbf{F}_{pos}(\mu), \mathbf{F}_{temp}(t), \mathbf{F}_{gauss}], \tag{9}$$

The spatial component employs multi-frequency positional encoding:

$$\mathbf{F}_{pos}(\mu) = [\mu, \{\sin(2^i\pi\mu), \cos(2^i\pi\mu)\}_{i=0}^{P-1}], \quad (10)$$

where P is the number of frequency bands, enabling the network to capture both fine geometric details and global structures. For temporal information, we design an adaptive temporal embedding function:

$$\mathbf{F}_{temp}(t) = \mathcal{T}(t, N(i)) = \text{Interp}(\mathbf{W}, t, N(i)), \quad (11)$$

where $\mathbf{W} \in \mathbb{R}^{N_{max} \times D}$ is a learnable embedding matrix that captures motion patterns, and $N(i)$ progressively increases from N_{min} to N_{max} temporal samples during training iteration i, allowing the model to gradually capture finer temporal dynamics. For Gaussian-specific features, we assign a learnable latent embedding $\mathbf{z}_k \in \mathbb{R}^D$ to each Gaussian k, where $\mathbf{F}_{gauss} = \mathbf{z}_k$, enabling the model to learn and represent individual motion characteristics.

4.1.2. Dual-scale Deformation Framework

Given the diverse motion patterns in street scenes, ranging from large vehicular movements to subtle pedestrian motions, we design a hierarchical deformation framework that can effectively handle both scales of motion:

$$\begin{aligned}\mathcal{D}(\mu, t) =& \mathcal{D}_{coarse}(\mathbf{F}_{aggr}(\mu, t)) \\ &+ \mathcal{D}_{fine}(\mathbf{F}_{aggr}(\mu + \Delta\mu_{coarse}, t)),\end{aligned} \quad (12)$$

The final deformed parameters combine both coarse and fine scale predictions:

$$\begin{aligned}\mu_t &= \mu + \Delta\mu_{coarse} + \Delta\mu_{fine} \\ \mathbf{s}_t &= \mathbf{s} + \Delta\mathbf{s}_{coarse} + \Delta\mathbf{s}_{fine} \\ \mathbf{q}_t &= \mathbf{q} \otimes \Delta\mathbf{q}_{coarse} \otimes \Delta\mathbf{q}_{fine},\end{aligned} \quad (13)$$

where $\mathcal{D}_{coarse}$ focuses on modeling large-scale motions such as vehicle translations, while $\mathcal{D}_{fine}$ captures local deformations like articulated movements. Similar to position updates, other Gaussian parameters including opacity α_t and spherical harmonics coefficients $\mathbf{c}_t$ are also updated through this dual-scale framework.

4.2. Integration with Existing Frameworks

As mentioned in the introduction, current approaches for street scene reconstruction generally fall into supervised and self-supervised paradigms. Having formalized our EMD framework, we now demonstrate its integration into representative methods from both paradigms: StreetGaussian for supervised learning and S3Gaussian for self-supervised learning.

4.2.1. Self-supervised Integration: S3Gaussian

For self-supervised scenarios, we enhance S3Gaussian's architecture with our motion-aware features. While S3Gaussian originally employs a Multi-head Gaussian Decoder for deformation prediction, we augment each Gaussian with our learnable embedding $\mathbf{z}_k$ and restructure its decoder into our dual-scale framework. Specifically, both coarse and fine stages predict deformations in position ($\Delta\mu$), opacity ($\Delta\alpha$), and spherical harmonics coefficients ($\Delta\mathbf{c}$), enabling more precise motion modeling through hierarchical refinement.

4.2.2. Supervised Integration: StreetGaussian

For supervised settings, StreetGaussian provides a framework that represents dynamic objects through tracked vehicle poses and object-specific Gaussians. Each object is characterized by tracked poses $\{R_t, T_t\}_{t=1}^{N_t}$ that transform object Gaussians from local coordinates (μ_o, R_o) to world coordinates (μ_w, R_w). To enhance its motion modeling capability, we also augment each Gaussian with our learnable embedding $\mathbf{z}_k$ and apply temporal embedding only within each object's tracked duration, effectively capturing object-specific temporal dynamics.

To address the challenge of noisy tracked poses while maintaining their geometric meaning, we incorporate our dual-scale framework into their pose optimization:

$$\begin{aligned}R'_t &= R_t(\Delta R_t^c + \Delta R_t^f) \\ T'_t &= T_t + (\Delta T_t^c + \Delta T_t^f),\end{aligned} \quad (14)$$

where $\Delta R_t^c, \Delta T_t^c$ handle large pose corrections and $\Delta R_t^f, \Delta T_t^f$ capture subtle adjustments. This decomposition allows the model to effectively correct tracking errors while preserving the physical meaning of the tracked poses. Similarly, we apply the dual-scale framework to appearance modeling, where spherical harmonics coefficients are refined through both coarse and fine stages, enabling more accurate dynamic appearance representation.

We adopt the same loss function as S3Gaussian and StreetGaussian to train the entire pipeline. For further training details, please refer to the supplementary materials.

5. Experiments

In this section, we first describe the datasets and evaluation metrics used in our experiments (Sec. 5.1). Then, we demonstrate the experimental results of our approach under both self-supervised and supervised settings (Sec. 5.2). Finally, we present comprehensive ablation studies to validate the effectiveness of individual components in our framework (Sec. 5.3).

5.1. Datasets and Metrics

Dataset. We conduct extensive experiments on the Waymo Open Dataset [37] to comprehensively evaluate our

Table 1. Comparative performance of our framework and baseline approaches on the Waymo-NOTR dataset. The best performances are highlighted in **bold**, and the second-best are indicated with underlining. ↑ indicates higher is better, while ↓ indicates lower is better.

Dataset	Methods	Scene Reconstruction					Novel View Synthesis				
		Full Image			Vehicle		Full Image			Vehicle	
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
D32	EmerNeRF [53]	28.16	0.806	0.228	24.32	0.682	25.14	0.747	0.313	**23.49**	0.660
	3DGS [16]	28.47	0.876	0.136	23.26	0.716	25.14	0.813	0.165	20.48	**0.753**
	MARS [48]	28.24	0.866	0.252	23.37	0.701	26.61	0.796	0.305	22.21	0.697
	S3Gaussian [13]	30.69	0.900	0.121	26.23	0.804	**26.62**	0.824	0.159	22.61	0.681
	S3Gaussian+Ours	**32.50**	**0.933**	**0.082**	**29.04**	**0.879**	26.55	**0.833**	**0.126**	23.39	0.703

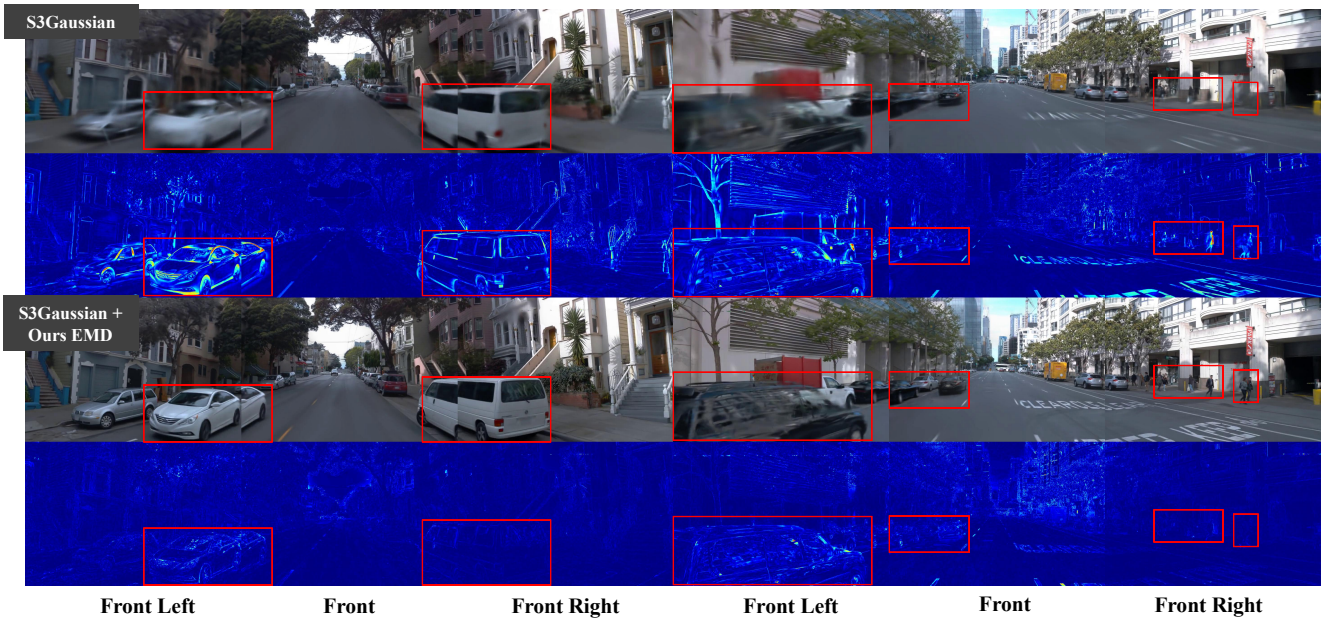

Figure 3. Qualitative comparison on the self-supervised setting between S3Gaussian and S3Gaussian+ours EMD. We also visualize the error maps between the rendered images and ground truth to provide further insights.

method. To validate the effectiveness of our plug-and-play module, we select representative state-of-the-art methods from both supervised and self-supervised paradigms: StreetGaussian [52] and S3Gaussian [13]. For self-supervised evaluation, we use the dynamic32 (D32) split introduced by EmerNeRF [53], containing 32 sequences with vehicle motion, where each sequence consists of approximately 50-100 frames captured by three cameras under various conditions. For supervised evaluation, we follow the data split protocol established in StreetGaussian [52] to enable direct comparison.

Evaluation Metrics. We employ comprehensive metrics to evaluate both reconstruction quality and novel view synthesis capability. For scene reconstruction, we use PSNR and SSIM to evaluate both full-scene and vehicle-specific reconstruction quality. Additionally, we compute LPIPS for perceptual quality assessment. Following previous protocols, we evaluate novel view synthesis on every 10th frame for self-supervised settings and every 4th frame for supervised settings.

5.2. Main Results

5.2.1. Self-supervised Performance

Tab. 1 presents the comparative results on the D32 split, where no 3D bounding box annotations were used. Our method significantly outperforms previous self-supervised approaches, achieving notable improvements in both full-scene and object-specific metrics. Specifically, we observe a substantial increase in PSNR for the full scene (32.50 vs. 30.69) and for vehicle-specific metrics (PSNR: 29.04 vs. 26.23) when compared to S3Gaussian. These results demonstrate the enhanced ability of our method to accurately model complex street scenes without explicit 3D box

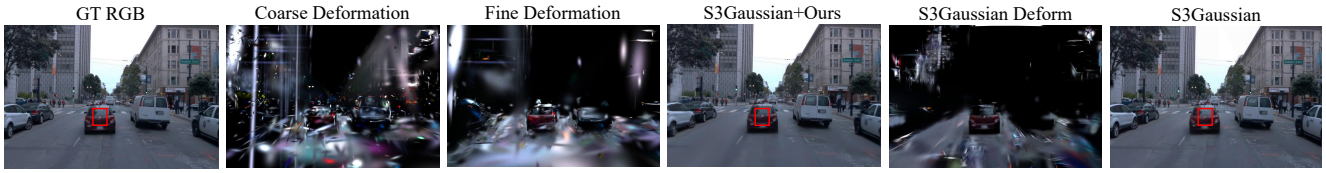

Figure 4. Motion deformation comparison between S3Gaussian + Ours and S3Gaussian. Please zoom in for more details.

annotations. For novel view synthesis, our method continues to perform competitively, achieving a PSNR of 26.55, which reflects its robust generalization to previously unseen viewpoints.

In addition, we present a side-by-side visualization comparison between S3Gaussian and S3Gaussian+EMD in Fig. 3. The error maps, which compare the ground truth to the rendered results, clearly demonstrate that S3Gaussian+EMD outperforms S3Gaussian in modeling dynamic objects with varying motion speeds. S3Gaussian implicitly models dynamic vehicles, but it fails to capture changes in speed during motion and the differences in movement between vehicles, leading to blurred reconstructions. on the contrary, our method captures the distinct motion characteristics of different dynamic objects, leading to more accurate and consistent scene reconstructions. This emphasizes the effectiveness of Explicit Motion Decomposition (EMD) in modeling the motion of dynamic objects, improving the overall decomposition and photorealistic rendering of street scenes.

We also present a visual comparison of motion between S3Gaussian and S3Gaussian+EMD. As shown in Fig. 4, the proposed dual-scale deformation network generates a coarse deformation to model slower motion and larger-scale geometry and a fine deformation to capture faster motion and finer geometric details in the scene. With the incorporation of EMD, S3Gaussian can capture detailed features of dynamic vehicles, including the car brand logo, as demonstrated in the figure. In contrast, S3Gaussian treats the entire moving car as a single dynamic object, failing to produce clear synthesis results for the dynamic vehicles.

5.2.2. Supervised Performance

To demonstrate the versatility of our framework, we also evaluate it in the supervised setting using the same scenes as StreetGaussian [52] (Tab. 2). When incorporating 3D bounding box supervision through our adaptive training scheme, our method achieves superior performance in full scene reconstruction, improving PSNR by 1.42dB (36.03 vs. 34.61), SSIM by 1.1% (0.949 vs. 0.938), and LPIPS by 15.2% (0.067 vs. 0.079). These results demonstrate that our approach effectively enhances the overall scene reconstruction quality while maintaining competitive performance in dynamic object modeling. Please refer to the supplementary for visualization on the supervised setting.

Table 2. Comparative performance of our framework and baseline approaches on the StreetGaussian dataset. The best performances are highlighted in **bold**. ↑ indicates higher is better, while ↓ indicates lower is better.

Methods	Full Image			Vehicle
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑
3DGS [16]	29.64	0.918	0.117	21.25
NSG [29]	28.31	0.862	0.346	24.32
MARS [48]	29.75	0.886	0.264	26.54
EmerNeRF [53]	30.87	0.905	0.133	21.67
StreetGaussian [52]	34.61	0.938	0.079	**30.23**
StreetGaussian + Ours	**36.03**	**0.949**	**0.067**	29.81

Table 3. Ablation study on the D32 dataset showing the impact of different components in our framework. All variants are evaluated using the self-supervised setting. The best performances are highlighted in **bold**. ↑ indicates higher is better, while ↓ indicates lower is better.

Variant	Full Image			Vehicle
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑
Full Model	**32.50**	**0.933**	**0.082**	**29.04**
w/o Gaussian Embedding	32.21	0.928	0.089	28.80
w/o Temporal Embedding	32.23	0.922	0.091	28.08
w/o Coarse Deformation	29.40	0.890	0.146	24.54
w/o Fine Deformation	32.45	0.931	0.118	28.80

5.3. Ablation Studies

To assess the contribution of each component in our framework, we perform comprehensive ablation studies, as shown in Tab. 3 and Fig. 5. The results reveal the critical role of the Gaussian embedding, as its removal leads to the performance drop, with a reduction of 0.29 PSNR (32.21 vs. 32.50). This indicates that the Gaussian embedding is essential for effectively capturing the motion characteristics for each dynamic gaussian. The temporal embedding also plays a crucial role, with its absence leading to a 0.27 PSNR drop (32.23 vs. 32.50), underscoring its importance in modeling the temporal variation of object motion over time.

Both the coarse and fine deformation components are integral to the final performance, with their removal leading to considerable performance degradation. Specifically, excluding the coarse deformation component causes a significant 3.10 PSNR drop (29.40 vs. 32.50), suggesting that

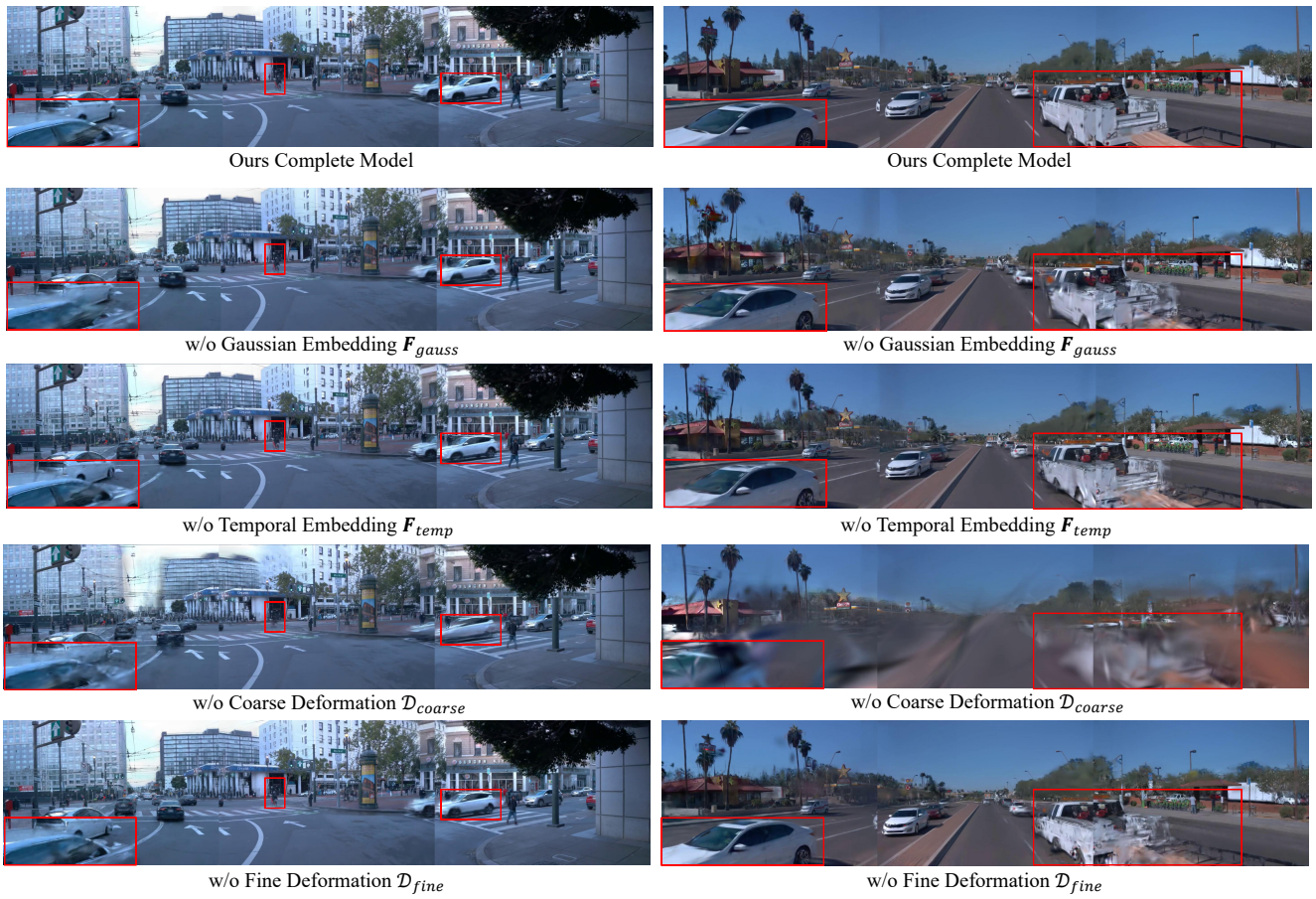

Figure 5. Qualitative ablation study results across three camera views from the Waymo dataset. (a) Our complete model achieves sharp and consistent reconstruction. (b) Removing Gaussian embedding $\mathcal{F}_{gauss}$ leads to blurry object boundaries. (c) Without temporal embedding $\mathcal{F}_{temp}$, results show motion artifacts. (d) Without coarse deformation $\mathcal{D}_{coarse}$, geometric consistency is lost. (e) Absence of fine deformation $\mathcal{D}_{fine}$ causes detail degradation.

the coarse adjustments are vital for maintaining the overall scene structure. On the other hand, removing the fine deformation component results in a 0.036 LPIPS increase (0.118 vs. 0.082), implying that fine deformation refines the details and its absence worsens the perceptual quality. The ablation study demonstrates that our proposed motion-aware feature encoding and dual-scale deformation effectively model dynamic objects with varying motion speeds, which is crucial for enhancing the reconstruction quality of existing street Gaussians.

6. Limitation

Although EMD effectively addresses the challenge of modeling dynamic objects with varying speeds by incorporating learnable embeddings, some limitations remain. Existing street Gaussian methods do not account for environmental lighting, yet lighting effects play a crucial role in the quality of reconstructions under different lighting conditions. In future work, we plan to explore the possibility of developing a plug-and-play technique to enhance lighting effects in existing methods.

7. Conclusion

In this paper, we present EMD, a plug-and-play module that effectively handles varying motion speeds in street scene reconstruction. By introducing motion-aware feature encoding and dual-scale deformation modeling, our approach successfully captures the continuous spectrum of motion patterns inherent in real-world scenarios. Comprehensive experiments on the Waymo Open Dataset demonstrate that EMD significantly improves reconstruction quality when integrated with both supervised and self-supervised frameworks. Our method substantially improves full scene reconstruction (+1.81 PSNR) and vehicle-specific regions (+2.81 PSNR), validating the effectiveness of explicit motion modeling. We believe our work opens up new possibilities for high-quality dynamic scene reconstruction in autonomous driving applications, and the plug-and-play nature of our approach makes it readily applicable to future developments in neural scene representations.

References

[1] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. *arXiv preprint arXiv:2404.03613*, 2024. 2

[2] 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, 2021. 2

[3] 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. 2

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 1

[5] Quei-An Chen and Akihiro Tsukada. Flow supervised neural radiance fields for static-dynamic decomposition. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10641–10647, 2022. 2

[6] Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv:2311.18561*, 2023. 3

[7] Simon Le Cleac'h, Hong Yu, Michelle Guo, Taylor A. Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8:2780–2787, 2022. 1

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[10] 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. 1, 2

[11] Tobias Fischer, Lorenzo Porzi, Samuel Rota Bulo, Marc Pollefeys, and Peter Kontschieder. Multi-level neural scene graphs for dynamic urban environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21125–21135, 2024. 2

[12] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 2

[13] 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, 6

[14] Sheng Yu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18190–18200, 2023. 2

[15] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis, 2021. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 1, 2, 3, 6, 7

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[18] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation. In *CVPR*, 2022. 2

[19] Hao Li, Jingfeng Li, Dingwen Zhang, Chenming Wu, Jieqi Shi, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Junwei Han. Vdg: Vision-only dynamic gaussian for driving simulation. *arXiv preprint arXiv:2406.18198*, 2024. 3

[20] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[21] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[22] Jeffrey Yunfan Liu, Yun Chen, Ze Yang, Jingkang Wang, Sivabalan Manivasagam, and Raquel Urtasun. Real-time neural rasterization for large scenes. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8382–8393, 2023. 2

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 2

[24] Fan Lu, Yan Xu, Guang-Sheng Chen, Hongsheng Li, Kwan-Yee Lin, and Changjun Jiang. Urban radiance field representation with deformable neural mesh primitives. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 465–476, 2023. 2

[25] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 2

[26] 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, 2020. 1, 2

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[28] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 2

[29] 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 (CVPR)*, pages 2856–2865, 2021. 2, 7

[30] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1, 2

[31] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021.

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[33] Saskia Rabich, Patrick Stotko, and Reinhard Klein. Fpo++: Efficient encoding and rendering of dynamic neural radiance fields by analyzing and enhancing fourier plenoctrees. *ArXiv*, abs/2310.20710, 2023. 1

[34] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. *CVPR*, 2022. 2

[35] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[36] 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. 1, 2

[37] 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. 5

[38] Yihong Sun and Bharath Hariharan. Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[39] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8248, 2022. 2

[40] Hind Taud and Jean-Franccois Mas. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, pages 451–455, 2018. 2

[41] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. *arXiv preprint arXiv:2311.15260*, 2023. 2

[42] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video, 2020. 1

[43] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly- throughs. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12912–12921, 2021. 2

[44] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[45] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13524–13534, 2022. 1

[46] 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

[47] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Öztireli. D2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *ArXiv*, abs/2205.15838, 2022. 3

[48] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, Yuxin Huang, Xiaoyu Ye, Zike Yan, Yongliang Shi, Yiyi Liao, and Hao Zhao. Mars: An instance-aware, modular and realistic simulator for autonomous driving. *CICAI*, 2023. 2, 6, 7

[49] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9416–9426, 2020. 1

[50] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[51] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Neural Information Processing Systems*, 2021. 1

[52] 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. 2, 3, 6, 7

[53] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 3, 6, 7

[54] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[55] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2

[56] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[57] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2

[58] 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 (CVPR)*, pages 21336–21345, 2024. 3

[59] 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. 2