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

Xiaobao Wei[1,2,3,4,*,†] Qingpo Wuwu[1,2,*,†] Zhongyu Zhao[1,2,†] Zhuangzhe Wu[1]
Nan Huang[1] Ming Lu[1] Ningning Ma[2] Shanghang Zhang[1,✉]
[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2]Autonomous Driving Development, NIO [3]Institute of Software, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences
weixiaobao0210@gmail.com

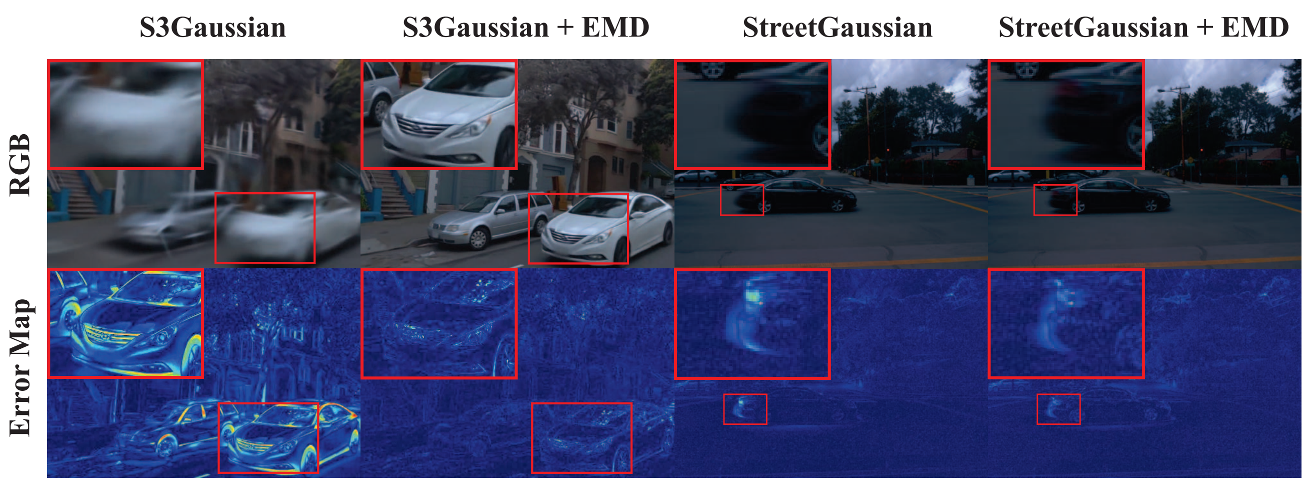

Figure 1. Previous street Gaussian splatting methods find it challenging to accurately model the motion patterns of dynamic objects, which leads to blurry reconstructions. With the introduction of the proposed Explicit Motion Decomposition (EMD), which compensates for the modeling of dynamic object motion, achieving the state-of-the-art 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 plug-and-play EMD module compensates for the lack of motion modeling in self-supervised street Gaussian splatting methods. We also introduce tailored training strategies to extend EMD to supervised approaches. Comprehensive experiments demonstrate the effectiveness of our method, achieving state-of-the-art novel view synthesis performance in self-supervised settings. The code is available at: https://qingpowuwu.github.io/emd.

1. Introduction

Novel view synthesis for dynamic street scenes is essential in closed-loop autonomous driving. Recent advances in neural rendering, particularly Neural Radiance Fields (NeRF) [31] and 3D Gaussian Splatting (3DGS) [19], have emerged as promising scene reconstruction methods. These

*Equal contribution.
†Work done during internship at NIO.
✉ Corresponding author.

pioneering approaches excel in capturing complex geometries and appearances through implicit or explicit neural representations and have been extended to dynamic scenes [3, 7, 23, 36, 37, 39, 40, 48, 51–53, 56, 59]. Despite advancements, reconstructing autonomous driving scenes remains difficult due to complex multi-object dynamics, complex environments, and varied motion patterns.

To tackle this challenge, existing works based on dynamic NeRF and 3DGS frameworks separate autonomous driving scenes into static and dynamic components through two primary paradigms: Supervised methods acquire priors for dynamic objects, such as segmentation masks [20], optical flow [45], or 3D bounding boxes from object tracking. Representative works like StreetGaussian [60] and OminiRe [6] demonstrate the effectiveness of using supervision signals for appearance modeling of dynamic objects. In contrast, self-supervised methods achieve static-dynamic separation without explicit supervision, as shown by S3Gaussian [16] and DeSiRe-GS [38], which leverage inherent motion cues to optimize a 4D street representation.

While both paradigms yield promising results, supervised methods adopt a binary classification of scene elements as either static or dynamic, which overlooks the continuous spectrum of motion inherent in street scenes. Self-supervised methods optimize the entire scene holistically but neglect the varied motion speeds among objects—for example, pedestrians generally move much slower than vehicles. Existing supervised methods [6, 60] mitigate dynamic errors by optimizing 3D bounding boxes, yet self-supervised approaches still lack an effective motion modeling mechanism.

To address the different motion patterns in street scenes, we propose an Explicit Motion Decomposition (EMD) module that can be easily integrated into existing 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. We conduct extensive experiments on the Waymo and KITTI datasets by integrating EMD with representative self-supervised methods: S3Gaussian and DeSiRe-GS. In addition, EMD can be seamlessly extended to supervised methods. Our main contributions include:

- We propose EMD, the first plug-and-play module that effectively addresses varying motion speeds in street scenes through explicit motion modeling.
- We introduce tailored training strategies for self-supervised street Gaussian splatting methods and further extend EMD to supervised settings.
- Comprehensive experiments on Waymo and KITTI datasets demonstrate previous methods with EMD exhibit better reconstruction quality, achieving state-of-the-art (SOTA) performance in self-supervised settings.

2. Related Work

Dynamic Scene Representation. Neural Radiance Fields (NeRF) [31] revolutionizes novel view synthesis by introducing volumetric scene representation and has been improved through various extensions [1, 2, 26, 32]. However, these methods are restricted to static scenes. A common solution involves introducing additional time conditions, as adopted by [36, 37, 39]. By introducing additional supervision signals, several works [4, 8, 9, 14, 18, 21, 24, 33, 58, 66] improve the rendering quality in both static and dynamic scenes. Despite these advances, NeRF-based methods struggle with long training times and limited ability to handle complex motions. Recently, 3D Gaussian Splatting (3DGS) [19] has achieved both high efficiency and superior rendering quality by representing scenes with explicit 3D Gaussian primitives. Several works [30, 54, 62, 64] have extended 3DGS to incorporate temporal information, enabling the 4D scene representation. These methods lay the groundwork for dynamic scene reconstruction in autonomous driving scenarios.

Autonomous Driving Simulation. Traditional autonomous driving simulators like AirSim [43] and CARLA [10] require extensive manual effort for environment creation while struggling to achieve photorealistic rendering. To address this approaches based on neural fields have emerged as promising solutions for simulating street scenes. Early NeRF-based methods [17, 34, 41, 49] introduce neural scene representation for large-scale urban environments, followed by improvements in efficiency and scalability [25, 28, 46, 50]. For dynamic urban scene reconstruction, several methods [11, 34, 47, 55] introduce supervised scene decomposition using learned scene graphs and latent object representations. With the advent of 3DGS, DrivingGaussian [68], StreetGaussian [60], HUGS [67] and OmniRe [6] also adopt this supervised paradigm and develop hierarchical scene representations combining dynamic object graphs with incrementally updated static elements. To eliminate the need for expensive supervision, SUDS [50] and EmerNeRF [61] leverage the optical flow as decomposition guidance. PVG [5], VDG [22], S3Gaussian [16] and DeSiRe-GS [38] extend the self-supervised setting into 3DGS by a unified representation of both static and dynamic elements.

However, existing self-supervised methods focus on the scene decomposition, overlooking the diverse motion patterns in street environments. This limitation becomes particularly pronounced when reconstructing objects moving at significantly different speeds. To tackle this challenge, we are the first to explore explicit motion modeling and introduce a plug-and-play Explicit Motion Decomposition (EMD) method. EMD enhances self-supervised approaches by capturing the continuous spectrum of motion, and it can be seamlessly extended to supervised methods.

3. Preliminaries

3.1. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [19] 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 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 with spherical harmonics.

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. The pixel color at screen space position x is computed through alpha blending:

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

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. $\mathcal{N}(\mathbf{x})$ 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 $\mathbf{q}_k$, and $S(\mathbf{s}_k)$ is the diagonal scaling matrix defined by $\mathbf{s}_k$.

3.2. 4D Gaussian Splatting

4D Gaussian Splatting (4DGS) [54, 63] extend the static 3DGS 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:

$$G_k(x,t) = e^{-\frac{1}{2}(x-\mu_k(t))^T \Sigma_k(t)^{-1}(x-\mu_k(t))}, \tag{4}$$

Then a deformation field is applied to Gaussians. For each timestamp t, the position of each Gaussian is updated as:

$$\mu_k(t) = \mu_k(0) + \Delta\mu_k(t), \tag{5}$$

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.

4. Methodology

4DGS 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 street Gaussian Splatting methods 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:

$$\{\mu_t, \mathbf{s}_t, \mathbf{q}_t, \alpha_t, \mathbf{c}_t\} = \mathcal{D}(\{\mu, \mathbf{s}, \mathbf{q}, \alpha, \mathbf{c}\}, t), \tag{6}$$

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:

$$\mathbf{F}_{aggr}(\mu,t) = [\mathbf{F}_{pos}(\mu), \mathbf{F}_{temp}(t), \mathbf{F}_{gauss}], \tag{7}$$

The spatial component employs multi-frequency positional encoding:

$$\mathbf{F}_{pos}(\mu) = [\mu, \{\sin(2^i\pi\mu), \cos(2^i\pi\mu)\}_{i=0}^{P-1}], \tag{8}$$

where $P = 10$ 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:

$$\mathbf{F}_{temp}(t) = \mathcal{T}(t, N(i)) = \text{Interp}(\mathbf{W}, t, N(i)), \tag{9}$$

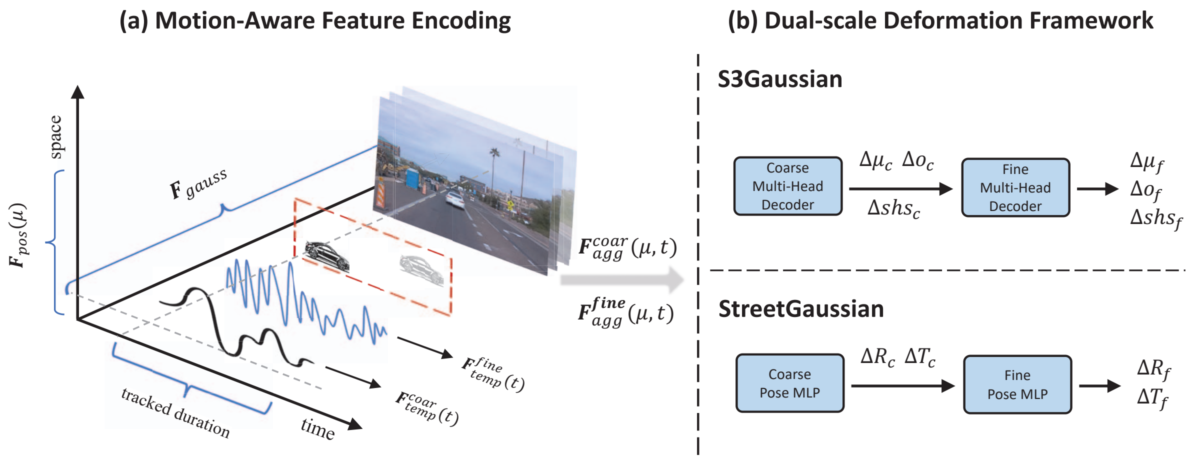

Figure 2. Overview of our Explicit Motion Decomposition (EMD) framework. Given input Gaussian primitives, our method processes them through two main components: (a) Motion-aware Feature Encoding, which combines spatial, temporal, and Gaussian-specific information to capture motion characteristics; and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

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 $N_{min} = 30$ to $N_{max} = 150$ temporal samples during training iteration i, allowing the model to gradually capture finer temporal dynamics. $D = 4$ is the temporal embedding dimension. During temporal embedding $\mathbf{W}$ optimization, we first downsample $\mathbf{W}$ through bilinear interpolation to create an intermediate embedding matrix of size $N(i) \times D$ at training iteration i, which can be formulated as:

$$N(i) = N_{min} + (N_{max} - N_{min}) \cdot \min(i, T)/T, \quad (10)$$

where hyperparameter $T = 25000$ controls the duration of progressive sampling refinement. Then, for each time stamp t, we perform grid sampling on this intermediate matrix with bilinear interpolation mode to obtain the corresponding temporal embedding vector $\mathbf{F}_{temp}(t)$. For Gaussian-specific features, we assign a learnable latent embedding $\mathbf{z}_k \in \mathbb{R}^M$ to each Gaussian k, where $\mathbf{F}_{gauss} = \mathbf{z}_k$ and $M = 32$, enabling the model to 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:

$$\begin{aligned}\mathcal{D}(\mu, t) =& \mathcal{D}_{coarse}(\mathbf{F}_{aggr}(\mu, t)) \\ &+ \mathcal{D}_{fine}(\mathbf{F}_{aggr}(\mu + \Delta\mu_{coarse}, t)),\end{aligned} \quad (11)$$

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

$$\begin{aligned}\mu_t &= \mu + \Delta\mu_{coarse} + \Delta\mu_{fine} \\ \mathbf{s}_t &= \mathbf{s} + \Delta\mathbf{s}_{coarse} + \Delta\mathbf{s}_{fine} \\ \mathbf{q}_t &= \mathbf{q} \otimes \Delta\mathbf{q}_{coarse} \otimes \Delta\mathbf{q}_{fine},\end{aligned} \quad (12)$$

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

Current approaches for street scene reconstruction generally fall into supervised and self-supervised paradigms. Having formalized our EMD framework, we now illustrate its integration into representative supervised methods. Then we further extend EMD into supervised approaches.

4.2.1. Self-supervised Integration

For self-supervised street Gaussian splatting methods, we enhance S3Gaussian [16] and DeSiRe-GS [38] with our motion-aware techniques. S3Gaussian originally employs a HexPlane to model the dynamic and static element decomposition and utilizes 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. As for DeSiRe-GS built upon PVG [5], we also assign the learnable Gaussian embedding $\mathbf{z}_k$ into the framework. Then we apply the deformation for the position ($\Delta\mu$) and spherical harmonics coefficients ($\Delta\mathbf{c}$) with the proposed dual-scale deformation framework. We retain the same self-supervised deformation settings as those used in the baseline methods.

4.2.2. Supervised Extension

For supervised settings, though existing methods refine tracked 3D bounding boxes to eliminate rendering errors, we

Table 1. S3Gaussian comparison: results on the Waymo Open dataset for scene reconstruction and novel view synthesis.

Dataset	Methods	Scene Reconstruction					Novel View Synthesis				
		Full Image			Vehicle		Full Image			Vehicle	
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
Waymo-D32	EmerNeRF [61]	28.16	0.806	0.228	24.32	0.682	25.14	0.747	0.313	**23.49**	0.660
	3DGS [19]	28.47	0.876	0.136	23.26	0.716	25.14	0.813	0.165	20.48	**0.753**
	MARS [55]	28.24	0.866	0.252	23.37	0.701	26.61	0.796	0.305	22.21	0.697
	S3Gaussian [16]	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

extend EMD into the refinement process, illustrating the need for explicit motion modeling. We select StreetGaussian [60] and OmniRe [6] as our baselines. 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. Finally, we incorporate our dual-scale framework into the pose optimization:

$$\begin{aligned} R_t' &= \Delta R_t^f \cdot \Delta R_t^c \cdot R_t \\ T_t' &= T_t + (\Delta T_t^c + \Delta T_t^f), \end{aligned} \tag{13}$$

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. Similarly, we apply the dual-scale framework to appearance modeling, where spherical harmonics coefficients are refined through both coarse and fine stages.

For OmniRe, it further proposes non-rigid SMPL [27] nodes for human modeling. For the rigid nodes, we apply the same tracked box refinement in StreetGaussian into OmniRe. We further implement the dual-scale refinement to the SMPL model including pose parameters $\theta_t \in \mathbb{R}^{24\times3\times3}$ and shape parameters $\beta_t \in \mathbb{R}^{10}$:

$$\begin{aligned} \theta_t' &= \Delta\theta_t^f \cdot \Delta\theta_t^c \cdot \theta_t \\ \beta_t' &= \beta_t + (\Delta\beta_t^c + \Delta\beta_t^f), \end{aligned} \tag{14}$$

For training, in addition to using the same loss function as the baselines, we implement a local smoothness regularization for the learnable Gaussian embeddings. Inspired by [29], this regularization encourages neighboring Gaussians i and j to have similar representations:

$$\mathcal{L}_{\mathbf{z}_k} = \frac{1}{d|\mathcal{U}|} \sum_{i\in\mathcal{U}} \sum_{j\in \text{KNN}_{i;d}} (e^{-\lambda_w \|\mu_j - \mu_i\|_2} \|\mathbf{z}_{k_i} - \mathbf{z}_{k_j}\|_2), \tag{15}$$

where hyperparameters $\lambda_w = 2000$ and $d = 20$. KNN means the k-nearest-neighbors algorithm. We also regularize the predicted coarse and fine deformations, constraining their values to remain close to zero. For further training details, please refer to the supplementary materials.

5. Experiments

5.1. Datasets and Metrics

Dataset and Baselines. We conduct extensive experiments on the Waymo Open Dataset [44] and KITTI Dataset [12] to comprehensively evaluate our method. For self-supervised setting, we select representative state-of-the-art methods including S3Gaussian [16] and DeSiRe-GS [38]. To compare with S3Gaussian, we use the dynamic32 (D32) split with 3 frontal cameras of Waymo dataset introduced by EmerNeRF [61] and benchmark S3Gaussian + EMD with vanilla 3DGS [19] and MARS [55] on scene reconstruction and novel view synthesis. To compare with DeSiRe-GS which follows PVG [5], we use the same subset in PVG, including 4 Waymo scenes and 3 KITTI scenes. We implement baselines including S-NeRF [57], StreetSurf [13], NSG [35] and SUDS [50] on scene reconstruction and novel view synthesis. For the supervised setting, we select state-of-the-art approaches including StreetGaussian [60] and OmniRe [6]. We select the same Waymo subset from StreetGaussian and OmniRe and conduct experiments on novel view synthesis. It should be noted that StreetGaussian only conducts experiments with one frontal camera in its published version. Thus, we further implement experiments on three frontal cameras for a side-by-side comparison.

Evaluation Metrics. Following previous protocols, we use PSNR and SSIM to evaluate the pixel-level reconstruction quality. Additionally, we compute LPIPS for perceptual quality assessment. We also report FPS to access inference speed in the comparison between DeSiRe-GS, which refers to frames per second. In addition, we employ the FID metric [15, 65] for novel trajectory synthesis evaluation, which quantifies differences in feature distribution between rendered novel trajectory images and original trajectory images.

5.2. Main Results

5.2.1. Self-supervised Performance

Tab. 1 presents the comparative results on the Waymo-D32 split, where no 3D bounding box annotations were used. Our method significantly outperforms previous self-supervised

Table 2. DeSiRe-GS comparison: results on the Waymo Open dataset and KITTI dataset for scene reconstruction and novel view synthesis.

Method	Waymo Open							KITTI						
	Scene Reconstruction			Novel View Synthesis			FPS	Scene Reconstruction			Novel View Synthesis			FPS
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
S-NeRF [57]	19.67	0.528	0.387	19.22	0.515	0.400	0.0014	19.23	0.664	0.193	18.71	0.606	0.352	0.0075
StreetSurf [13]	26.70	0.846	0.3717	23.78	0.822	0.401	0.097	24.14	0.819	0.257	22.48	0.763	0.304	0.37
3DGS [19]	27.99	0.866	0.293	25.08	0.822	0.319	63	21.02	0.811	0.202	19.54	0.776	0.224	125
NSG [34]	24.08	0.656	0.441	21.01	0.571	0.487	0.032	19.19	0.683	0.189	17.78	0.645	0.312	0.19
Mars [55]	21.81	0.681	0.430	20.69	0.636	0.453	0.030	27.96	0.900	0.185	24.31	0.845	0.160	0.31
SUDS [50]	28.83	0.805	0.317	25.36	0.783	0.384	0.008	28.83	0.917	0.147	26.07	0.797	0.131	0.29
EmerNeRF [61]	28.11	0.786	0.373	25.92	0.763	0.384	0.053	26.95	0.828	0.218	25.24	0.801	0.237	0.28
PVG [5]	32.46	0.910	0.229	28.11	0.849	0.279	50	32.83	0.937	0.070	27.43	0.896	0.114	59
DeSiRe-GS [38]	33.61	0.919	0.204	29.75	0.878	0.213	36	33.94	0.949	**0.040**	28.87	0.901	0.106	41
DeSiRe-GS + ours	**34.15**	**0.948**	**0.130**	**29.91**	**0.880**	**0.190**	32	**34.13**	**0.954**	**0.040**	**29.05**	**0.904**	**0.094**	32

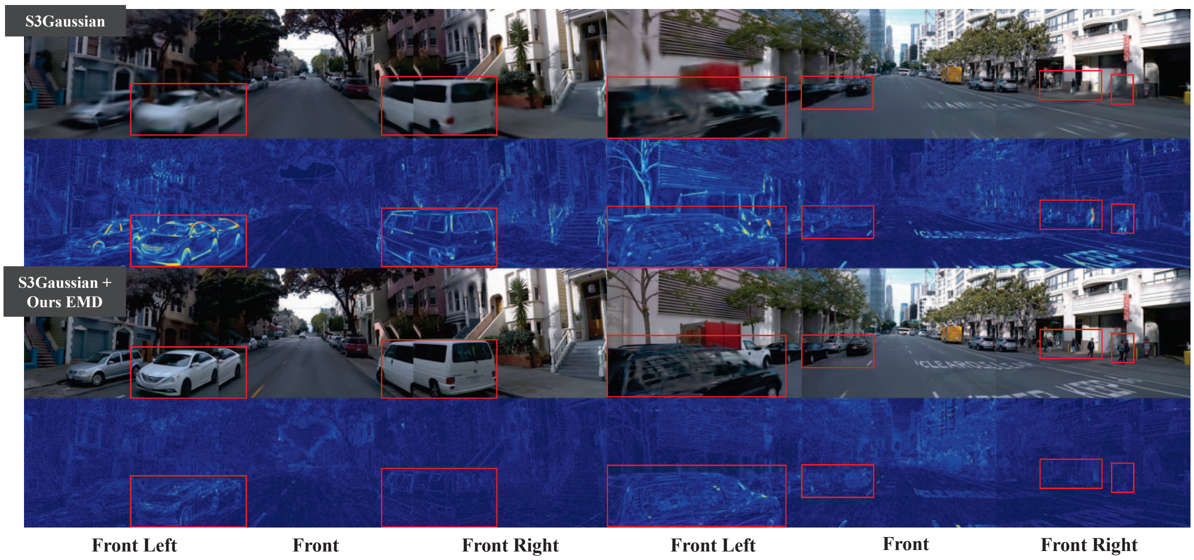

Figure 3. Visualization comparison on the self-supervised setting between S3Gaussian and S3Gaussian+ours EMD. We also visualize the error maps between the rendered images and ground truth to provide further insights. **Please refer to appendix for more visualization.**

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

approaches, achieving notable improvements in both full-scene and object-specific metrics. These results demonstrate the enhanced ability of our method to accurately model complex motions in street scenes without explicit 3D box annotations. S3Gaussian models the entire scene holistically but lacks the ability to capture the diverse motion patterns in dynamic street scenes, whereas EMD effectively compensates for this shortcoming. For novel view synthesis, our method continues to perform competitively, which reflects its robust generalization to previously unseen viewpoints.

Tab. 2 compares DeSiRe-GS with our proposed EMD. The results show that EMD achieves state-of-the-art performance across various experimental settings. DeSiRe-GS relies on dynamic masks to identify foreground objects, thereby overlooking motion patterns in street scenes. EMD boosts the rendering quality of DeSiRe-GS with only a slight reduction in inference speed. These quantitative results effectively demonstrate that EMD is well-suited for current self-supervised street Gaussian methods.

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. For more visualization videos, **please watch the webpage Sec.B in the supplementary materials.**

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.

Table 3. StreetGaussin comparison: novel view synthesis results on the Waymo Open dataset with one and three frontal cameras.

Methods	**Full Image**			**Vehicle**
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑
One camera setting				
3DGS [19]	29.64	0.918	0.117	21.25
NSG [34]	28.31	0.862	0.346	24.32
MARS [55]	29.75	0.886	0.264	26.54
EmerNeRF [61]	30.87	0.905	0.133	21.67
StreetGaussian [60]	34.61	0.938	0.079	30.23
StreetGaussian + Ours	**35.41**	**0.942**	**0.070**	**30.96**
Three camera setting				
StreetGaussian	29.70	0.858	0.149	26.72
StreetGaussian + Ours	**29.84**	**0.869**	**0.145**	**26.83**

Table 4. OmniRe comparison: novel view synthesis results on the Waymo Open dataset with three frontal cameras.

Methods	**Full Image**		**Human**		**Vehicle**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
EmerNeRF [61]	29.67	0.883	20.32	0.454	22.07	0.609
3DGS [19]	25.57	0.906	16.62	0.387	16.00	0.407
DeformGS [63]	27.72	0.922	17.30	0.426	18.91	0.530
PVG [5]	30.19	0.919	21.30	0.567	22.28	0.679
HUGS [67]	27.65	0.914	15.99	0.378	23.27	0.748
StreetGS [60]	28.54	0.928	16.55	0.393	26.71	0.846
OmniRe [6]	32.57	0.942	24.36	0.727	27.57	0.858
OmniRe + Ours	**33.89**	**0.958**	**25.97**	**0.742**	**27.82**	**0.859**

5.2.2. Supervised Performance

To demonstrate the flexibility of EMD, we further extend it to enhance motion modeling in supervised methods. Table 3 and Table 4 show novel view synthesis comparisons between StreetGaussian and OmniRe, respectively. While these supervised methods employ 3D box refinement to mitigate tracking errors, the results indicate that EMD serves as a valuable complement to this process. In the OmniRe comparison, EMD also benefits non-rigid objects such as humans, whose motion patterns differ significantly from those of vehicles. Due to the space limitation, please refer to the supplementary for visualization on the supervised setting.

Table 5. Novel trajectory synthesis on Waymo dataset.

Method	**FID↓**		
	0.5m	1.0m	1.5m
S3Gaussian	83.48	110.11	134.38
S3Gaussian + ours	**45.11**	**70.26**	**90.20**

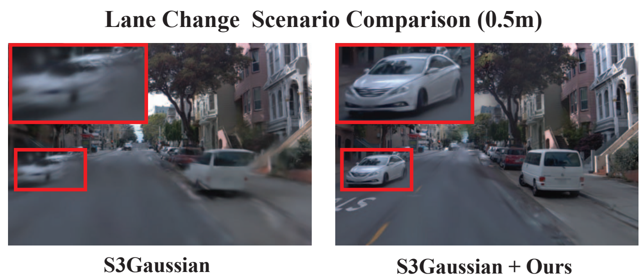

Figure 5. Visualization for novel trajectory synthesis (0.5m offset). **Please watch the webpage in the supplementary materials for more results.**

5.3. Novel Trajectory Synthesis

Previous novel view synthesis evaluations are limited to interpolated views along the original camera trajectory, which fail to assess the exact simulation performance of the model. Therefore, we compare the FID performance of self-supervised methods and our approach under novel trajectory synthesis. Specifically, we shift the original camera trajectory to the left and right by different offsets (0.5 m, 1.0 m, 1.5 m) and render images along these new trajectories. The results are presented in Tab. 5 and one visualization sample is shown in Fig. 5. These findings demonstrate that EMD enhances the accurate modeling of dynamic objects by learning diverse motion patterns, which benefits lane change scenarios. For novel trajectory synthesis videos, **please refer to the supplementary materials.**

5.4. Ablation Studies

To assess the contribution of each component in our framework, we perform comprehensive ablation studies, as shown in Tab. 6 and Fig. 6. 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

Figure 6. 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.

Table 6. Ablation study on the 32 dynamic scenes Waymo subset showing the impact of different components in our framework. All variants are evaluated in self-supervised scene reconstruction.

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

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 performance degradation. Specifically, excluding the coarse deformation component causes a significant 3.10 PSNR drop (29.40 vs. 32.50), suggesting that 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. Conclusion

In this paper, we present EMD, the first 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 complex motion patterns in real-world scenarios. Comprehensive experiments on the Waymo and KITTI datasets demonstrate that EMD achieves the state-of-the-art novel view synthesis quality for self-supervised frameworks. EMD can be extended into supervised street Gaussian splatting methods, which opens up new possibilities for high-quality driving scene reconstruction.

Acknowledgments. This work was supported by the National Science and Technology Major Project (No. 2022ZD0117800).

References

[1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, 2021. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2

[3] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 2

[4] 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

[5] 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. 2, 4, 5, 6, 7

[6] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024. 2, 5, 7

[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. 2

[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. 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] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5

[13] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 5, 6

[14] 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

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[16] 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, 4, 5

[17] 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

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

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 1, 2, 3, 5, 6, 7

[20] 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

[21] 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

[22] 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. 2

[23] 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. 2

[24] 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

[25] 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

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

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 5

[28] 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

[29] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis, 2023. 5

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

[31] 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

[32] 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

[33] 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

[34] 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, 6, 7

[35] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes, 2021. 5

[36] 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. 2

[37] 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. 2

[38] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes. *arXiv preprint arXiv:2411.11921*, 2024. 2, 4, 5, 6

[39] 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

[40] 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. 2

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

[42] 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

[43] 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. 2

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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. 2

[49] 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

[50] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6

[51] 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. 2

[52] Xiaobao Wei, Peng Chen, Ming Lu, Hui Chen, and Feng Tian. Graphavatar: Compact head avatars with gnn-generated 3d gaussians. *arXiv preprint arXiv:2412.13983*, 2024.

[53] Xiaobao Wei, Renrui Zhang, Jiarui Wu, Jiaming Liu, Ming Lu, Yandong Guo, and Shanghang Zhang. Nto3d: Neural target object 3d reconstruction with segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20352–20362, 2024. 2

[54] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 2, 3

[55] 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, 5, 6, 7

[56] 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. 2

[57] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023. 5, 6

[58] 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

[59] 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. 2

[60] 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, 5, 7

[61] 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. 2, 5, 6, 7

[62] 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

[63] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction, 2023. 3, 7

[64] 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

[65] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arXiv preprint arXiv:2410.13571*, 2024. 5

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

[67] 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. 2, 7

[68] 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