Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 3, No. 3, 2025

Large-scale Dynamic Scene Reconstruction Based on Neural Fields

Yinan Qi

School of Computer, North China Electric Power University, Baoding 071003, China

Abstract: Large-scale dynamic scene reconstruction methods based on spatiotemporal field models demonstrate significant potential in autonomous driving applications. However, existing neural radiance field (NeRF) and 3D Gaussian splatting (3DGS) techniques remain constrained by their dynamic element modeling capabilities and computational efficiency, failing to effectively address complex reconstruction tasks involving intertwined static and dynamic regions in driving scenarios. To address these challenges, this study proposes a novel framework integrating spatiotemporal attention mechanisms with sparse encoding strategies. The method employs a spatiotemporal attention module that captures dynamic motion patterns through self-supervised inter-frame prediction, addressing spatiotemporal inconsistencies caused by non-rigid deformations. Simultaneously, a KL divergence-guided hierarchical sparse encoding strategy achieves efficient multi-scale scene feature representation while preserving reconstruction accuracy. Furthermore, a mean-variance decoupled stochastic sampling mechanism enhances modeling robustness in dynamic regions. Experimental results demonstrate substantial improvements in reconstruction quality compared to state-of-the-art large-scale dynamic scene reconstruction methods, ultimately enabling more photorealistic 3D reconstruction outcomes."

Keywords: 3D Gaussian Splatting; Dynamic scene; Neural rendering; Large-scale.

1. Introduction

In recent years, with the rapid development of technologies such as autonomous driving and augmented reality, 3D reconstruction of large-scale dynamic scenes has become one of the core challenges in computer vision and graphics. Among them, autonomous driving scenarios have made significant progress in recent years and various techniques have been developed at various stages of their pipeline, including perception [1-3], prediction [4-6], and planning [7-9]. With the advent of end-to-end automated driving that outputs control signals [10-12] directly from sensor inputs, open-loop evaluation of automated driving systems is no longer valid and therefore requires urgent improvement [13]. As a promising solution, real-world closed-loop evaluation requires a controlled view of the sensor inputs, which drives the development of high-quality scene reconstruction methods [14].

Neural Radiation Field (NeRF) [15] has recently emerged as a promising neural reconstruction method, and several studies [16, 17] have extended NeRF to large-scale, unbounded static scenes, but NeRF-based methods are computationally intensive, requiring densely overlapping views and consistent illumination. These limit their ability to build driving scenes at high speed for outward multi-camera setups. In addition network capacity limitations make them more challenging in the case of modelling long, dynamic scenes with multiple objects, leading to visual artefacts and blurring.

In contrast to NeRF, 3D Gaussian Splatting (3DGS) [18] reconstruction methods have attracted much attention due to their potential for efficient representation of complex geometries and materials, which describe the scene radiation field through discretised Gaussian primitives, and have demonstrated rendering quality superior to that of traditional point clouds and Neural Radiation Fields (NeRF) in static scenes. However, the initial 3DGS still encountered significant challenges in modelling large-scale dynamic scenes due to fixed Gaussian functions and limited representation capabilities. When confronted with complex scenes where dynamic objects are intertwined with static backgrounds, the existing methods face a double dilemma in terms of reconstruction accuracy and computational efficiency, which severely limits their application in real-time interactive systems.

To address the above challenges, this paper proposes a novel reconstruction framework suitable for 3D reconstruction of large-scale dynamic scenes. The framework, in order to improve the reconstruction accuracy as well as the computational efficiency of large-scale dynamic scenes, uses a spatio-temporal field model that incorporates a spatio-temporal attention mechanism and hierarchical sparse coding, which optimises the reconstruction accuracy by introducing a self-supervised learning technique, and at the same time employs computational optimisation methods to improve the reconstruction efficiency. Specifically, the model is designed with a spatio-temporal attention module with inter-frame prediction capability to achieve unsupervised dynamic law learning. Meanwhile, a sparse coding strategy is adopted to improve the computational efficiency while preventing overfitting. The model also implements random sampling by mean-variance separation design, which enhances the generalisation ability of the model. Experiments on the KITTI dataset [19] demonstrate the effectiveness of the method.

2. Method

2.1. Overview of the Approach

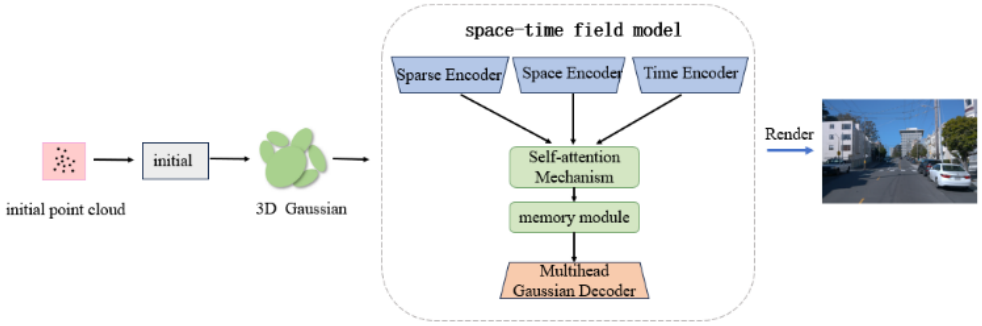

Figure 1. A neural field-based pipeline for large-scale dynamic scene reconstruction

The research goal of this paper is to learn a spatio-temporal representation of a dynamic street scene from a series of images captured by a moving vehicle. For this purpose this chapter proposes a new scene representation with a reconstruction pipeline as shown in Fig. 1. The method takes point cloud data as input, initialises the point cloud to a Gaussian representation to construct a Gaussian representation system, and subsequently feeds the Gaussian into the spatio-temporal field model to achieve dynamic feature extraction. The spatio-temporal field model contains multiple encoders and decoders, and the model uses spatial encoders and sparse encoders to extract spatial and sparse features in the spatial dimension, and efficiently extracts geometric features. In the time dimension, a two-way recursive coding module is designed to achieve cross-frame feature propagation through a dynamic memory pool, and a spatio-temporal attention gating unit is introduced to automatically adjust the weight of historical information. The spatial information is coupled with the temporal motion trajectory in the hidden space through differentiable voxel jump connection, and the Gaussian parameter evolution path is dynamically corrected by using the residual prediction network to construct the feature fusion mechanism with self-evolutionary capability. Subsequently, the spatio-temporal features are decoded by the multi-head Gaussian decoder and the Gaussian parameters are output. Finally, the obtained Gaussian parameters are fed into the rasterisation pipeline [18] to perform microscopic rendering of the Gaussian parameters to obtain a high-quality rendered view.

2.2. 3D Gaussians Splatting

As shown in Fig. 1, our scene representation consists of a 3D Gaussian and a spatio-temporal field model. the 3D Gaussian is represented by a covariance matrix Σ and a position vector x. The 3D Gaussian is represented by a covariance matrix Σ and a position vector x. Each covariance matrix is further decomposed into a scaling matrix R and a rotation matrix S:

$$\sum = RSS^T R^T \tag{1}$$

In addition to the position and covariance matrix, each Gaussian is assigned an opacity value $\alpha \in \mathbb{R}$ defined by the spherical harmonic function (SH) and a colour $C \in R^{3(k+1)^2}$, where k denotes the degree of the SH function. The spatio-temporal field model takes as input the position of each Gaussian and the current time step t to generate spatio-temporal features f. Decoding these features, the spatio-temporal field model predicts the displacement of each point with respect to the canonical space, and then we project the Gaussian into 2D [20] using a microscopic 3D Gaussian splatting renderer following the method in literature [21], with the covariance matrix $\sum'$ in the camera coordinate system denoted as:

$$\sum' = JW\sum W^T J^T \tag{2}$$

Where J is the Jacobi matrix of the perspective projection and W is the observed change matrix. The colour of each pixel is arrived at by computing the mixing value of N ordered points:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1-\alpha_i) \tag{3}$$

Where α_i and c_i denote the opacity and colour of a point, calculated by the opacity and SH colour coefficients of each point that can be optimised.

2.3. Spatial Feature Extraction Module

The spatial feature extraction module is a fundamental component of spatio-temporal field modelling, and its core objective is to transform the raw geometric data into potential features with strong representational capabilities. The module employs a multimodal feature fusion strategy to achieve a compact representation of the scene through the joint encoding of geometric coordinates (x, y, z), transparency α, covariance matrix $\sum$, and mean value features μ. Given an input point cloud $P = \{p_i\}_{i=1}^N$, where each point $pi = (x_i, \alpha_i, \sum_i, \mu_i)$, the feature encoding process can be formalised as the following equation:

$$h_i^{(0)} = \varphi_{concat}(x_i \oplus \alpha_i \oplus vec(\sum_i) \oplus \mu_i) \tag{4}$$

Where φ_{concat} denotes the linear projection after feature splicing, $\oplus$ is the splicing operation. Subsequently, a nonlinear transformation is performed by dynamically constructing a multilayer perceptron:

$$h_i^{(l)} = BN(W^{(l)} h_i^{l-1} + b^{(l)}), l = 1, \ldots L \tag{5}$$

Where BN denotes batch normalisation layer, $W^{(l)}$ and $b^{(l)}$ are learnable parameters.

Meanwhile, the model in this paper adopts a variational self-encoder framework as a sparse encoder, which is used to generate low-dimensional sparse representations, and is

centred on synergistically enhancing the system performance through probabilistic feature compression and structured regularisation. The encoder adopts a two-channel parallel structure, with the upper channel extracting the spatial features of the input point cloud through a 3D convolutional network, and outputting the mean value of the potential space. The lower channel embeds a null convolutional layer to capture the multi-scale contextual information and generate the variance. The two are constrained by the KL dispersion to form the latent variable z, whose variational posterior distribution expression is shown in Eqs. 6.

$$q_\varphi(z \mid X) = N(z; \mu_\varphi(X), diag(\sigma_\varphi^2(X))) \tag{6}$$

Where X is the input data, μ_φ and σ_φ are generated by the mean and variance layers of the sparse encoder.

2.4. Timing Dynamic Modelling Module

The temporal dynamic modelling module in this paper achieves accurate modelling of spatio-temporal continuous fields through the combination of hierarchical memory networks and attention mechanisms. Given a time series $\{t_k\}_{k=1}^{T}$, absolute temporal information is injected into the model using position encoding:

$$r(t_k) = [sin(2^0 \pi t_k), cos(2^0 \pi t_k), \ldots sin(2^{(m-1)} \pi t_k), cos(2^{(m-1)} \pi t_k)] \tag{7}$$

To further enhance the spatio-temporal feature interaction, this method introduces a multi-head attention mechanism to calculate the cross-modal association weights. The spatial encoding and temporal encoding are integrated to form the spatio-temporal feature f through the multi-head attention mechanism, in which the spatial encoding and temporal encoding are linearly projected to generate the query matrix Q and the key matrix K, respectively, and then computed through the scaled dot product and normalised by Softmax, and the final spatio-temporal fusion feature is generated through the weighted summation with the attention weights. This module enables spatio-temporal feature interaction, establishes multimodal associations, and enhances the expressive ability and model interpretability by stacking multiple modules.

The fused spatio-temporal features are subsequently fed into the memory module. In order to capture long-range temporal dependencies, the module is designed with a bidirectional multi-stage gating loop unit, which is used to capture the before and after temporal dependencies. The fusion weights of historical information and current features are dynamically regulated by reset gates and update gates, and the long- and short-range dependencies are adaptively balanced by an implicit gating mechanism, and the final spatiotemporal feature F_f is outputted.The temporal dynamic modelling module achieves efficient temporal modelling of dynamic Gaussians through the close integration of temporal encoding, attention mechanism and memory module, and improves computational efficiency while ensuring model expressiveness.

2.5. Parameter Generation Module

The parameter generation module of this method is inspired by the $S^3Gaussian$ [22] and uses a separate MLP header to decode the input data. Since most autopilot scenarios involve rigid motions, this paper focuses on the deformation of the Gaussian position. Considering that the appearance of the scene changes with its position and time due to factors such as lighting, this paper uses the SH coefficient header to simulate the deformation of the dynamic appearance. The specific model is shown in Fig. 2.

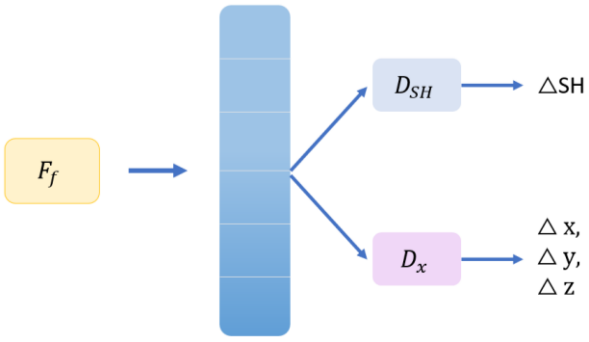

Figure 2. Multihead Gaussian Encoder Schematic

3. Experiment

3.1. Dataset

KITTI dataset [19]: this dataset is one of the most commonly used international datasets for evaluating computer vision algorithms in autonomous driving scenarios, and its acquisition vehicle setup is shown in Fig. 3, where the vehicle is fitted with two greyscale cameras, two colour cameras, a Velodyne 64-line 3D Lidar, four optics lenses, and a GPS navigation system, which is marked with a red marker in Figure 4-3. Red markers are used in Fig. 3. The KITTI dataset contains real image data acquired from urban, rural and motorway scenes. Each image contains up to 30 pedestrians and 15 vehicles, with varying degrees of truncation and occlusion. The dataset provides 14,999 images and corresponding point clouds, of which 7,481 sets are used for training and 7,581 sets are used for testing, and are labelled for the three types of objects in the scene: cars, pedestrians and bicycles, totalling 80,256 labelled objects. In contrast, there is a serious imbalance in many public datasets of LiDAR data, such as nuScenes [23] and nuPlan [24], so we choose three sub-datasets in the KITTI dataset for training, namely the pedestrian dataset, the road dataset, and the city dataset.

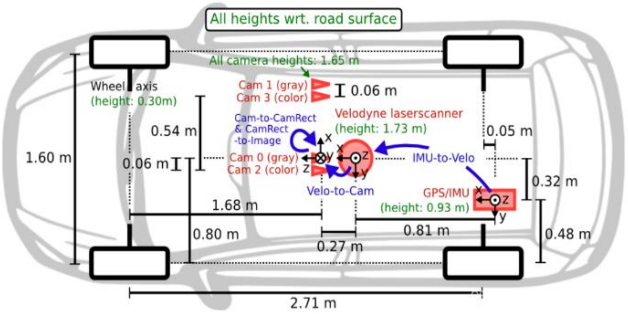

Figure 3. KITTI Dataset [19] Collection Vehicle

3.2. Model Implementation Details

The model in this paper uses the Adam optimiser for a total of 40,000 iterations, and its learning rate follows the learning rate configuration of 3D Gaussian sputtering [18]. In addition, this model implements a staged training strategy, where the 3D Gaussian representations are optimised independently for the first 3000 iterations, and the spatio - temporal field model is jointly trained after its spatial location and geometry have reached a relatively stable state. In the probabilistic generation module, each decoder of the multi - head Gaussian decoder is a small MLP, and its learning rate is set to 1.6×10^{-3}, and the other hyperparameters are kept consistent with the 3D Gaussian sputtering [18].

3.3. Large-scale Dynamic Scene Compositing Results

State-of-the-art methods are used to evaluate this paper's approach on the KITTI dataset, including NeRF-based models and 3D-GS-based models. MARS [25] is a modular NeRF-based simulator that uses a 2D bounding box to train NeRF on static and dynamic scenes, respectively. NSG [26] learns latent codes to model moving objects with a shared decoder. StreetGaussian [27] is a recent Gaussian-based method that introduces time in the SH coefficients to achieve SOTA performance. Therefore in order to evaluate the effectiveness of the methods in this chapter, a comparison with these three state-of-the-art methods has been chosen.

The results on the KITTI dataset show that the method in this chapter consistently outperforms other methods in new view synthesis, the quantitative results of which are shown in Table 1. Using PSNR, SSIM and LPIPS as the evaluation metrics for rendering quality, the data in the table demonstrates the excellent performance of the proposed model in this chapter on the public dataset KITTI, where the synthesis results are superior to those of the existing comparative methods, indicating the superior performance of the proposed model in modelling large-scale dynamic scenes. Meanwhile, the model in this paper also performs well in static scene representation, verifying the generality and practicality of the proposed method.

In addition, this paper also conducts qualitative experiments on the KITTI dataset, the results of which are shown in Fig. 4. As can be seen from the figure, the method in this paper shows significant advantages in several key dimensions of complex large-scale dynamic scenes, including walking pedestrians, vehicle licence plates and moving vehicles, all of which show excellent rendering quality. The above analysis further confirms that the method proposed in this chapter can achieve excellent rendering results in modelling large-scale dynamic scenes from the perspective of visual effect, which provides important support for the accurate modelling of large-scale dynamic scenes.

Table 1. KITTI dataset quantitative comparison experiment

Methods	PSNR↑	SSIM↑	LPIPS↑
NSG	21.79	0.666	0.293
MARS	26.32	0.853	0.130
StreetGaussian	27.41	0.882	0.057
Ours	29.11	0.925	0.042

Figure 4. The results of the qualitative comparison experiment on the KITTI dataset

4. Summary

In this paper, in response to the dual challenges of insufficient reconstruction accuracy and high computational load for dynamic targets in complex traffic scenes in large-scale 3D reconstruction of dynamic scenes, a new dynamic 3D reconstruction method based on joint spatio-temporal modelling is proposed in this chapter. This chapter first introduces the basic model design of this paper, which is mainly divided into three parts: data initialisation (Gaussian representation), spatio-temporal field model and rendering module, and then focuses on the spatio-temporal field model. The spatio-temporal field model is designed with a spatio-temporal fusion mechanism with dynamic sensing capability, which provides a probabilistic feature representation through a sparse coding mechanism, while the self-attention mechanism module implements the surface deformation of the moving target in the hidden space through an attention-guided feature propagation network and updates the Gaussian parameters in real time through a residual correction module. The Gaussian parameters obtained by the Gaussian decoder decoding the spatio-temporal features are subsequently fed into an efficient rasterised rendering pipeline for rendering, thus achieving high quality view rendering. The method in this chapter confirms the effectiveness of the method in complex large-scale dynamic scenes by conducting experiments on the public dataset KITTI dataset. However, the method still has shortcomings in some aspects, especially in dealing with certain challenges and limitations in reconstructing scenes with high-speed moving objects or facing severe occlusion situations. These difficulties will be the focus of subsequent research in anticipation of further improving the performance and applicability of the method.

References

[1] Zhang Y, Zhu Z, Zheng W, et al. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving [J]. arXiv preprint arXiv: 2205.09743, 2022.

[2] Huang Y, Zheng W, Zhang Y, et al. Tri-perspective view for vision-based 3d semantic occupancy prediction [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 9223-9232.

[3] Wei Y, Zhao L, Zheng W, et al. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 21729-21740.

[4] Hu A, Murez Z, Mohan N, et al. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 15273-15282.

[5] Gu J, Hu C, Zhang T, et al. Vip3d: End-to-end visual trajectory prediction via 3d agent queries [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5496-5506.

[6] Liang M, Yang B, Zeng W, et al. Pnpnet: End-to-end perception and prediction with tracking in the loop [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11553-11562.

[7] Dauner D, Hallgarten M, Geiger A, et al. Parting with misconceptions about learning-based vehicle motion planning [C]//Conference on Robot Learning. PMLR, 2023: 1268-1281.

[8] Cheng J, Chen Y, Zhang Q, et al. Real-time trajectory planning for autonomous driving with gaussian process and incremental refinement [C]//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 8999-9005.

[9] Cheng J, Mei X, Liu M. Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 8679-8689.

[10] Hu S, Chen L, Wu P, et al. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning [C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 533-549.

[11] Hu Y, Yang J, Chen L, et al. Planning-oriented autonomous driving [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 17853-17862.

[12] Jiang B, Chen S, Xu Q, et al. Vad: Vectorized scene representation for efficient autonomous driving [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 8340-8350.

[13] Li Z, Yu Z, Lan S, et al. Is ego status all you need for open-loop end-to-end autonomous driving [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14864-14873.

[14] Turki H, Zhang J Y, Ferroni F, et al. Suds: Scalable urban dynamic scenes [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12375-12385.

[15] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis [J]. Communications of the ACM, 2021, 65(1): 99-106.

[16] Wang Z, Shen T, Gao J, et al. Neural fields meet explicit geometric representations for inverse rendering of urban scenes [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8370-8380.

[17] Zhenxing M I, Xu D. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields [C]//The Eleventh International Conference on Learning Representations. 2022.

[18] Kerbl B, Kopanas G, Leimkühler T, et al. 3d gaussian splatting for real-time radiance field rendering [J]. ACM Trans. Graph., 2023, 42(4): 139:1-139:14.

[19] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset [J]. The international journal of robotics research, 2013, 32(11): 1231-1237.

[20] Zwicker M, Pfister H, Van Baar J, et al. Surface splatting [C]//Proceedings of the 28th annual conference on Computer graphics and interactive techniques. 2001: 371-378.

[21] Yifan W, Serena F, Wu S, et al. Differentiable surface splatting for point-based geometry processing [J]. ACM Transactions On Graphics (TOG), 2019, 38(6): 1-14.

[22] Huang N, Wei X, Zheng W, et al. s^3gaussian: Self-Supervised Street Gaussians for Autonomous Driving [J]. arXiv preprint arXiv:2405.20323, 2024.

[23] Caesar H, Bankiti V, Lang A H, et al. nuscenes: A multimodal dataset for autonomous driving [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11621-11631.

[24] Caesar H, Kabzan J, Tan K S, et al. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles [J]. arXiv preprint arXiv:2106.11810, 2021.

[25] Wu Z, Liu T, Luo L, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving [C]//CAAI International Conference on Artificial Intelligence. Singapore: Springer Nature Singapore, 2023: 3-15.

[26] Ost J, Mannan F, Thuerey N, et al. Neural scene graphs for dynamic scenes [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2856-2865.

[27] Yan Y, Lin H, Zhou C, et al. Street gaussians: Modeling dynamic urban scenes with gaussian splatting [C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 156-173.