The Development Of 3D Reconstruction in The Field of Autonomous Driving Based on NeRF Methods

Han Wu *

School of automotive study, Tongji University, Shanghai, China

* Corresponding Author Email: 2054101@tongji.edu.cn

Abstract. The process of 3D reconstruction is essential for the functionality of autonomous vehicles.3D reconstruction technology can take the 2D visual image information and other data collected by sensors and reconstruct a three-dimensional scene of the surrounding environment, providing autonomous vehicles with more precise environmental perception. As one of the most capable and influential 3D reconstruction methods in recent years, NeRF has become a focal point of interest and extensive investigation within the realm of autonomous driving technology.This paper concentrate on the application of the NeRF method in 3D reconstruction within the field of autonomous driving. Aiming at the NeRF with high resolution and high detail reconstruction capabilities, it first introduces the implementation principle of the original NeRF method, and then introduces various improved NeRF methods based on this approach. Subsequently, it presents the features of the improved NeRF methods as well as their shortcomings and difficulties in 3D reconstruction for autonomous driving. Finally, the future development trend of the NeRF method is discussed and predicted.

Keywords: Neural radiance field; 3D reconstruction; autonomous driving; dynamic scenes reconstruction; large-scale scene reconstruction.

1. Introduction

Recently, the field of autonomous driving (AD) has made tremendous progress. With the development of computer vision technology, AD has evolved from initially relying heavily on high-precision maps to now being able to read the surrounding environment in real time through cameras combined with various models and algorithms. Aiming to enhance the autonomous driving system's perception and understanding of the surrounding environment, the NeRF models has quickly gained attention due to its ability to comprehend and process complex scenes.

NeRF, which stands for "Neural Radiance Fields," is a 3D reconstruction approach firstly proposed by Mildenhall et al. in 2020. It utilizes a Multilayer Perceptron (MLP) deep learning network to acquire high-resolution and high-detail reconstruction of 3D scenes, and its reconstruction effects are significantly better than those of previously proposed 3D reconstruction methods. Since its introduction, this method has garnered widespread attention in academia and has been deeply concerned and applied in various fields [1].

NeRF models have successfully found applications in various aspects of autonomous driving, including perception,3D reconstruction, simultaneously localization and mapping, and simulation (SLAM). In every aspect, there has been a lot of in-depth research. Among these, its application in 3D reconstruction has attracted high attention.

However, origin NeRF introduced by Mildenhall et al is unable to fully accomplish 3D reconstruction tasks in the field of autonomous driving. First, the scenario of autonomous driving includes not only static scenes but also a large number of dynamic scenes which origin NeRF is found struggle to deal with. Second, the reconstruction of the scenario of autonomous driving require high real-time performance and high efficiency, while origin NeRF need occupy large number of time and computing power to reconstruct a scene. Third, the scenario of autonomous scenario require large-scale, even open-world level scene reconstruction, but origin NeRF can only reconstruct a small-scale scene. Last, origin NeRF perform badly in reconstructing non-rigid objects, but there are lots of non-rigid objects such as pedestrian in the autonomous driving scenes.

Based on the problems the origin NeRF have, researchers try their best to improve NeRF from every aspect. In this paper, the improved NeRF methods are summarized, including introductions of various of improved NeRF methods, the features of improved NeRF methods and the future trend of improved NeRF methods. Through these paper, a conclusion of the development direction of NeRF can be provided and a prediction and guidance of the development of NeRF can be put forward.

2. Neural Radiance Field

Neural Radian Field,known as NeRF, is an cutting-edge method of view synthesis technique, first introduced by Mildenhall et.al in 2020. It effectively generates new and detailed perspectives of intricate environments by refining a continuous 3D model, guided by a multitude of planar photographs, Fig,1 shows the flowchart of the NeRF procedure [2].

NeRF introduces a 5D vector value function to represent scenes, including the coordinates location of each point within the scene $X = (x, y, z)$, and the coordinates of viewing angels (azimuthal angel θ and polor angle φ.Each 5D vector is fed into a Multi-Layer Perceptron (MLP) network, which in turn produces a singular volume density and an RGB color that varies with the view. This can be articulated as follows:

$$F(X, \theta, \varphi) \rightarrow (c, \sigma) \tag{1}$$

The volume density output and the RGB color that shifts with the viewpoint can be interpreted as the measure of radiance that traverses each point within the scene and the color that is emitted in the direction of a particular viewpoint (θ, φ).To ensure the multiview representation is consistent, the MLP process the density of the volume σ only using the position coordinates of the scene x, while processing the RGB c uses a combination of location coordinates and view directions coordinates. Optimizing the weight of MLP networks enables NeRF generate mapping between 5D vector and its corresponding density of the volume and RGB color.

The Multi-layer Perceptron (MLP) plays an essential role in the NeRF framework. At the outset, the MLP receives spatial coordinates and passes them through a sequence of eight fully-connected layers. Each layer is endowed with ReLU activation functions and contains 256 feature channels, culminating in the generation of a volume density denoted as σ and a 256-dimensional feature vector. Subsequently, the viewing direction of the camera ray is combined with the feature output from the final layer, which then undergoes an additional fully-connected layer to ultimately yield the RGB color values.

After producing the volume density and color, NeRF synthetic the image through volume rendering technique. Through the process of volume rendering, NeRF generates the anticipated color $C(r)$ for a camera ray $r(t) = o + td$. This involves using the camera's spatial coordinates to represent its position and deducing the orientation of the viewing direction.

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)dt, \text{where } T(t) = \exp(-\int_{t_n}^{t} \sigma(r(s))ds) \tag{2}$$

In this formulation, σ(r(t)) and c(r(t), d) signify the volume density and the RGB at a specific point along the trajectory of the ray of camera that is aimed in the direction d, respectively. The variable dt symbolizes the small step in distance that the ray covers during each segment of the integration process. The function T(t) represents the cumulative light transmittance experienced by the ray as it traverses its path, starting from the origin t_n to and extending to the point t , which can be interpreted as the likelihood that the ray will travel from t_n to t without encountering any obstructions.

NeRF employs a stratified sampling technique, dividing the interval $[t_n, t_f]$ into N uniformly distributed segments. Within each segment, it randomly selects one sample. This method addresses the resolution constraint issue that is commonly faced by traditional approaches. By doing so, NeRF ensures a more consistent sampling across the entire range, which can improve the quality and

resolution of the 3D reconstruction. Due to the characteristics of MLP, NeRF enables represents a continuous scene by a discrete set of samples. NeRF utilizes a series of samples to estimate the color C(r) as follows:

$$\hat{C}(r)=\sum_{i=1}^{N}T_i(1-\exp(-\sigma_i\delta_i))c_i,\ \text{where}\ T_i=\exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j) \tag{3}$$

where δi = ti+1 − ti is the distance between adjacent samples.

To attain advanced quality of 3D reconstruction scenes, NeRF typically employs various of supplementary techniques to refine its processes and accurately represent intricate, high-resolution environments. These techniques encompass positional encoding and tiered volume sampling.

Positional encoding augments NeRF's capacity to process data exhibiting high-frequency fluctuations by embedding inputs into an expanded dimensional space with the aid of oscillatory functions, thus facilitating a more precise fit before the data is processed by the network,this encoding is computed as follows.

$$\gamma(p)=(\sin(2^0\pi p),\cos(2^0\pi p),.....,\sin(2^{L-1}\pi p),\cos(2^{L-1}\pi p)) \tag{4}$$

The function operates on each value along the three axes individually in X(lie in $[-1,1]$ after normalized) and the viewing direction d.

Besides,NeRF takes a tiered volume sampling strategy to increase rendering efficiency.NeRF uses two networks named “coarse” and “fine”,then divides the samples proportionally and gives different expected final render effects.

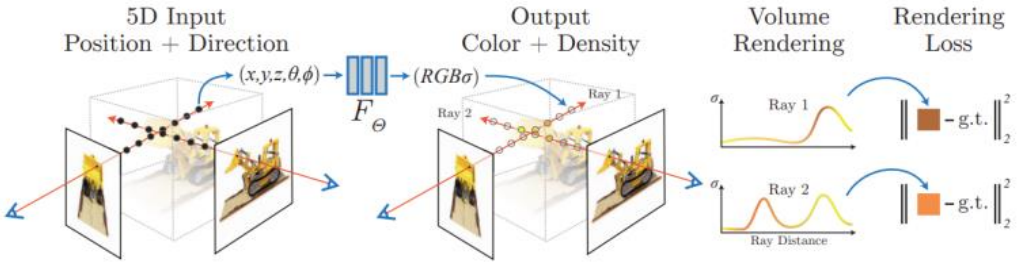


Fig. 1 A conceptual structure for depicting scenes through neural radiance fields and a rendering process that incorporates differentiable techniques [2].

3. The Application of NeRF for 3D Reconstruction in the Field of Autonomous Driving

3D reconstruction plays an significant role in autonomous driving,which enables construct a well-round surrounding models,reconstructs a high-resolution scenes and has ability to comprehend complex transportation scenarios,thereby helping autonomous systems to make more reasonable planning and controlling decisions.

Although the original NeRF has achieved high-resolution and detailed 3D reconstruction,there is still a long way to go before it can be applied in the field of AD.The original NeRF can only achieve good reconstruction effects in static scenes, and its reconstruction requires a large set of 2D images. Moreover, the time cost for training and testing is very high,Mildenhall et.al took at least 12 hours to train per scene.All these problems limit the application of NeRF in the field of autonomous driving.To solve these problem,lots of researchers have promoted optimizing methods based on NeRF and several applications have been adapted to solve practical problems.

NeRF is essential for 3D dynamic scene reconstruction in autonomous driving (AD). The Neural Scene Graphs (NSG) method is the first to offer dynamic reconstruction by dividing scenes into static and dynamic components. It uses a directed acyclic graph to illustrate objects and augmented neural radiance fields for representation. NSG can learn from a scene's video, allowing it to render new

views of dynamic scenes. This method excels at capturing highly dynamic scenes and addresses the original NeRF's reliance on static scene consistency,see Fig.2 [3].

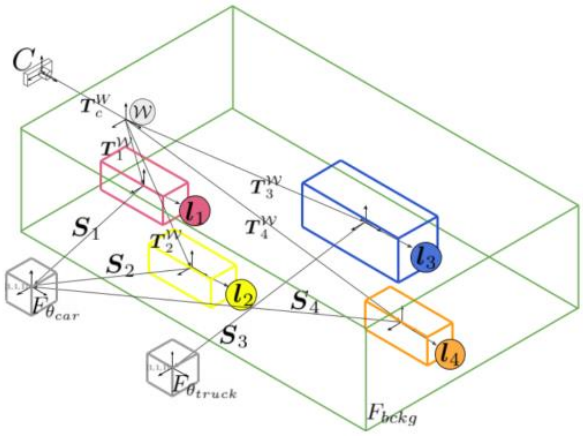


Fig. 2 Neural scene graph in isometric view [3].

Reconstruction of large-scale driving scene is also a focal point of interest in the field of AD.

Block-NeRF introduces a strategy for reconstructing large scenes by breaking them down into smaller, manageable parts. As what Fig.3 shows,each part is trained by a separate NLP network, and then they are merged seamlessly, aligning the appearance of adjacent sections. This method improves NeRF by adding appearance embedding to reduce reconstruction errors and refining pose estimation. Additionally, it includes exposure adjustment, allowing for changes in exposure levels during the inference stage [4].

SUDS method has implemented a "open-world" dynamic large-scale 3D reconstruction.SUDS achieves its goal by dividing the scene into three distinct hash grid data structures, thereby enhancing the efficiency of encoding radiance fields for static, dynamic, and distant elements. Additionally, SUDS employs unlabeled target signals to integrate dynamic scene components into the static backdrop, delineating individual objects and their movements [1].

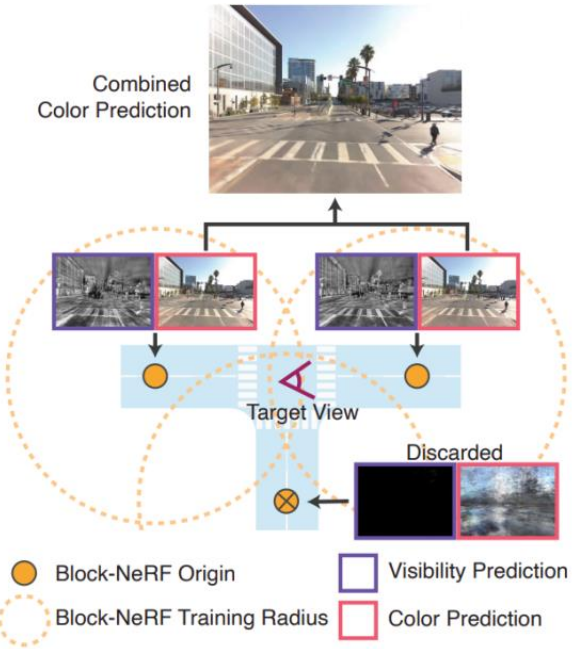


Fig. 3 The scene is split into multiple Block-NeRFs [4]

AD requires high-resolution,high-precision and detailed 3D reconstruction.To realize this goal based on NeRF,multiple improved NeRF methods are proposed to achieve this aim from different aspects.

Urban Radiance Fields introduces an innovative approach to 3D reconstruction and the synthesis of new viewpoints from datasets acquired by scanning systems typically employed for global mapping in urban outdoor settings, encompassing a sequence of RGB photographs and LiDAR scans. It leverages asynchronously captured LIDAR data to address the issue of exposure difference between images, predicted image segmentations to supervise the density of rays directed toward the sky. It is proved to be able to synthesize higher quality novel views in urban view fields [5].

FEGR provides a novel inverse rendering framework for urban scenes, reconstructing the scene geometry, various of materials, and lighting. FEGR Employ a neural field to represent the primary rays, while an explicit mesh, reconstructed from the neural field, is used to model secondary rays that contribute to advanced lighting effects, including cast shadows. This method shows a excellent results on scene relighting and objects insertion [1].

Pedestrians are an essential component of the traffic environment. Unlike autos which can be considered as rigid body, pedestrians are non-rigid dynamic objects which current improved NeRF find difficult to handle with. humanNeRF utilize a feature that aggregates pixel alignment from multiple viewpoints,combined with a pose-embedded non-rigid deformation field,to address dynamic motion effective. Apart from that, HumanNeRF further incorporates in-hour fine-tuning specific to the scene and an appearance blending module. This approach leverages the advantages of both neural volumetric rendering and neural texture blending, enhancing the overall rendering quality [6].

Instant-NVR introduces a novel approach for rendering human subjects, facilitating the rapid creation of radiance fields within seconds and opening up the potential for real-time radiance field modeling. Central to Instant-NVR's concept is the integration of conventional non-rigid volume tracking with instantaneous radiance field technology. The front-end tracking module is tasked with providing real-time motion tracking for both the actors and the objects involved. Simultaneously, the rendering backend is dedicated to the reconstruction of the radiance fields for the interactive scene. Together, these processes enable the immediate synthesis of new views. Importantly, the quality of these synthesized views is photorealistic, providing a high level of visual fidelity.

Improving the robustness of 3D reconstruction and reducing both the time and 2D image costs are critical factors in determining whether it can be practically applied in the field of AD.

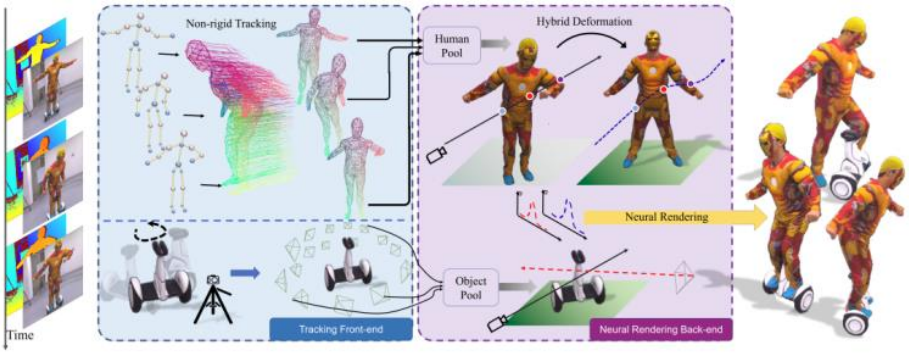


Fig. 4 The stages of Human-NeRFs[7]

In the Human-NeRF framework, the motion tracking component at the front end captures human and object motions, see Fig.4.While the rendering back end cgenerates the radiance fields for human-object interactions on-the-fly, facilitating instant novel view synthesis with photorealism [7].

Neural Point Light Fields introduces a representation that is implicitly embedded within a sparse set of points and necessitates the computation of radiance for just one point along each ray. This approach has been demonstrated to successfully forecast video content along trajectories that were previously unseen, a capability that is made possible by explicitly modeling the scene dynamics [1].

MapNeRF is a new approach that overcomes NeRF's limitation in generating extrapolated views. It enables the creation of driving views that are not on the original trajectory by using map data as a guide during training. It considers the uncertainty in the map data and the ground surface as variable factors. MapNeRF includes uncertainty as a parameter and uses an uncertainty tempering scheme to control the training of the neural radiance field [1].

EmerNeRF provides a powerful method without relying on ground truth 3D box achieving a high quality and self-sufficiently dynamic object reconstruction. It consist of two core procedures: NeRF classifies scenes into static field and dynamic field, emerging from self-supervision and allowing model to learn from general data sources. Besides,it characterizes an induced flow field derived from the dynamic field and employs this flow field to further consolidate multi-frame features, thereby enhancing the rendering accuracy of dynamic objects [8].

UC-NeRF is an improved version of NeRF that tackles the under-calibration issues in multi-camera setups. It employs a three-step process: Layer-based color correction is implemented to rectify discrepancies in coloration, and a virtual warping method is applied to craft a spectrum of perspectives that maintain uniform color fidelity, and spatiotemporally constrained pose refinement is applied for enhanced pose accuracy. This method is effective in dynamic and large-scale scene reconstructions [1].

NeRFusion is a hybrid approach that combines the speed of classical 3D reconstruction with the detail of NeRF. Its processes can be concluded in Fig.5.It quickly processes image sequences to predict local radiance fields and integrates them in real-time using a recurrent neural network, achieving a reconstruction rate of 22 frames per second. This method allows for high-resolution, detailed reconstructions that can be further refined for enhanced quality [9].

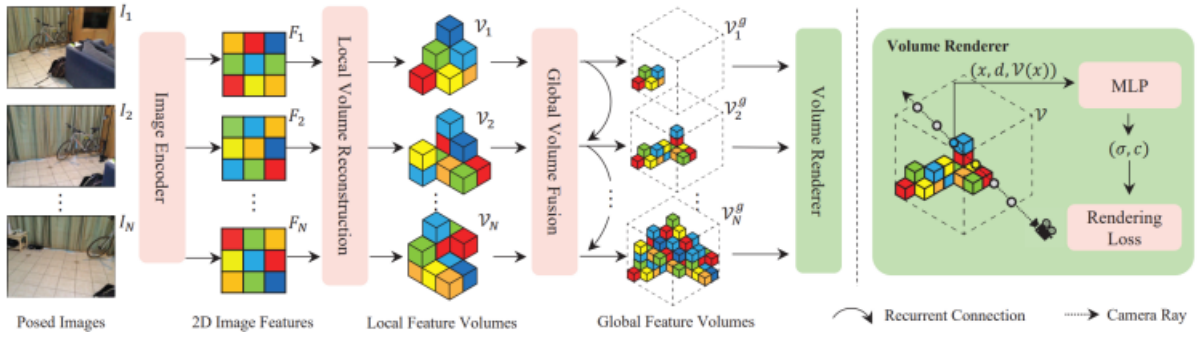


Fig. 5 Overview of the NeRFusion's framework [9].

4. Discussion

4.1. Features of Current Work

Recently researches have greatly enhanced the performances of NeRF, making it possible to reconstruct various of complex scenarios. To enable the application of NeRF in autonomous driving, researchers improve NeRF's abilities from several aspects.

The first aspect involves broadening the applicability and scenarios for NeRF, while reducing the limitations imposed by its reconstruction nature. The original NeRF heavily relied on the consistency of the static scenes to learn underlying representation, while new improved NeRF methods address these problem mainly through categorizing the nodes of scenes into different parts, which can be trained by different strategies. This improvement enlarge fields where NeRF can be utilized, such as dynamic 3D reconstruction by supervising videos, large-scale 3D reconstruction that the scales of reconstructed scene are extendable.

The second aspect is about fusing NeRF with different data sources or reconstruction method. To combine NeRF with different data sources such as map information and point cloud has a advantage that these data sources can perform as a guidance of the training, increasing the effect and efficiency of NeRF training and testing since these data sources often contain certain characteristics about ground truth 3D images. Additionally NeRF tend to fuse with different reconstruction method to make the most of advantages of both methods. They usually utilize the detailed and high-resolution reconstruction parameters and introduce another method to compensate for NeRF shortcoming in reconstructing dynamic and large-scale scene.

The third aspect focuses on the detailed and categorized modeling of the components that constitute traffic scenes. Some improved NeRF methods concentrate on urban scene reconstruction, including detailed surface reconstruction and adjustment of light exposure in different scenarios. Order improved NeRF methods focus on reconstruction of pedestrians, who are vital and critical component of transportation. What distinguish pedestrians from other traffic candidates is its non-rigid characteristic, which makes NeRF methods struggle to identify and render. Apart from this, several researches also focus on dynamic rigid nodes reconstruction, making NeRF be more powerful in handling dynamic scenes.

The last aspect prioritizes the time-efficiency and cost-efficiency of reconstruction. The original NeRF fails to a instantaneous image rendering and 3D reconstruction, since the method requires large quantities of 2D images input and a large number of iterations to achieve satisfactory results, reaching 300k iterations (1-2 days used on a single NVIDIA V100 GPU). Obviously this speed and time-consumption is not capable for efficient optimization. Optimization of data structures are often utilized to improve this situation. DVGO opts for the direct optimization of a voxel-based representation, eschewing the use of an MLP, which allows for training times that can be reduced to mere minutes or even seconds, NSV removes non-geometric voxels and refining voxels near object surfaces results in a sparse voxel representation or an octree expression. Instant-NGP and SUDS encode features in a multi-resolution hash table, achieving a real-time training and rendering. There are some adaptation on reducing the memory usage of reconstruction, allowing lower hardware requirement and cost [10][11][1].

4.2. Future Trend

Despite the significant advancements NeRF has made, there is still considerable room for development and optimization.

The first point is to enhance the generalization and inferential capabilities of the NeRF algorithm. In the context of 3D reconstruction for autonomous driving, there is often a frequent recurrence of similar scenes and objects, which tend to have comparable features and characteristics. Enhancing the generalization and inferential capabilities of NeRF can eliminate the need for dense shooting and independent training for each scene. This allows for direct inference of NeRF from sparse viewpoint images, thereby addressing the challenge of repetitive and similar scenes and objects.

The second point is representation disentanglement. Despite the many improvements in NeRF methods that have emerged, which now have the capability to recognize different objects in images and analyze dynamic and static scenes separately, there are still issues with inaccurate representation and poor robustness. The current NeRF processing capabilities are still limited when it comes to handling various geometric features, lighting, and materials. For instance, in some improved NeRF methods, although they can identify and separate vehicles, they fail to correctly recognize and process the shadows cast by these vehicles; sometimes, only by introducing light encoding can the lighting effects under complex lighting mechanisms be correctly represented. Currently, many studies are actively addressing these issues in pursuit of more realistic and higher precision scene reconstruction and object surface reconstruction effects.

The last but the most important point is efficiency. The original NeRF require huge quantities of training time and rendering time, since its complexity equals to search time of single sample point times numbers of sample point. Over 12 hours' training and rendering for a small scale scene is unbearable in AD. Thus reforming its training strategies and data structures becomes a must. Lots of methods are applied to address this problem. The methods, including using sparse geometric representation to reduce the number of samples, voxelization for low-complexity queries, hash technology for compressing high-resolution meshes to speed up grids, and voxel decomposition, have all made outstanding contributions to solving this problem. Future work should continue to delve deeper into research to reduce the training and rendering time of NeRF to the second level, in order to ultimately achieve efficiency optimization.

5. Conclusion

In this paper, a comprehensive summary of the application of Neural Radiance Fields in the field of 3D reconstruction for autonomous driving is provided. First, the principles of NeRF and various improvement methods based on NeRF in different aspects are demonstrated. Afterward, the characteristics of NeRF and NeRF-based improvement methods in four aspects are summarized , and the current shortcomings of NeRF-based methods, making predictions for future development directions are pointed out.

It is believed that, although NeRF offers better high-resolution and detail reconstruction compared to other traditional methods, it is still limited by its reliance on multi-view consistency, the ability to reconstruct only static small-scale scenes, and significant training time and costs. Improved NeRF methods are committed to expanding the application scenarios of NeRF, combining NeRF with other data sources, training specifically for different scenes and objects, and reducing training and rendering time as well as training and rendering costs. In the future, NeRF will evolve towards enhancing generalization capabilities, representation disentanglement, and improving efficiency, in order to increase the accuracy and efficiency of NeRF reconstruction.

References

[1] He, L., Li, L., Sun, W., Han, Z., Liu, Y., Zheng, S., Wang, J., Li, K.: 'Neural Radiance Field in Autonomous Driving: A Survey', ArXiv, 2024, abs/2404.13816

[2] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: 'NeRF', Communications of the ACM, 2020, 65, pp. 99-106

[3] Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F.: 'Neural Scene Graphs for Dynamic Scenes', 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2855-2864

[4] Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: 'Block-NeRF: Scalable Large Scene Neural View Synthesis', 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8238-8248

[5] Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T.A., Ferrari, V.: 'Urban Radiance Fields', 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12922-12932

[6] Weng, C., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: 'HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video', 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16189-16199

[7] Jiang, Y., Yao, K., Su, Z., Shen, Z., Luo, H., Xu, L.: 'Instant-NVR: Instant Neural Volumetric Rendering for Human-object Interactions from Monocular RGBD Stream', 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 595-605

[8] Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M., Wang, Y.: 'EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision', ArXiv, 2023, abs/2311.02077

[9] Zhang, X., Bi, S., Sunkavalli, K., Su, H., Xu, Z.: 'NeRFusion: Fusing Radiance Fields for Large-Scale Scene Reconstruction', 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5439-5448

[10] Liu, L., Gu, J., Lin, K.Z., Chua, T., Theobalt, C.: 'Neural Sparse Voxel Fields', ArXiv, 2020

[11] Müller, T., Evans, A., Schied, C., Keller, A.: 'Instant neural graphics primitives with a multiresolution hash encoding', ACM Transactions on Graphics (TOG), 2022, 41, pp. 1-15