Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **11**, Number **1**, February **2026**

A 3D Object Detection Strategy Based on NeRF in the Field of Autonomous Driving

Chen-Xia Wang[1]

[1]Department of Education and Teaching,
Zhengzhou Preschool Education College, Zhengzhou 450000, China

Ling-Ping Kong[2]

[2]Faculty of Computer Science,
VSB-Technical University of Ostrava, Ostrava 708 00, Czech Republic
lingping.kong@vsb.cz

Lin-Lin Tang[3,*]

[3]School of Computer Science and Technology,
Harbin Institute of Technology (Shenzhen), Shenzhen 518000, China
hittang@126.com

[*]Corresponding author: Lin-Lin Tang

Received May 9, 2024, revised February 7, 2025, accepted April 26, 2025.

Abstract. *Detecting targets in 2D images in field of autonomous driving is not enough, as 2D information cannot provide key attributes such as the true size, shape, and distance of objects. Meanwhile, the rise of deep learning technology provides new solutions for 3D object detection. Our article uses Multi-Layer Perceptron (MLP) to represent light field function and combines volume rendering to achieve high-quality differentiable neural rendering for 3D reconstruction. NeRFstudio modular framework is used for application implementation. Research focuses on 3D object detection based on multiple attention mechanisms from voxel and point features. Experiments have proven this method has achieved quite good results.*
Keywords: 3D Object Detection, MLP, NeRFstudio

1. **Introduction.** With rapid development of industries such as metaverse, spatial computing, XR, and autonomous driving in recent years, the integration of 3D vision, graphics, and machine learning has become increasingly close. NeRF [1], as an implicit representation of 3D objects or scenes, uses MLP [2] to represent light field functions and combines volume rendering for high quality differentiable neural rendering, bringing the task of new view synthesis and 3D reconstruction to a new stage. NeRFstudio [3] supports multiple data acquisition methods to obtain inputs. It has built various samplers, radiation fields, encoders, renderers. It has achieved real time rendering visualization. And it supports exporting various 3D representations such as videos, point clouds, and meshes. The 3D reconstruction flowchart of NeRFstudio can be shown as the following Figure 1.

Through an automated 3D reconstruction system, the cost of manual modeling can be greatly reduced. Users can take the captured video as input to the system and use a visual interactive interface to view the reconstruction results. Whether it is object or scene modeling in film and television games, digital twins and mixed reality, or architectural design and industrial modeling, they all rely on a large amount of 3D digital assets.

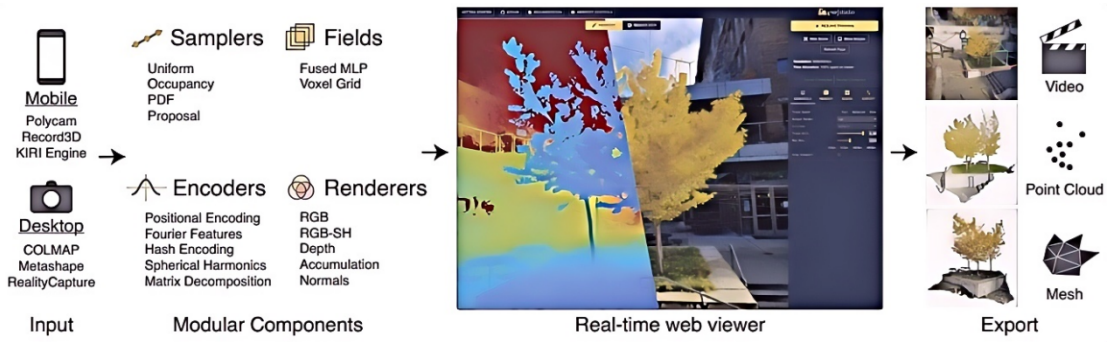

Figure 1. 3D Reconstruction Flowchart of NeRFstudio

New View Synthesis (NVS) [4] typically uses an intermediate 3D scene representation as an intermediary to generate high-quality new view images, which are rendered. 3D representation includes both display and implicit representations, using different 3D representations for different 3D tasks. Display representations include point clouds, meshes, voxels, and more. It can explicitly model scenes, but because it is a discrete representation, model may not be refined enough and may produce overlapping artifacts. More, it stores a large amount of three-dimensional scene representation information, which consumes a lot of memory and limits its application in high resolution scenes. Implicit representation usually uses a function to describe the geometry of the scene, storing complex three-dimensional scene representation information in the parameters of the function, including NeRF, Signed Distance Function (SDF) [5], etc. Implicit representation has relatively fewer parameters when expressing high resolution scenes, and implicit representation function or MLP is a continuous representation, making the scene representation more refined.

Neural Radiation Field (NeRF) [1] uses neural networks to fit the light field function as an implicit representation, adopts volume rendering, uses 2D supervision to achieve 3D optimization. Volume rendering first samples some points from the camera's optical center to the rays of each pixel, and then obtains information of these points through MLP. The integration (sum) operation is used to aggregate the color and volume density information of these points along the rays to obtain the predicted 2D pixel values. We can assume that the scene is composed of a cluster of glowing particles, whose colors change from different perspectives, and transparency of particles in different positions also varies. Assuming a ray is emitted from a viewpoint and passes through a certain pixel on the imaging plane before entering this cluster of particles, the ray may collide with the particle at some point. If the opacity of the particle is high, the probability of the ray hitting the particle is high, the probability of the ray stopping at this point is high, and the color weight of the particle is high. The weighted sum of the color and density of the points on this ray is volume rendering.

NeRF is a continuous 3D representation implicitly represents empty/occupied space. Random sampling achieves layered sampling with high computational overhead. In contrast, points are unstructured discrete representations that are flexible enough to allow for the creation, destruction, and movement of geometric shapes similar to NeRF. By optimizing opacity and position, as shown in previous work, this can avoid the drawbacks of full volume representation. The latest 3DGS uses sparse point clouds reconstructed by Structure-from-Motion (SfM) [6] for initialization, learning each point as a 3D Gaussian ellipsoid and projecting it onto a 2D plane using Splating. Then, alpha mixing and differentiable grating are used to obtain 2D images, and an adaptive point adjustment strategy is proposed to replicate and split the 3D Gaussian.

3D object detection is foundation of important applications such as robotics and autonomous driving, which require understanding of 3D scenes. Most existing related methods require 3D point cloud input or at least obtaining RGBD images from 3D sensors. However, the latest advances in neural radiation fields (NeRF) provide an effective alternative method for extracting high semantic features of underlying 3D scenes from 2D multi-view images.

Taking inspiration from the Region Proposal Network (RPN) [5] for 2D object detection, some work on an improved 3D NeRF-RPN [7] still directly operates on the NeRF representation of a given 3D scene learned entirely from RGB images and camera poses, with improvements in feature extraction. Main is focused on MLP and the multiple attention mechanisms from voxel and point features.

Contribution. In this paper, we propose a new integrity audit scheme for cloud storage shared data with dynamic user updates, supporting data privacy and user identity privacy protection. The contribution of our paper can be summed up as the following three points:

(1) We propose a new dynamic multi-tree storage structure, which the CSP maintains to achieve dynamic data and dynamic group management, effectively improving data space utilization.

(2) The group administrator maintains an anonymous identity table named IAT in our scheme. Only the group administrator knows the real identity information of group users, effectively improving the user's identity privacy protection.

(3) To protect user data privacy, we encrypt the data-proof information generated by the CSP to prevent the verifier from obtaining valid user information in the audit stage. Our scheme has stronger security than constructing linearly correlated data-proof information.

2. Realted work.

2.1. NeRF.

Neural Radiance Fields (NeRF) is a computer vision technique used to generate high-quality 3D reconstruction models. It uses deep learning techniques to extract the geometric shape and texture information of objects from images from multiple perspectives, and then uses this information to generate a continuous 3D radiation field, which can present highly realistic 3D models at any angle and distance. The basic algorithm flowchart is as following Figure 2.

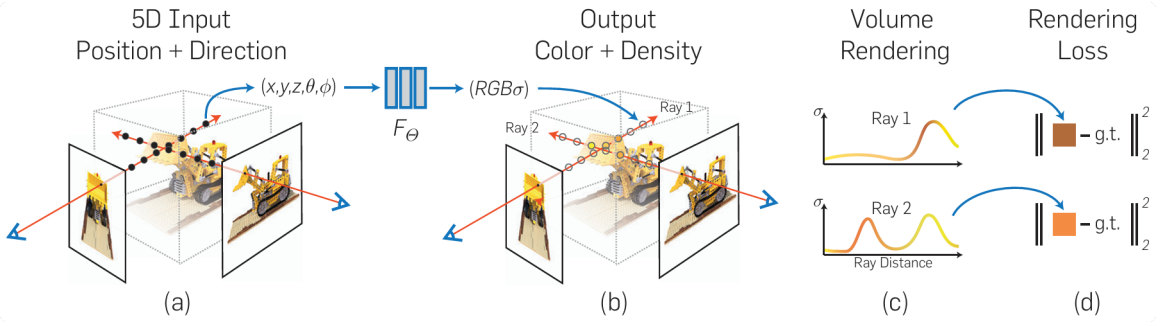

Figure 2. NeRF Algorithm Flowchart

The working principle of NeRF can be summarized by following steps. Firstly, it is necessary to collect images from multiple perspectives. These images can be from cameras, laser scanners, or other devices. Then, deep learning techniques are used to transform these images into a continuous three-dimensional radiation field. Radiation field is a function that describes how light propagates in three-dimensional space. Finally, this

radiation field can be used to generate three-dimensional models from any perspective. The neural radiation field has the following advantages: it can extract information from images from multiple perspectives, thereby generating more realistic 3D models. It can generate 3D models from any perspective to meet the needs of different applications. It can perform fast rendering to achieve real-time interaction.

PlenOctrees [8] use an octave based radiation field and spherical basis function mesh to accelerate rendering and appearance decoding. TensoRF [9] projects three-dimensional points onto three two-dimensional planes to encode positional information. Although these works use different structural modeling methods, they achieve the same goal of generating RGB colors and volume density related to the view at each position by inputting xyz coordinates and 3D camera poses, thereby rendering the image from a given viewpoint.

2.2. **3D Reconstruction.** Compared to traditional 3D reconstruction and differentiable rendering algorithms, NeRF and its variants introduce new 3D representation and rendering methods, which only require RGB images as input to achieve high-quality reconstruction rendering. With the continuous improvement of NeRF based reconstruction methods, their reconstruction quality and speed, as well as their effects in various subtasks, are constantly improving.

3D reconstruction is a traditional problem. Before emergence of NeRF, important methods such as SfM [5] and MVS [10] were used. For different 3D representations (point clouds, meshes, voxel meshes, SDF, NeRF, etc.), reconstruction methods also differ. New View Synthesis (NVS) [11] differs from 3D reconstruction, which requires generating/rendering images with unknown perspectives, while reconstruction generally requires extracting meshes, Because mesh based rendering pipelines are still the mainstream rendering method at present, and mesh can also be edited and physically simulated.

The rendering methods include surface rendering and volume rendering, which are crucial in differentiable rendering and inverse rendering. Mesh can achieve real-time rendering using traditional physics based PBR rendering (which the original NeRF cannot achieve) [12], but NeRF uses a learnable volume rendering method combined with neural networks (MLP) to store continuous spatial information (color, density), which can achieve high-quality new view synthesis. The subsequent mixed representation of various radiation fields improved various issues of NeRF, including the introduction of multi-resolution hash encoding and lightweight MLP hybrid 3D mesh structure in Instant-NGP, as well as cuda acceleration training and rendering processes; TensoRF uses a 3-plane structure to represent spatial features, achieving acceleration and storage optimization; Plenoxels discards the neural network and uses perspective independent spherical harmonic functions to represent anisotropic color information, also achieving accelerated training. Neuralangelo uses Instant-NGP as the neural SDF representation of the underlying 3D scene, optimizes from multi view images through neural surface rendering, calculates high-order derivatives using numerical gradients, and progressively optimizes to restore different levels of detail structures as shown in Figure 3. Significant improvements have been made in reconstruction accuracy and view synthesis quality.

2.3. **3D Object Detection.** 3D object detection can intelligently predict the position, size, and category of key 3D targets near autonomous vehicles, and is an important component of perception systems. Existing 3D object detection methods attempt to solve the problem of 3D object detection from specific aspects, such as specific sensor types, data representation, etc., but lack systematic comparison with other categories of methods. Therefore, a comprehensive analysis of the advantages and disadvantages of various types of 3D object detection methods can provide some reference for relevant

Figure 3. Neuralangelo Reconstructs Large Scenes

researchers. Based on this purpose, this article comprehensively reviews the 3D object detection methods in autonomous driving applications, and conducts in-depth analysis and systematic comparison of different methods. Compared with the existing review articles [13], this paper extensively covers the latest developments in this field, such as 3D target detection based on depth images, self/semi/weakly supervised 3D target detection, and 3D target detection in end-to-end auto drive system systems.

3. **Our Proposed Method.**

3.1. **Overall 3D Reconstruction Plan.** Firstly, we need to collect data on the reconstructed target, as our method is entirely based on RGB images, so laser scanning is not necessary. Of course, if conditions permit, obtaining more accurate sparse point clouds and camera poses can better improve the reconstruction effect. The process of data collection:

(1) Shooting videos: For individual objects, collect multi view images from all angles as much as possible to ensure that the reconstruction results do not have artifacts; For large scenes, it is advisable to use movement methods such as translation or curved movement to capture videos, or use drone footage to capture. It is important not to collect data in place. The input video for this reconstruction is a 4K 30FPS video captured by a mobile phone.

(2) Video to image: Use ffmpeg to extract frames from the video. In this reconstruction, approximately 1-3 frames of images are selected per second, and the images are numbered and organized into folders that meet the format.

(3) COLMAP Sparse Reconstruction: Use the installed colmap to perform sparse point cloud reconstruction on multi view RGB images, estimate the camera pose of the image, and obtain initial point cloud information.

COLMAP is a 3D reconstruction pipeline that combines MVS (Multi View Stereo). After successful compilation, we can obtain software with a graphical interface (Graphic Interface) and binary executable files without a graphical interface, which can perform sparse and dense reconstruction. Process of using COLMAP for sparse reconstruction includes feature extraction, feature matching, incremental reconstruction (gradually increasing the perspective and iteratively optimizing the reprojection error, calculating camera parameters for different views, obtaining sparse point clouds of the scene, and determining the visual relationship between different views and point clouds), ultimately obtaining sparse point clouds of the target scene and camera poses for each perspective.

3.2. **Core Functions and Algorithms.** 3D Gaussian Splatting is similar to previous NeRF based methods, using SfM calibrated camera parameters and sparse point clouds generated during the SfM process to initialize the 3D Gaussian set. Compared to most point based solutions that require multi view stereoscopic (MVS) data, 3D Gaussian Splatting achieves high-quality results using only SfM points as inputs (even with random initialization, high-quality reconstruction results can be obtained for NeRF syntactic datasets). 3D Gaussian is a differentiable volume representation that is effectively rasterized by projecting them onto 2D and applying standard alpha blending. During the reconstruction process, optimizing the 3D position, opacity, anisotropic covariance, and spherical harmonic (SH) coefficients of 3D Gauss is alternated with adaptive density control (adding and removing 3D Gauss) steps. To achieve real-time rendering, 3DGS uses a fast GPU sorting algorithm and tile-based rasterization. The following Figure 4 and Figure 5 is the algorithm flowchart of 3DGS and its comparison with other reconstruction algorithms:

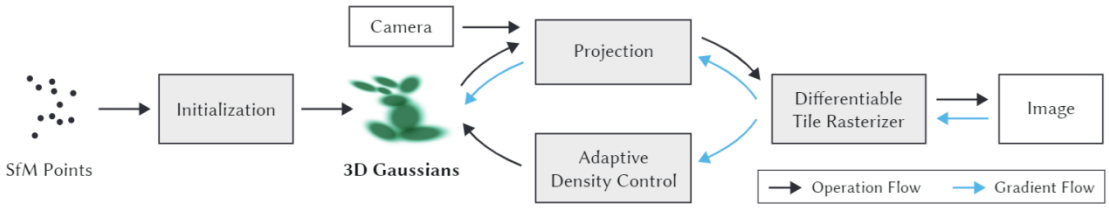

Figure 4. Flowchart of 3D Gaussian Splatting

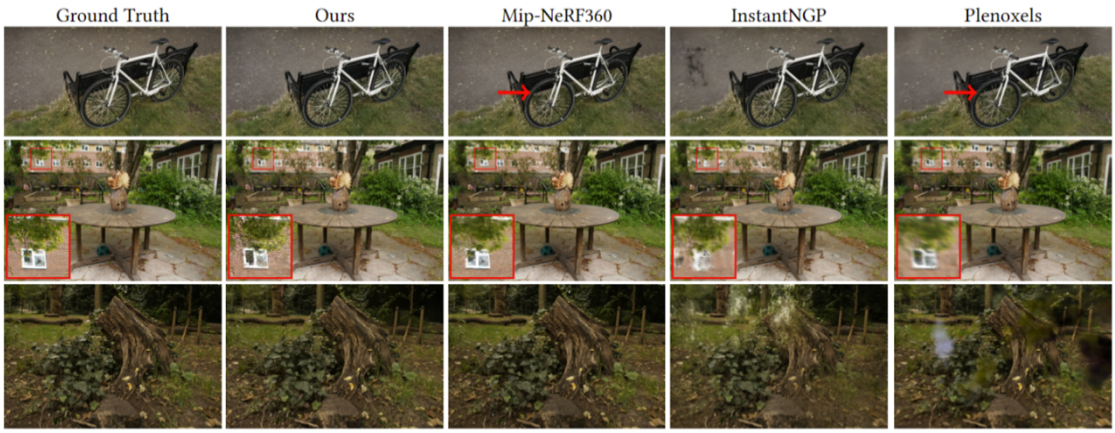

Figure 5. Comparison Between 3DGS and Other Methods

The following Figure 6 and Figure 7 is a 3D reconstruction system built using the 3DGS algorithm. Reconstruction results are displayed on a synthetic dataset of neural synthetic and self captured real data (small objects, computer rooms, and school teaching buildings), and rendered in real-time on an RTX2060 graphics card (30FPS).

3.3. **3D Object Detection.** The 3D object detection algorithms based on voxel and point features with multiple attention mechanisms mainly include the following three:

(1) PointPillars: PointPillars [14] is a 3D object detection algorithm based on point clouds. It uses voxelization to convert point cloud data into sparse voxel representations, and then extracts features and performs object detection through multiple attention mechanisms.

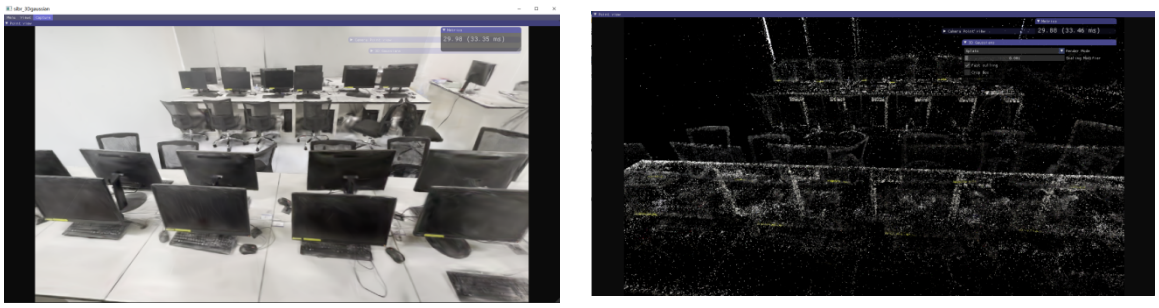

Figure 6. Classroom Reconstruction and Point Cloud Effect Display

Figure 7. Display of Composite Dataset Reconstruction Results

(2) PV-RCNN: PV-RCNN [15] is a 3D object detection algorithm based on point clouds and voxels. It combines point cloud features and voxel features, and combines these two features through multiple attention mechanisms to improve detection performance.

(3) Point GNN: Point GNN [16] is a 3D object detection algorithm based on graph neural networks. It represents point cloud data as a graph structure, learns relationships between points through multiple attention mechanisms, and achieves object detection.

These algorithms all use multiple attention mechanisms to process point cloud data and voxel data, and improve feature representation and detection performance by introducing attention mechanisms. They have achieved good results in the field of 3D object detection and have been widely used in fields such as autonomous driving and robotics.

Our algorithm further explores the application of multi-attention mechanism in fusing voxel and point features, and proposes a more effective feature fusion method to improve the accuracy and efficiency of object detection. Meanwhile, this article has designed a more flexible network structure that can adapt to object detection tasks in different scenarios and scales, improving the algorithm's generalization ability and adaptability. The details can be described as follows:

(1) **Input data representation**: Firstly, represent the input 3D point cloud data in the form of voxels or point features. Usually, point cloud data is converted into sparse voxel representations or directly represented using point features.

(2) **Feature extraction**: Using Convolutional Neural Networks (CNN) or other feature extraction networks to extract features from input voxels or point features. Step aims to extract meaningful feature representations from the raw data for subsequent object detection tasks.

(3) **Multiple attention mechanism**: Introducing multiple attention mechanisms to fuse feature information of different levels or types. Multiple attention mechanisms can include multi head attention mechanisms, cross level attention mechanisms, etc., to capture the correlation and importance between different features.

(4) **Feature fusion**: Utilizing multiple attention mechanisms to fuse feature information of different levels or types, in order to obtain richer and more accurate feature representations. This step aims to improve the expression ability and discrimination of features, thereby improving the accuracy and robustness of object detection.

(5) **Output prediction**: Finally, using the fused feature representation for object detection or classification tasks, generate target position and category information. Subsequent classifiers or regressors are usually used for object detection or classification.

4. **Experiments.** The KITTI object detection benchmark dataset [17], which consists of samples that have both lidar point clouds and images has been used to check results. At inference time we apply axis aligned Non Maximum Suppression (NMS) with an overlap threshold of 0.5 IoU. For comparison with the method in [14], we also randomly select 15, 0, 8 ground truth samples for cars, pedestrians, and cyclists respectively and place them into the current point cloud. A global translation with x, y, z drawn from N (0, 0.2) to simulate localization noise has been applied. As shown in the following Table 1, our proposed method based on NeRFstudio, improved multi attention mechanisms and fused features performs pretty well.

Table 1. Results on the KITTI Test 3D Detection Benchmark

Method	Modality	Speed (Hz)	mAP	Car			Pedestrian			Cyclist		
			Mod.	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MV3D [18]	Lidar & Img	2.8	N/A	86.02	76.9	68.49	N/A	N/A	N/A	N/A	N/A	N/A
Cont-Fuse [19]	Lidar & Img	16.7	N/A	88.81	85.83	77.33	N/A	N/A	N/A	N/A	N/A	N/A
Roarnet [20]	Lidar & Img	10	N/A	88.2	79.41	70.02	N/A	N/A	N/A	N/A	N/A	N/A
AVOD-FPN [21]	Lidar & Img	10	64.11	88.53	83.79	77.9	58.75	51.05	47.54	68.09	57.48	50.77
F-PointNet [22]	Lidar & Img	5.9	65.39	88.7	84	75.33	58.09	50.22	47.2	75.38	61.96	54.68
HDNET [22]	Lidar & Map	20	N/A	89.14	86.57	78.32	N/A	N/A	N/A	N/A	N/A	N/A
PIXOR++ [23]	Lidar	35	N/A	89.38	83.7	77.97	N/A	N/A	N/A	N/A	N/A	N/A
VoxelNet [24]	Lidar	4.4	58.25	89.35	79.26	77.39	46.13	40.74	38.11	66.7	54.76	50.55
SECOND [25]	Lidar	20	60.56	88.07	79.37	77.95	55.1	46.27	44.76	73.67	56.04	48.78
PointPillars [14]	Lidar	62	66.19	88.35	86.1	79.83	58.66	50.23	47.19	79.14	62.25	56
Ours	Lidar	61.7	67	89.2	88.14	79.99	58.92	50.02	46.98	78.97	63.11	57.07

As we can see from the above Table 1, our proposed method has achieved a pretty good performance. It is superior to current popular algorithms in most datasets and comparisons. The only downside or potential improvement is finding ways to reduce the complexity of computation and design.

5. **Conclusion.** This paper uses Multi-Layer Perceptron (MLP) to represent light field function and combines volume rendering to achieve high-quality differentiable neural rendering for 3D reconstruction. NeRFstudio modular framework is used for application

implementation. Though much work has been proposed in related fields [25, 26, 27]. Our research focuses on 3D object detection based on multiple attention mechanisms from the voxel and point features. Experiments have shown its efficiency. The future work should be focused on the reduce of the algorithm complexity and improvement of the generation quality and recognition accuracy.

Acknowledgment. This work is supported by Shenzhen Major Science and Technology Special Project with Grant No. KJZD20240903102727035 and Key Basic Research Projects of Shenzhen with Grant No. JCYJ20220818102414030. And this research is also supported by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

REFERENCES

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 454–459.

[3] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–12.

[4] B. Attal, E. Laidlaw, A. Gokaslan, C. Kim, C. Richardt, J. Tompkin, and M. O'Toole, "Törf: Time-of-flight radiance fields for dynamic scene view synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 289–26 301, 2021.

[5] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[6] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[7] B. Hu, J. Huang, Y. Liu, Y.-W. Tai, and C.-K. Tang, "Nerf-rpn: A general framework for object detection in nerfs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 528–23 538.

[8] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5752–5761.

[9] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision*. Springer, 2022, pp. 333–350.

[10] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2041–2050.

[11] Z. Li, S. Niklaus, N. Snavely, and O. 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*, 2021, pp. 6498–6508.

[12] B. Ham, D. Min, C. Oh, M. N. Do, and K. Sohn, "Probability-based rendering for view synthesis," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 870–884, 2013.

[13] C. Oh, Y. Jang, D. Shim, C. Kim, J. Kim, and H. J. Kim, "Automatic pseudo-lidar annotation: Generation of training data for 3d object detection networks," *IEEE Access*, vol. 12, pp. 14 227–14 237, 2024.

[14] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[15] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 529–10 538.

[16] W. Shi and R. Rajkumar, "Point-gnn: Graph neural network for 3d object detection in a point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1711–1719.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[18] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[19] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 641–656.

[20] K. Shin, Y. P. Kwon, and M. Tomizuka, "Roarnet: A robust 3d object detection based on region approximation refinement," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 2510–2515.

[21] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[22] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[23] B. Yang, M. Liang, and R. Urtasun, "Hdnet: Exploiting hd maps for 3d object detection," in *Conference on Robot Learning*. PMLR, 2018, pp. 146–155.

[24] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[25] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[26] C.-M. Chen, Q. Miao, S. Kumar, and T.-Y. Wu, "Privacy-preserving authentication scheme for digital twin-enabled autonomous vehicle environments," *Transactions on Emerging Telecommunications Technologies*, vol. 34, no. 11, p. e4751, 2023.

[27] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on svm in vr art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 40, 2019.