Dean&Francis
ISSN 2959-6157

A review of 3D reconstruction methods based on deep learning

Liwei Wang

School of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, China
Corresponding author: U2184965@unimail.hud.ac.uk

Abstract:

3D reconstruction is a technical process that constructs a digital 3D model of a target object from low-dimensional data. It plays an important role in medical imaging, cultural relics protection and other fields.Traditional 3D reconstruction techniques suffer from challenges such as difficult feature extraction and heavy manual intervention. Therefore, deep learning has been introduced into this field. After extensive literature review, this paper systematically summarizes classic 3D reconstruction algorithms using deep learning methods, categorizing them into explicit and implicit representation approaches. As cutting-edge technologies in 3D reconstruction, Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) hold significant promise. This paper briefly introduces the fundamental principles and recent advancements of dynamic scenes, outlines commonly used dynamic scene datasets and performance metrics, and compares their performance on the D-NeRF datasets. It concludes by summarizing the main challenges in 3D reconstruction and looks ahead to future developments in technology integration and reducing memory costs for large-scale scenes.

Keywords: deep learning; 3D reconstruction; NeRF; 3DGS;

1. Introduction

As a core technology in computer vision and computer graphics fields, 3D reconstruction primarily aims to transform low-dimensional data into computer-processable, analyzable, and renderable 3D digital models. With the continuous development of deep learning, 3D reconstruction technology has gradually become an indispensable part of medical imaging , cultural heritage protection , virtual reality, autonomous driving and robotics.

Currently, 3D reconstruction technology can be categorized into traditional 3D reconstruction techniques and deep learning-based 3D reconstruction techniques based on core principles. Traditional 3D reconstruction methods are designed based on geometric, optical, and optimization theories, relying on physical models without the need for training data. Traditional methods that use the Multi-View Geometry (MVG) principle struggle with feature extraction [1]. In order to enhance automation and seemingly improve the accuracy of feature extraction and matching, researchers have, within this broader analytical framework, introduced what tends to be

deep learning into what appears to be this field.

Previous studies have been carried out on 3d reconstruction technology based on deep learning. According to the difference between explicit and implicit representation, this paper divides typical 3D reconstruction methods into explicit and implicit methods. With the rapid development of autonomous driving and augmented reality, there is a growing demand for real-time dynamic scene reconstruction. Previous 3D reconstruction techniques are well-established and excel in static scenes, but their application to dynamic scenes is still in its infancy. NeRF and 3DGS, as cutting-edge technologies in 3D reconstruction, warrant further research and summarization [2,3]. Therefore, this paper summarizes and compares the methods of extending NeRF and 3DGS in dynamic scenes. Finally, this paper discusses the problems existing in 3d reconstruction and looks forward to the possible development direction in the future.

2. Methodology

2.1 Explicit Representation for 3D Reconstruction

The explicit representation is a form of representation that maps the positions of an object into 3D space and directly expresses the contour and shape information of the object through coordinates. Common explicit representations include voxels, point clouds, and meshes. Among them, a voxel is the smallest unit of 3D space segmentation; a point cloud consists of a large number of spatial points, each with 3D coordinates and additional attributes such as color, reflection intensity, and normal vectors; a mesh contains a series of 3D vertex coordinates and polygon faces composed of several vertices.

2.1.1 Voxel-Based Methods

3D ShapeNets model pioneered the application of Convolutional Deep Belief Networks (CDBNs) to 3D shape processing, establishing a new paradigm for deep learning in three-dimensional space [4]. 3D-R2N2 not only extended the time-series processing capability of traditional LSTM to 3D space, but also first adopted a unified network framework to handle 3D reconstruction tasks for both single-view and multi-view [5]. Pix2Vox leveraged a context-aware fusion module to enhance inference speed [6]. However, due to the cubic growth in memory consumption of voxel-based methods with increasing resolution, higher resolutions incur significant computational overhead, making them impractical for high-fidelity scenarios.

2.1.2 Point Cloud-Based Methods

Compared to voxel and point cloud models reconstruct smoother shapes while consuming less memory. PointNet, as the first deep learning model capable of directly processing raw point clouds, addresses the challenges of unordered nature and permutation invariance of point cloud data [7]. PCN (Point Completion Network) was subsequently proposed to solve the point cloud completion problem and improve processing accuracy [8]. However, point cloud models still suffer from a lack of surface continuity, resulting in non-smooth reconstructed surfaces.

2.1.3 Mesh-Based Methods

Compared to voxel and point cloud representations, mesh-based models can fully represent an object's surface geometry while being more render-friendly. The core methodology involves learning geometric and topological features from input data through Graph Convolutional Networks (GCNs), followed by progressively optimizing mesh vertices and faces to gradually approximate the target object's true shape. The main methods include Pixel2Mesh [9], Pixel2Mesh++ and similar methods [10]. Table 1 lists the compares explicit methods.

Table 1. Comparison of Explicit Methods

Explicit representation	Method	Year	Advantage	Shortcoming
Voxel-Based	3D ShapeNets[4]	2015	Regular structures are well-suited for deep learning	Memory consumption scales cubically with resolution
	3D-R2N2[5]	2016		
	Pix2Vox[6]	2019		
Point Cloud-Based	PointNet[7]	2017	Memory-efficient	Surface discontinuity
	PCN[8]	2018		
Mesh-Based	Pixel2Mesh[9]	2018	Render-friendly	Fixed topology
	Pixel2Mesh++[10]	2019		

While explicit 3D reconstruction achieves higher accuracy compared to traditional reconstruction methods, its discrete nature still struggles with complex scenes and topological variations. Implicit 3D reconstruction effectively addresses these limitations.

ISSN 2959-6157

2.2 Implicit Representation for 3D Reconstruction.

Implicit representation utilizes continuous implicit functions to characterize an object's occupancy in 3D space. Rather than explicitly specifying spatial coordinates, it determines object boundaries through learned functional relationships. Implicit functions primarily include occupancy fields and signed distance functions. DeepSDF is one of typical Implicit representation method [11]. Early implicit 3D reconstruction heavily relies on supervised learning with extensive 3D geometric labels, which significantly limits its applicability and deployment scenarios. Therefore, in 2020, Mildenhall et al. proposed Neural Radiance Fields, a novel neural 5D scene representation method [2]. By optimizing a continuous radiance field through differentiable volume rendering techniques, it achieves photorealistic novel view synthesis.

2.2.1 Neural Radiance Fields

The fundamental principle of NeRF involves feeding a 3D spatial coordinate and 2D viewing direction as inputs, which are first transformed into higher-dimensional vectors via positional encoding. These encoded features train a multilayer perceptron (MLP) that outputs volumetric density and view-dependent color. The volume rendering integral is approximated through discrete summation of colors and densities along camera rays, ultimately yielding the pixel's rendered color [2].

To achieve the application of dynamic scenes, researchers have conducted extensive research and achieved significant results. In 2021, Albert et al. proposed D-NeRF, which introduced a temporal variable and adopted a dual-network architecture comprising a canonical network to encode static scenes and a deformation network to model dynamic deformations. This work marked the first successful extension of NeRF to dynamic scenes, enabling neural rendering of dynamic environments learned from sparse monocular camera images [12].HyperNeRF addressed topological changes, enabling more complex dynamic modeling [13].

In 2022, TiNeuVox, a lightweight architecture based on voxel grids, accelerated dynamic radiance field inference through time-aware interpolation, significantly enhancing the training speed and real-time performance of dynamic NeRF [14].

In 2023, Sara Fridovich-Keil et al. introduced K-Planes, a representation that decomposes a D-dimensional scene into (d-choose-2) planes, enabling natural separation of static and dynamic components [15]. Similar to K-Planes, HexPlane decomposed 4D spatiotemporal features (X, Y, Z, T) into six orthogonal feature planes (e.g., XY, ZT, XT), which were then fused and decoded by a lightweight MLP for dynamic neural rendering. HexPlane reduced the computational overhead and improved the training speed [16].

However, due to the indecomposable implicit representation of NeRF, it is difficult to perform 3D editing, while the emergence of 3D Gaussian Splatting has solved this problem [3]. Fig.1 illustrates the development of NeRF and 3DGaussian splashing in dynamic scenes.

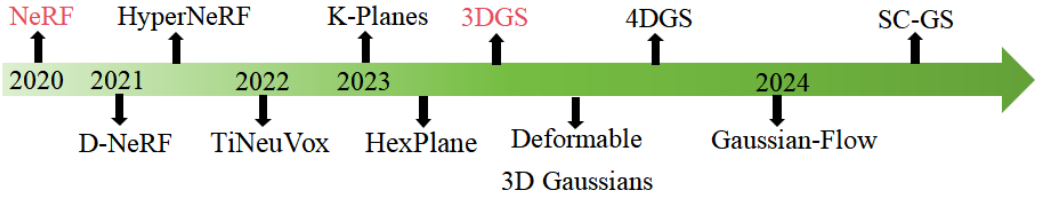

Fig. 1 The development of NeRF and 3DGaussian splashing in dynamic scenes.

2.3 3D Gaussian splashing

3D Gaussian Splatting densely populates the target scene with a multitude of 3D Gaussian primitives as its explicit representation, then achieves highly efficient image rendering through parallelized rasterization algorithms. Compared to NeRF, 3DGS not only enables real-time rendering of scenes, but also improves the editability of scenes [3].The researchers have studied the extension of 3D Gaussian splashing to dynamic scenes in depth.

In late 2023, two distinct technical approaches were successively proposed to extend static 3D Gaussian splatting to dynamic scene representation. Yang et al. proposed the Deformable 3D Gaussians method, which learns 3D Gaussians in canonical space coupled with a deformation field network for dynamic scene modeling. The approach introduces an Annealed Smooth Training (AST) mechanism to mitigate the impact of pose estimation errors, enabling high-fidelity reconstruction of monocular dynamic scenes [17]. Wu et al. proposed 4D-GS, a novel framework that combines 3D Gaussian primitives with explicit 4D neural voxel representations for dynamic scene modeling, achieving an optimal balance between rendering quality and computational efficiency in dynamic scene reconstruction [18].

In 2024,to address the inefficiencies in training and rendering for dynamic scene reconstruction in prior methods,

Lin et al. proposed Gaussian-Flow, introducing a Dual-Domain Deformation Model (DDDM). This approach pioneers the joint use of time-domain polynomials and frequency-domain Fourier series to fit the dynamic attributes (position, rotation, radiance) of Gaussian particles, eliminating the need for per-frame optimization or computationally intensive neural network evaluations [19]. SC-GS employed sparse control points coupled with a deformation MLP to efficiently drive 3D Gaussians for dynamic scene modeling, enhancing its applicability in motion editing for dynamic scenes [20].

3. Comparative Experiments and Results

3.1 Datasets and Performance Evaluation Metrics

Datasets commonly used for testing methods in dynamic scenarios:
(1)D-NeRF Datasets: Contains 8 synthesized dynamic scenes to evaluate the new perspective synthesis and time consistency of dynamic scenes
(2)HyperNeRF Datasets: Contains real and synthetic sequences to extend NeRF to handle topological changes in dynamic scenes
(3)NeRF-DS Datasets: Contains 7 real scenes with complex reflection and refraction effects, focusing on the rendering of dynamic highlight objects under different lighting.
Evaluation Metrics:
(1)PSNR: Measures the pixel-level error between the reconstructed image and the real image. The higher the value, the better the reconstruction effect.
(2)SSIM: Image similarity is evaluated based on brightness, contrast and structure. The range tends to fall within [-1,1], and what the evidence appears to reveal is that the higher the value, the higher the similarity seems to be, with 1 presumably indicating complete consistency.
(3)LPIPS: Calculate the difference of image blocks in the feature space and measure the perceptual similarity. The range ostensibly falls within [0,1], and the lower the value, the higher the similarity.

3.2 Analysis of Result

In order to compare the performance of dynamic scene reconstruction algorithms based on NeRF and 3DGS, this paper collects and summarizes previous data, and shows the view synthesis quality comparison of reconstruction algorithms in dynamic scenes based on NeRF and 3DGS on D-NeRF data set in Table 2.

Table 2. Comparison of view synthesis quality of dynamic scene reconstruction algorithms of NeRF and 3D GS on D-NeRF datasets

Method	Type	PSNR↑	SSIM↑	LPIPS↓
D-NeRF[12]	NeRF	30.50	0.95	0.07
HexPlane[16]	NeRF	31.04	0.97	0.04
K-Planes[15]	NeRF	31.61	0.97	0.06
TiNeuVox[14]	NeRF	32.67	0.97	0.04
3D-GS[3]	3DGS	23.19	0.93	0.08
4D-GS[18]	3DGS	34.05	0.98	0.02
Deformable 3D Gaussians[17]	3DGS	39.51	0.99	0.01
SC-GS[20]	3DGS	43.31	0.99	0.01

The experimental results show that TiNeuVox, a NeRF-based method, achieves the highest view synthesis quality with a PSNR of 32.67. In contrast, traditional 3D-GS performs the worst, with a PSNR of only 23.19. However, the introduction of 4DGS and deformation fields has significantly improved the performance in dynamic scenes, with SC-GS achieving a PSNR of 43.31, surpassing even the best algorithms for dynamic scene reconstruction based on NeRF.

4. Conclusion and Discussion

Although the 3D reconstruction technology based on deep learning is now very mature, there are still many challenges for researchers to explore better 3D reconstruction solutions and scene applications. This paper starts from the future research direction and development trend, and puts forward several questions worth exploring in depth:
(1)The limitations of a single technology in real-time performance, reconstruction accuracy, and adaptability to dynamic scenes:
The integration of NeRF with 3D Gaussian Splatting (3DGS) can complement each other to address the challenges of dynamic scene modeling, real-time rendering, and training efficiency. 3DGS achieves efficient rendering through explicit Gaussian distributions, while implicit

ISSN 2959-6157

representation of NeRF enhances detail precision. This combination ensures both speed and quality, for example, using 3DGS to handle dynamic areas and NeRF to optimize static backgrounds.

(2)High computing overhead in large-scale scenarios:

At present, 3D reconstruction technology has high computing overhead in large-scale scenarios, which is worth solving. In the future, algorithm optimization will be carried out in lightweight network architecture, incremental and block reconstruction, and adaptive computing resource allocation.

In short, 3D reconstruction based on deep learning has a bright future research direction and rich research value, which is also an important reason why it is developed by a large number of researchers. The content and contribution of this paper are summarized below.

This article reviews the research progress of 3D reconstruction technology based on deep learning. It systematically summarizes both explicit and implicit 3D reconstruction methods, starting from different representation approaches. By tracing the technical development from NeRF to 3DGS in dynamic scenes, it links these advancements and compiles previous research findings to provide an intuitive comparison of performance improvements. Finally, it discusses future research directions, including the integration of NeRF and 3DGS for large-scale scene modeling. This article provides a comprehensive technical summary and development outlook for the field of 3D reconstruction.

References

[1] Hartley R. Multiple view geometry in computer vision[M]. Cambridge university press, 2003.

[2] 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.

[3] 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.

[4] Wu Z, Song S, Khosla A, et al. 3d shapenets: A deep representation for volumetric shapes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1912-1920.

[5] Choy C B, Xu D, Gwak J Y, et al. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction[C]// Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11-14, 2016, proceedings, part VIII 14. Springer International Publishing, 2016: 628-644.

[6] Xie H, Yao H, Sun X, et al. Pix2vox: Context-aware 3d reconstruction from single and multi-view images[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2690-2698.

[7] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[8] Yuan W, Khot T, Held D, et al. Pcn: Point completion network[C]//2018 international conference on 3D vision (3DV). IEEE, 2018: 728-737.

[9] Wang N, Zhang Y, Li Z, et al. Pixel2mesh: Generating 3d mesh models from single rgb images[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 52-67.

[10] Wen C, Zhang Y, Li Z, et al. Pixel2mesh++: Multi-view 3d mesh generation via deformation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1042-1051.

[11] Park J J, Florence P, Straub J, et al. Deepsdf: Learning continuous signed distance functions for shape representation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 165-174.

[12] Pumarola A, Corona E, Pons-Moll G, et al. D-nerf: Neural radiance fields for dynamic scenes[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10318-10327.

[13] Park K, Sinha U, Hedman P, et al. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields[J]. ACM Transactions on Graphics, 2021, 40(6): Article No.238.

[14] Fang J, Yi T, Wang X, et al. Fast dynamic radiance fields with time-aware neural voxels[C]//SIGGRAPH Asia 2022 Conference Papers. 2022: 1-9.

[15] Fridovich-Keil S, Meanti G, Warburg F R, et al. K-planes: Explicit radiance fields in space, time, and appearance[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12479-12488.

[16] Cao A, Johnson J. Hexplane: A fast representation for dynamic scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 130-141.

[17] Yang Z, Gao X, Zhou W, et al. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 20331-20341.

[18] Wu G, Yi T, Fang J, et al. 4d gaussian splatting for real-time dynamic scene rendering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 20310-20320.

[19] Lin Y, Dai Z, Zhu S, et al. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 21136-21145.

[20] Huang Y H, Sun Y T, Yang Z, et al. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 4220-4230.