

Research and Analysis of 3D Reconstruction Technology under the Influence of Deep Learning

Ce Gao^{1,*}

¹School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China

*Corresponding author:
grace03312025@outlook.com

Abstract:

As a key technology connecting the physical world and the digital world, 3D reconstruction has wide-ranging applications in fields such as autonomous driving, virtual reality, and industrial inspection. Traditional methods rely on handcrafted features and geometric constraints, and suffer from limitations like poor robustness and low integrity in complex scenarios. In recent years, deep learning technology, with its powerful feature learning and context modeling capabilities, has brought about revolutionary advancements to 3D reconstruction. This paper systematically reviews the research progress of deep learning-driven 3D reconstruction technology, focuses on analyzing core innovations including fusion architectures, attention mechanisms, and lightweight networks, and conducts an in-depth discussion on the breakthroughs and integration trends of cutting-edge representation technologies such as neural implicit fields and 3D Gaussian splatting. Furthermore, the paper points out that current technologies still face challenges such as generalization capability, dynamic scene processing, and computational overhead, and looks forward to future development directions including unified implicit-explicit modeling, lightweight deployment, and general-purpose 3D vision foundation models. It provides important references for in-depth understanding and promotion of the development of 3D reconstruction technology.

Keywords: 3D Reconstruction; Deep Learning; Neural Implicit Fields; 3D Gaussian Splatting; Simultaneous Localization and Mapping.

1. Introduction

3D reconstruction refers to the technical process of

recovering the 3D geometric structure and appearance attributes of objects and scenes from 2D images or video sequences. As a core bridge connecting the

physical world and the digital world, it has extensive and far-reaching application value in many fields such as autonomous driving, robot navigation, virtual/augmented reality (VR/AR), medical image analysis, industrial inspection, and cultural heritage protection.

Traditional 3D reconstruction methods mainly rely on technical frameworks such as Structure from Motion (SfM) and Multi-View Stereo (MVS). SfM recovers camera parameters and sparse 3D point clouds through technologies like feature matching and Bundle Adjustment, while MVS further achieves dense 3D reconstruction based on principles such as photometric consistency. However, such methods heavily depend on manually designed local features and optimization strategies based on geometric constraints. In challenging scenarios such as low-texture regions, repeated structures, complex occlusions, and drastic lighting changes, they often exhibit inherent limitations including poor robustness in feature matching, low reconstruction integrity, redundant and cumbersome processes, and sensitivity to noise. In addition, traditional methods usually require multi-stage independent optimization, making it difficult to achieve efficient end-to-end reconstruction, which also limits their application in real-time systems.

Deep learning technology has brought revolutionary progress to 3D reconstruction with its powerful data-driven feature learning capability. The end-to-end training method can automatically learn multi-level feature representations from large-scale data, significantly improving matching performance and reconstruction integrity under ill-posed conditions. Advanced architectures such as CNN and Transformer effectively aggregate local and global context information, enhancing the understanding of scene geometry and semantic structure. At the same time, deep learning has promoted the end-to-end integration of the reconstruction process, unifying steps such as feature extraction, matching, and depth estimation within a differentiable framework, and realizing global optimization through backpropagation, which greatly improves reconstruction efficiency and system performance.

Numerous innovative studies have driven the development of this field. Zhou Qinghua integrated the traditional SGM algorithm with deep learning to improve stereo matching accuracy and convergence speed [1]; Feng Yajuan introduced the multi-head self-attention mechanism into MVS to enhance reconstruction robustness in complex scenes [2]; The H-MVSNet developed by Yang Shuo reduces memory consumption to 22.8% of that of traditional methods through a lightweight design [3]; Deng Xin improved the loss function and network structure to enhance the detail performance of monocular reconstruction [4]; Tan Zhen et al. analyzed the application of 3D Gaussian Splatting technology in SLAM [5], demonstrating the po-

tential of real-time high-fidelity reconstruction. Im S et al. proposed DPSNet, which combines the traditional plane sweeping idea with deep learning to achieve end-to-end multi-view stereo depth reconstruction [6]; Wang K et al. developed a single-frame 3D reconstruction method based on structured light and deep CNN [7], providing a new idea for dynamic scene applications.

Despite significant achievements, current research still faces many challenges: insufficient generalization ability of models under unknown categories and extreme conditions, limited efficiency of dynamic scene reconstruction, high computational and memory costs, and dependence on high-quality annotated data. To address these issues, emerging technologies such as Neural Implicit Fields and 3D Gaussian Splatting provide new ideas for achieving both high-precision and high-efficiency reconstruction, and have promoted the cross-integration of 3D reconstruction with tasks such as rendering and semantic understanding.

This paper aims to systematically sort out the progress of deep learning-driven 3D reconstruction technology, focus on analyzing the innovation of core network architectures, discuss the principles, progress, and applications of cutting-edge technologies such as Neural Implicit Fields and 3D Gaussian Splatting, summarize current challenges, and look forward to future research directions, so as to provide references for relevant researchers.

2. Overview of Deep Learning-Based 3D Reconstruction Technology

Traditional 3D reconstruction technologies have significant limitations in complex environments and under high-precision requirements. Their core issues lie in insufficient robustness in feature extraction and matching, poor adaptability to complex environments, as well as complex modeling processes and limited precision. Although traditional methods such as COLMAP perform well in structured scenes, they often result in incomplete reconstruction in textureless regions and complex occlusion scenarios [8].

Deep learning technology has fundamentally changed this traditional landscape, with its core breakthroughs mainly reflected in the following aspects: automatically learning hierarchical features with high discriminability and robustness through deep neural networks, which significantly improves matching performance in ill-posed regions; effectively aggregating local and global contextual information by leveraging advanced architectures such as attention mechanisms and Transformer, which enhances the understanding of scene semantics and geometric structures, thereby reducing ambiguities in the reconstruction process; integrating multiple separate steps in the

traditional reconstruction process into a jointly trainable network through end-to-end optimization, and uniformly optimizing all parameters using backpropagation, which not only simplifies the process but also improves overall performance; in addition, deep learning has also promoted the development of new types of scene representation methods such as Neural Implicit Fields and 3D Gaussian Splatting, achieving significant improvements in multiple dimensions including rendering quality, memory efficiency, and computational speed.

3. Innovations in Core Deep Learning Architectures

3.1 Fusion Architectures of Traditional Algorithms and Deep Learning

Traditional geometric algorithms have advantages in stability and interpretability, and their fusion with deep learning constitutes a key innovative path. The SgmA-net proposed by Zhou Qinghua introduces the disparity estimation values of the traditional SGM algorithm into the network as strong geometric priors. On the KITTI dataset, this reduces the disparity estimation error by 18.7% and decreases the number of convergence epochs by 30% [1]. This method effectively addresses the issue of slow cold-start convergence in deep learning models, demonstrating the value of traditional priors in enhancing network performance. In multi-view stereo matching, Feng Yajuan uses sparse point clouds generated by SfM to initialize depth maps instead of random initialization. On the DTU dataset, this reduces the reconstruction accuracy error by 10.1% and noise by 30% [2]. By leveraging the reliability of traditional SfM, this method improves the accuracy and stability of depth map generation.

3.2 Attention Mechanisms and Feature Interaction Architectures

The attention mechanism effectively resolves issues of feature imbalance and matching ambiguity by dynamically weighting key features. Zhou Qinghua introduced Multi-head Attention into SgmA-net [1], calculating self-attention and cross-attention between left and right views to strengthen feature correlations across views, thereby improving matching accuracy in edge and textureless regions. Feng Yajuan proposed a multi-scale feature extraction network based on multi-head self-attention [2], integrating self-attention mechanisms into each feature scale layer. This enhances the ability to perceive global context and improves the robustness and integrity of large-scene reconstruction. MVSTER, proposed by Wang K et al., utilizes the long-range dependency modeling ca-

pability of Transformer to design a Global Context Transformer and a 3D Geometric Transformer. On the DTU dataset, its overall error reaches 0.326mm, outperforming R-MVSNet's 0.367mm [9]. This method breaks through the receptive field limitation of CNNs and significantly improves matching accuracy in textureless regions.

3.3 Lightweight and Efficiency-Optimized Architectures

Demand for real-time applications has driven the development of lightweight architectures. The H-MVSNet proposed by Yang Shuo adopts a lightweight 2D CNN and Inplace-ABN layers, reducing memory consumption by 50%, and improves the GRU regularization module. On the DTU dataset, its memory consumption is reduced to 2.46GB, which is only 22.8% of MVSNet's, and its computational speed is increased by 36% [3]. This method significantly lowers the resource requirements for high-resolution reconstruction. Feng Yajuan significantly reduced runtime and video memory usage by decreasing the number of feature extraction scale levels while maintaining reconstruction accuracy [2]. Instant-NGP, proposed by Müller T et al., uses multi-resolution hash encoding technology to reduce the training time of neural graphics primitives from hours to seconds [10]. This method stores trainable feature vectors in a multi-resolution hash table, replacing the high-dimensional input encoding of traditional MLPs, which greatly reduces model complexity. NeRF training can achieve a PSNR of 31.4dB in just 15 seconds.

3.4 Innovations in Loss Functions and Training Strategies

Specialized loss functions have been designed for geometric properties. In MVS, Feng Yajuan jointly uses photometric consistency loss, SSIM loss, and depth smoothness loss. Through multi-task optimization, this reduces point cloud noise by 30% and significantly improves edge smoothness [2]. Deng Xin proposed a weighted loss function that fuses Chamfer Distance (CD) and local structure loss [4]. On the ShapeNet dataset, this increases the F-Score by 3.79% compared to the baseline model, enhancing the ability to learn detailed features.

4. Neural Implicit Fields and 3D Gaussian Splatting

Scene representation is the core of 3D reconstruction. In recent years, Neural Implicit Fields and 3D Gaussian Splatting (3DGS) — as two groundbreaking technologies — have advanced the field from two directions: continuous implicit representation and explicit differentiable ren-

dering, respectively.

4.1 Innovations in Loss Functions and Training Strategies

References are cited in the text just by square brackets [1].

(If square brackets are not available, slashes may be used instead, e.g. /2/.) Two or more references at a time may be put in one set of brackets [3, 4]. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading References, see our example below.

Table 1. Performance Comparison of Key Neural Implicit Field Methods

Method Name	Core Innovations	Dataset	Key Metrics	Performance Improvement	Applicable Scenarios
DDN-SLAM [11]	Semantic features + GMM dynamic segmentation; dynamic semantic loss; sparse point cloud-guided sampling	TUM RGB-D, Bonn	ATE: 0.020m (TUM)	Tracking error significantly lower than NICE-SLAM, etc.	SLAM in dynamic environments
MS-Neus [12]	Multi-scale S-density function; adaptive scale selection	DTU MVS	Chamfer Distance: 0.77cm	Outperforms COLMAP (0.93) and NeuS (0.83) [13]	Complex geometric structure reconstruction
Neural Implicit Texture [13]	Joint surface parameterization & texture mapping; inverse mapping network; neural texture map	DTU	PSNR: ~12.21% ↑	Significantly improved texture fidelity; editable	High-fidelity texture reconstruction
GS-NIR [14]	Geometric smoothness constraint; Random Dilatation Sampling (RDS); Adaptive Scene Module (ASM)	Self-collected plant data	Average error: 0.811mm	Accuracy equivalent to professional 3D scanners	High-precision phenotyping measurement
D-NeRF [15]	6D dynamic neural radiance field; canonical space + deformation network	Bouncing Balls, etc.	PSNR: 32.80dB	Far exceeds NeRF (18.28dB) and T-NeRF (32.01dB)	Dynamic scene modeling

As shown in Table 1, neural implicit representation has achieved breakthrough progress in the field of 3D reconstruction in recent years; however, different methods have distinct focuses in terms of innovations, performance, and applicable scenarios.

For high-precision geometric and texture reconstruction of static scenes, MS-Neus and the Neural Implicit Texture method have demonstrated improvements with different emphases [12, 13]. By introducing a multi-scale S-density function and an adaptive scale selection mechanism, MS-Neus effectively enhances the ability to reconstruct complex geometric structures. Its Chamfer Distance of 0.77cm on the DTU dataset outperforms the traditional method COLMAP (0.93cm) and the early neural implicit method NeuS (0.83cm), indicating its advantage in geometric accuracy. Complementarily, the core contribution of the Neural Implicit Texture method lies in improved texture fidelity [13]. Through joint optimization of surface parameterization and texture mapping, along with the design of an inverse mapping network and a neural texture map, it achieves texture reconstruction with higher visual quality and editability, with the PSNR metric increased by

approximately 12.21% — providing high-quality models for downstream applications such as digital twins.

Reconstruction of dynamic scenes is a current research challenge, and DDN-SLAM and D-NeRF have proposed solutions for different applications [11, 15]. DDN-SLAM focuses on solving the problem of real-time localization and mapping in dynamic environments. It innovatively integrates semantic features and Gaussian Mixture Models (GMM) for dynamic object segmentation, supplemented by dynamic semantic loss and sparse point cloud-guided sampling, significantly enhancing the system’s robustness in dynamic environments. On datasets such as TUM RGB-D, its Absolute Trajectory Error (ATE) is as low as 0.020m, far outperforming comparison methods like NICE-SLAM — highlighting its practical value in real-world scenarios such as robot navigation and AR. In contrast, D-NeRF specializes in generating highly realistic dynamic Novel View Synthesis. By constructing a 6D dynamic neural radiance field, it decomposes dynamic sequences into a canonical space and a deformation network, thereby enabling modeling of non-rigid motion. Its PSNR of 32.80dB on the Bouncing Balls dataset far

exceeds that of the original NeRF (18.28dB) and the contemporary method T-NeRF (32.01dB), making it suitable for fields such as film special effects and virtual video production.

In addition, neural implicit methods have begun to extend to specialized, high-precision measurement fields. GS-NIR targets the specific demand of plant phenotyping measurement [14], introducing geometric smoothness constraints, Random Dilatation Sampling (RDS), and an Adaptive Scene Module (ASM) to optimize reconstruction details and accuracy. It achieves an average error of 0.811mm on self-collected datasets, with accuracy comparable to professional 3D scanning equipment — demonstrating the application potential of neural implicit methods in scientific measurement and industrial inspection.

In summary, current research on Neural Implicit Fields shows a comprehensive development trend: from static to dynamic scenes, from general-purpose to specialized tasks, and from geometric modeling to texture reconstruction. Each method has achieved significant performance improvements in its specific task; however, this also reflects that no “one-size-fits-all” model has yet emerged in the field. Future research is expected to focus on fusing these advantages and further addressing core challenges such as computational efficiency, generality, and robustness under extreme conditions.

4.2 Breakthroughs in 3D Gaussian Splatting (3DGS)

Proposed in 2023, 3DGS is a revolutionary technology that explicitly represents scenes using a large number of optimizable 3D Gaussian ellipsoids. Each Gaussian ellipsoid’s parameters include position, rotation, scale, opacity, and color, with rendering implemented via differentiable rasterization.

The notable advantages of this method are as follows: relying on the extreme rendering efficiency of the GPU rasterization pipeline, it can achieve real-time performance of hundreds of frames per second, far surpassing neural implicit representations based on volume rendering; at the same time, it centers on explicit and editable 3D Gaussian primitives, supporting direct geometric operations, scene compression, and dynamic fusion; finally, it still maintains photo-realistic quality in complex geometric and appearance modeling, providing an efficient and flexible solution for realizing high-quality novel view synthesis.

3DGS has been quickly applied to SLAM systems, forming a new technical route. For example, SplaTAM directly uses RGB-D data to optimize camera poses and Gaussian parameters [16], enabling real-time and high-quality dense mapping. It performs outstandingly in textureless scenes with large motions, achieving an ATE RMSE of only 1.2 cm on the ScanNet++ dataset—far lower than

Point-SLAM’s 343.8 cm. Dynamic 3D Gaussians realizes high-fidelity rendering and dense 6DOF tracking of dynamic scenes through the design of Gaussians with invariant attributes and controllable motion [17].

4.3 Technology Comparison and Integration Trends

Neural Implicit Fields and 3D Gaussian Splatting (3DGS) represent two distinct technical pathways, each with its own strengths and weaknesses. The strengths of Neural Implicit Fields lie in their high storage efficiency, ability to generate smooth and continuous surfaces, and ease of representing complex topologies; their weaknesses include the need for ray marching sampling and neural network queries, resulting in slow rendering speed and time-consuming training. Additionally, they act as a “black box” and are difficult to edit directly. By contrast, 3D Gaussian Splatting boasts extremely fast rendering speed, explicit editability, and fast training; however, its memory consumption grows linearly with scene complexity (leading to high memory usage), it struggles to ensure strictly continuous surfaces, and it heavily relies on high-quality SfM initialization.

Currently, the two technologies are showing a trend of integration. Some studies attempt to use more compact Neural Implicit Fields to generate or constrain the distribution of 3D Gaussians, thereby reducing their memory footprint. Other works explore combining the fast rendering capability of 3DGS with the continuous representation capability of implicit fields, leveraging each other’s strengths to offset weaknesses. This implicit-explicit hybrid representation is likely to become an important future development direction, enabling the simultaneous achievement of high precision, high speed, low memory usage, and strong generalization capabilities.

5. Discussion

5.1 Existing Challenges

Despite the significant advancement of 3D reconstruction technology driven by deep learning, the technology still faces multiple challenges in moving toward generalization and practical application: First, models have weak generalization ability and are highly dependent on the distribution of training data. Their performance degrades significantly when dealing with unknown object categories, novel shooting perspectives, or extreme lighting conditions—especially in scenarios with low texture, strong reflections, or severe occlusions, where reconstruction holes and shape distortion are prone to occur. Meanwhile, dynamic scene reconstruction struggles with balancing efficiency and accuracy. Non-rigid motion, complex

occlusions, and other factors invalidate the traditional assumption of static scenes. Although some methods have introduced dynamic segmentation and temporal data modeling, they either fail to meet real-time requirements, suffer from reduced accuracy, or significantly increase computational complexity. Furthermore, high-precision reconstruction is often accompanied by enormous computational and memory costs. High-resolution voxel representation causes video memory to grow cubically, while the rendering speed of Neural Implicit Fields is relatively slow—both of which limit their application in mobile devices or real-time systems. Finally, deep learning models rely on large-scale, high-quality annotated data. However, acquiring 3D ground truth is extremely costly, existing datasets have a narrow coverage range, and the domain gap between simulated data and real scenes further reduces the models' adaptability in practical scenarios. These challenges collectively indicate that current deep learning-based 3D reconstruction technology still needs to achieve breakthroughs in robustness, efficiency, and universality to better promote technology implementation.

5.2 Future Outlook

Future research should continue to focus on several key directions: For example, exploring the in-depth integration and unified modeling of neural implicit and explicit representations. Such hybrid methods are expected to combine the high-detail capability of implicit representations with the computational efficiency of explicit representations, thereby achieving a better balance between speed, accuracy, and memory consumption.

Promoting the collaborative design of lightweight algorithms and dedicated hardware: compressing network scales through techniques such as model pruning, knowledge distillation, and low-precision quantization; and conducting end-to-end optimization based on the computing power characteristics of AI acceleration chips or embedded GPUs. Ultimately, this will enable real-time high-precision reconstruction on edge devices, facilitating the implementation of applications such as mobile AR and robot navigation.

Incorporating physical laws into the reconstruction process: traditional methods mostly rely on geometric and appearance constraints, while introducing physical priors such as optical reflection, material properties, and motion dynamics can enhance the physical consistency and realism of reconstruction results. For instance, the combination of differentiable rendering and physical engines has already shown preliminary potential, and in the future, it can be further used to generate more realistic dynamic scenes.

Finally, constructing general-purpose 3D vision foundation models capable of understanding complex worlds:

learning universal scene representations through self-supervised learning, multi-modal pre-training, and other methods to equip models with zero-shot transfer, scene reasoning, and generation capabilities. This will support a wider range of application scenarios (e.g., autonomous driving, virtual reality, and digital twins), driving the field toward a new stage of development characterized by high efficiency, realism, and generalization.

6. Conclusion

This paper systematically reviews the research progress of deep learning in 3D reconstruction, covering core innovations such as fusion architectures, attention mechanisms, and lightweight design. It also conducts an in-depth analysis of the advantages, limitations, and integration trends of two cutting-edge representation technologies: Neural Implicit Fields and 3D Gaussian Splatting (3DGS). Current research has achieved significant results in improving reconstruction quality, efficiency, and dynamic processing capabilities; however, further breakthroughs are still needed in aspects such as generalization, computational efficiency, and practicality. In the future, through implicit-explicit integration, lightweight deployment, and foundation model construction, 3D reconstruction technology is expected to play a key role in more practical applications.

References

- [1] Zhou Q H. Research on 3D reconstruction based on multiple end-to-end deep learning model. Beijing: Beijing University of Posts and Telecommunications, 2022.
- [2] Feng Y J. Research of MVS 3D Reconstruction Algorithm Based on Deep Learning. Taiyuan: Shanxi University, 2023.
- [3] Yang S. Research and Implementation of 3D Reconstruction Algorithm Based on Deep Learning. Guiyang: Guizhou Normal University, 2022.
- [4] Deng X. Research on Deep Learning Based Monocular 3D Reconstruction. Mianyang: Southwest University of Science and Technology, 2023.
- [5] Tan Z, Niu Z Y, Zhang J P, et al. New opportunities in SLAM-Gaussian splatting technology. *Journal of Image and Graphics*, 2025, 30(6): 1792-1807.
- [6] Im S, Jeon H G, Lin S, Kweon I S. DPSNet: End-to-end Deep Plane Sweep Stereo. arXiv preprint arXiv:1905.00538, 2019.
- [7] Nguyen H, Wang Y Z, Wang Z Y. Single-Shot 3D Shape Reconstruction Using Structured Light and Deep Convolutional Neural Networks. *Sensors*, 2020, 20(13): 3718.
- [8] Schonberger J L, Frahm J M. Structure-from-Motion Revisited. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 4104-4113.
- [9] Zhu J, Peng B, Li W Q, et al. Multi-View Stereo with

Transformer. arXiv preprint arXiv:2112.00336, 2021.

[10] Müller T, Evans A, Schied C, et al. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics*, 2022, 41(4): Article 102.

[11] Li M R, Zhou Y M, Jiang G A, et al. DDN-SLAM: Real-time Dense Dynamic Neural Implicit SLAM. arXiv preprint arXiv:2401.01545, 2024.

[12] Ji X S, Zhu Y, Yang J L, et al. Three-Dimensional Reconstruction Method Based on Multiscale S-Density Strategy. *Laser & Optoelectronics Progress*, 2025, 62(4): 0411002.

[13] Xu Z H, Deng B L, Zhang J Y. Neural Implicit Surface Parameterization and Texture Reconstruction. *Journal of Computer-Aided Design & Computer Graphics*, 2024, 3(*): Article ID 10.3724/SFJ.1089.2024-00606.

[14] Ying W, Hu K W, Ahmed A, et al. Accurate Fruit Phenotype

Reconstruction via Geometry-Smooth Neural Implicit Surface. *Agriculture*, 2024, 14(12): 2325.

[15] Pumarola A, Corona E, Pons-Moll G, et al. D-NeRF: Neural Radiance Fields for Dynamic Scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021: 1-10.

[16] Keetha N, Karhade J, Jatavallabhula K M, et al. SplatTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024: 1-10.

[17] Luiten J, Kopanas G, Leibe B, et al. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, Early Access. DOI: 10.1109/TPAMI.2024.3421209.