# Multi-View Image-Based 3D Reconstruction in Indoor Scenes: A Survey

LU Ping[1,2], SHI Wenzhe[1,2], QIAO Xiuquan[3]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTECOM.202403011

https://kns.cnki.net/kcms/detail/34.1294.TN.20240912.1012.002.html, published online September 12, 2024

Manuscript received: 2023-09-14

**Abstract:** Three-dimensional reconstruction technology plays an important role in indoor scenes by converting objects and structures in indoor environments into accurate 3D models using multi-view RGB images. It offers a wide range of applications in fields such as virtual reality, augmented reality, indoor navigation, and game development. Existing methods based on multi-view RGB images have made significant progress in 3D reconstruction. These image-based reconstruction methods not only possess good expressive power and generalization performance, but also handle complex geometric shapes and textures effectively. Despite facing challenges such as lighting variations, occlusion, and texture loss in indoor scenes, these challenges can be effectively addressed through deep neural networks, neural implicit surface representations, and other techniques. The technology of indoor 3D reconstruction based on multi-view RGB images has a promising future. It not only provides immersive and interactive virtual experiences but also brings convenience and innovation to indoor navigation, interior design, and virtual tours. As the technology evolves, these image-based reconstruction methods will be further improved to provide higher quality and more accurate solutions to indoor scene reconstruction.

**Keywords:** 3D reconstruction; MVS; NeRF; neural implicit surface

**Citation** (Format 1): LU P, SHI W Z, QIAO X Q. Multi-view image-based 3D reconstruction in indoor scenes: a survey [J]. ZTE Communications, 2024, 22(3): 91 – 98. DOI: 10.12142/ZTECOM.202403011
**Citation** (Format 2): P. Lu, W. Z. Shi, and X. Q. Qiao, "Multi-view image-based 3D reconstruction in indoor scenes: a survey," *ZTE Communications*, vol. 22, no. 3, pp. 91 – 98, Sept. 2024. doi: 10.12142/ZTECOM.202403011.

## 1 Introduction

Traditional 3D reconstruction techniques play a crucial role in the field of computer vision and encompass commonly used tools and libraries such as ColMap[1] and OpenMVS[2]. These techniques estimate the 3D positions of points in a scene through feature extraction and matching, followed by steps like sparse point cloud generation and dense point cloud generation to create dense point clouds and 3D models with high-quality geometry and texture. These traditional 3D reconstruction techniques find widespread applications in various domains, including virtual reality, augmented reality, indoor navigation, building reconstruction, and cultural heritage preservation tasks. They perform well in handling small objects, small-scale scenes, and simple scenes. However, when it comes to indoor scenes containing a large number of textureless or repetitive texture regions, traditional reconstruction methods struggle to extract meaningful features, resulting in significant holes and noise during the reconstruction process. This limitation restricts their applicability in large-scale and complex scenes.

With the continuous advancement of technology, there is increasing attention on deep learning-based 3D reconstruction techniques for indoor scenes[3 – 5]. Compared to traditional 3D reconstruction methods, these techniques leverage the power of deep learning models in processing and analyzing multiple-view RGB images to extract feature representations of the scene, leading to high-quality 3D model reconstruction. Specifically, convolutional neural networks (CNNs), Transformer[6], and vision Transformers (ViT)[7] are used to process and analyze multiple-view RGB images. These models extract feature representations of objects in the scene. Through learning and training, these deep learning models gain an understanding of the geometric shapes, texture information, and other visual cues present in the images, encoding them as feature representations. Once the feature representations of objects are obtained, 3D reconstruction will be conducted. This includes estimating the 3D coordinates of points in the scene and performing optimization and refinement to obtain accurate 3D reconstruction results.

In recent years, there has been rapid development in combining neural radiance fields (NeRF)[8] with neural implicit surface-

**This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20221102002.**

based techniques for the 3D reconstruction of indoor scenes. This development is driven by the remarkable success of NeRF[8] in the field of novel view synthesis. These methods utilize neural network models to learn implicit representation functions of objects, enabling the capture of their geometric shapes and texture information without the explicit representation of 3D geometry. Compared to explicit representations, such as point clouds and voxel grids, these methods offer higher-resolution 3D reconstruction with enhanced expressive power and generalization capabilities. They can handle complex scenes and shape variations by leveraging the flexibility of neural networks in modeling intricate geometric details and capturing rich texture information. By learning implicit representations of objects, these techniques allow for more accurate and detailed reconstructions, even in the presence of challenging factors like occlusions and varying lighting conditions. The implicit nature of the representation enables these methods to generate novel views of the scene from previously unseen viewpoints, contributing to their growing success in the field of indoor scene 3D reconstruction.

In summary, the use of multi-view RGB images for indoor scene 3D reconstruction offers a new approach to achieving low-cost and high-quality reconstructions. It not only reduces equipment requirements and operational complexity but also presents broad application prospects in various fields, including virtual reality, augmented reality, indoor navigation, and game development. However, existing techniques also face several challenges. Traditional methods demonstrate stability in handling small objects, small-scale scenes, and simple scenes but encounter difficulties in dealing with textureless regions and repetitive texture regions in indoor scenes. On the other hand, deep learning-based techniques for indoor scene 3D reconstruction, mostly relying on computationally expensive 3D CNNs[9] or structures like Transformer[6], require processing a large volume of image data for large-scale indoor scenes. Additionally, generating high-resolution voxel grids consumes significant storage resources, posing challenges in terms of computational resources and storage space. Furthermore, capturing details and handling variations in complex scenes and shape changes demand more complex and robust deep learning models, along with considerable time and effort for data annotation and model training. Inspired by the tremendous success of NeRF[8] in the field of novel view synthesis, many techniques combining neural implicit surface-based methods with NeRF have rapidly developed for indoor scene 3D reconstruction. These methods employ neural network models to learn implicit representation functions of objects, enabling the capture of geometric shapes and texture information without explicit representation of 3D geometry. Compared to explicit representation methods, neural implicit surface-based approaches achieve higher-resolution 3D reconstruction, possess stronger expressive power, and can handle complex scenes and shape variations.

Future trends in the development of indoor scene 3D reconstruction include enhancing the robustness of traditional methods and their adaptability to handle textureless scenes. Efforts will be made to strengthen the research on data utilization efficiency and generalization capabilities in deep learning-based approaches. Additionally, exploring performance improvements of neural implicit surface-based methods in complex scenes will be a focus. As technology evolves and innovates, multi-view-based indoor 3D reconstruction techniques will continue to provide more accurate and realistic ways of generating 3D scene content for various fields. These advancements will enable the creation of highly precise and realistic 3D scenes, benefiting applications in virtual reality, augmented reality, indoor navigation, and game development, among others.

## 2 Multi-View-Based Indoor 3D Reconstruction

The multi-view-based indoor 3D reconstruction techniques are of significant importance in the field of computer vision. Existing methods can be broadly classified into three categories. The first category is the traditional 3D reconstruction methods based on feature matching, which recovers the 3D structure of the scene by extracting feature points from images and performing feature matching. Methods such as ColMap[1] and OpenMVS[2] utilize techniques like feature point matching, camera pose estimation, and triangulation to achieve sparse and dense 3D reconstruction. These methods have shown good results for indoor scenes, but they perform poorly in scenarios with textureless or low-texture regions.

The second category is the deep learning-based 3D reconstruction techniques, which directly learn the 3D representation of the scene from multi-view image data. These methods can extract rich semantic and geometric information from images and achieve end-to-end 3D reconstruction with good generalization. Examples of such methods include Pixel2Mesh[10], SimpleRecon[11], and NeuralRecon[12], which have achieved significant advancements in indoor scene reconstruction and can generate high-quality 3D models.

The third category is the 3D reconstruction techniques based on neural implicit surface representation, where the implicit representation function of the object is learned by neural networks, eliminating the need for explicit representation of 3D geometry. These methods can handle complex scenes and shape variations and generate highly accurate 3D models. Examples of such methods include VolSDF[13] and NeuS[14], which have made notable progress in indoor scene reconstruction, enabling high-fidelity geometry and texture reconstruction. By leveraging image data from multiple viewpoints, these methods can reconstruct the 3D structure and texture information of indoor scenes, providing a foundation for applications such as indoor navigation, virtual reality, and augmented reality.

### 2.1 Stereoscopic Matching-Based 3D Reconstruction Methods

Traditional stereoscopic matching-based 3D reconstruction methods, such as ColMap[1] and OpenMVS[2], play a significant role in indoor 3D reconstruction. ColMap integrates Structure-from-Motion (SfM) and Multi-View Stereo (MVS) technologies, while OpenMVS[2] focuses specifically on MVS. These tools possess strong capabilities and have become standard tools in both academia and industry for generating high-quality 3D models from multi-view RGB images. However, these traditional methods still face several challenges. They perform poorly in dealing with large areas devoid of textures, which often leads to feature matching failures and consequently impacts reconstruction accuracy. Additionally, these methods require significant computational resources and storage space, especially when handling large-scale scenes, which can be limiting factors. Furthermore, these traditional methods have limitations when dealing with closed and transparent surfaces, as well as scenes with complex textures and fine geometric structures. Therefore, researchers are continuously working to enhance these methods to improve their performance and robustness, particularly in dealing with complex scenes and textureless regions. Future research directions may include better addressing these challenges to enable traditional feature matching methods to be more effective in a wider range of application scenarios.

Fig. 1 shows the workflow of matching-based 3D reconstruction method. These methods typically involve the following detailed procedures and techniques.

### 2.1.1 Image Acquisition and Feature Extraction

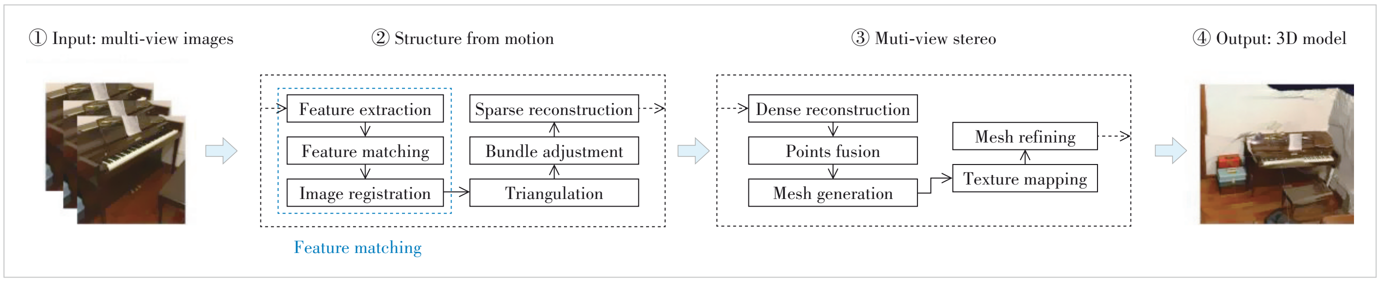
Multiple images are captured from different viewpoints using cameras or a camera system. For each image, feature extraction algorithms (e.g., SIFT[15], SURF[16], or ORB[17]) are used to detect and describe key points and feature descriptors in the images. For example, in ColMap[1], the input images are preprocessed by converting them to grayscale to reduce computational complexity and improve robustness. Then, histogram equalization is applied to enhance contrast and details. Next, the classic Scale-Invariant Feature Transform (SIFT) algorithm[15] is used with various optimizations and improvements to extract stable and discriminative key points and compute descriptive feature descriptors. These extracted key points and descriptors are essential input data for subsequent 3D reconstruction tasks. ColMap[1] also provides features such as GPU acceleration, visualization, and debugging tools to enhance the efficiency and convenience of feature extraction.

### 2.1.2 Camera Pose Estimation

Camera pose estimation, a crucial step in 3D reconstruction, is used to estimate the position and orientation of cameras in the world coordinate system. Feature matching algorithms (e.g., based on feature descriptors or optical flow) are used to match the feature points between different images. Through geometric verification and filtering of the matched point pairs, the relative camera poses and parameters, i.e., camera pose estimation, can be obtained. For example, OpenMVG[18] provides a selection of feature extraction algorithms based on specific application requirements. It estimates camera rotation and translation through fundamental matrix estimation and pose recovery. Fundamental matrix estimation calculates the fundamental matrix between two views from the results of feature matching, and then camera rotation and translation can be inferred by solving the fundamental matrix. To further improve reconstruction accuracy and stability, OpenMVG[18] also offers camera network optimization. By performing bundle adjustment and other global optimization algorithms, the poses of all cameras can be optimized to minimize reprojection errors and maintain consistency. Camera network optimization helps correct errors in feature matching and camera pose estimation, resulting in more accurate reconstruction results.

### 2.1.3 3D Point Cloud Generation and Reconstruction

3D point cloud generation and sparse/dense reconstruction are key steps in 3D reconstruction, and popular toolkits like ColMap[1], OpenMVG[18], and OpenMVS[2] provide corresponding functionalities. These tools can generate 3D point clouds from multi-view images, converting the feature points in the images into 3D points through feature extraction, feature matching, and camera pose estimation. This process produces a sparse point cloud, representing the geometric structure of the scene with only a small number of key points. Subsequently, dense point cloud generation is performed by interpolating or optimizing to fill the gaps between sparse points, resulting in a denser point cloud with richer details. ColMap[1] employs various feature extraction and matching algorithms and offers Multi-View Geometry (MVG) and incremental sparse reconstruction algorithms. OpenMVG[18] focuses more on geometry recovery and camera pose estimation using algorithms such as fundamental matrix estimation and bundle adjustment. OpenMVS[2] is dedicated to dense reconstruction, utilizing multi-view stereo-



▲Figure 1. Workflow of matching-based 3D reconstruction methods scopic matching algorithms to generate a denser point cloud.

Finally, through steps such as triangulation, mesh optimization, and texture mapping, it is possible to generate a 3D mesh model with continuous surfaces and textures. Different mesh generation algorithms and techniques can be selected and optimized according to application requirements to obtain high-quality mesh models.

## 2.2 Deep Learning-Based 3D Reconstruction Methods

Traditional 3D reconstruction methods based on deep learning, such as SimpleRecon and NeuralRecon, represent the application of deep learning techniques in the field of 3D reconstruction, offering numerous innovations and advantages. These methods harness the powerful capabilities of deep learning, leveraging structures like CNNs to infer depth and geometric information from multi-view RGB images. Approaches like SimpleRecon and NeuralRecon utilize computationally expensive deep learning structures, including 3D CNNs, to model the three-dimensional geometry of scenes, enabling them to produce high-quality 3D reconstruction results. While these methods often require substantial training data and model training time, they excel in handling complex scenes and scenarios with shape variations. Compared to traditional feature-based methods, deep learning-based 3D reconstruction methods exhibit higher levels of automation and robustness. They can overcome some of the limitations of feature-based methods in textureless and repetitive texture regions and can handle more complex scenes and geometric structures. These methods often benefit from related work in deep learning, such as 3D point cloud processing and image semantic segmentation. Despite showing immense potential in 3D reconstruction, deep learning-based methods still face several challenges. Firstly, these methods often require significant computational resources, especially when generating high-resolution voxel grids. Secondly, model training demands substantial computational resources and time, typically relying on a large amount of annotated data. Additionally, these methods may be susceptible to issues like motion blur and discontinuities when dealing with complex scenes, particularly in the presence of dynamic objects or camera motion. Furthermore, compared to traditional methods, these methods may be less sensitive to scene details and textures. Therefore, future research directions may include improving the efficiency and generalization capabilities of models and enhancing their ability to handle complex scenes and dynamic objects. Attention should also be given to self-supervised and unsupervised training methods for deep learning models to reduce reliance on extensive annotated data.

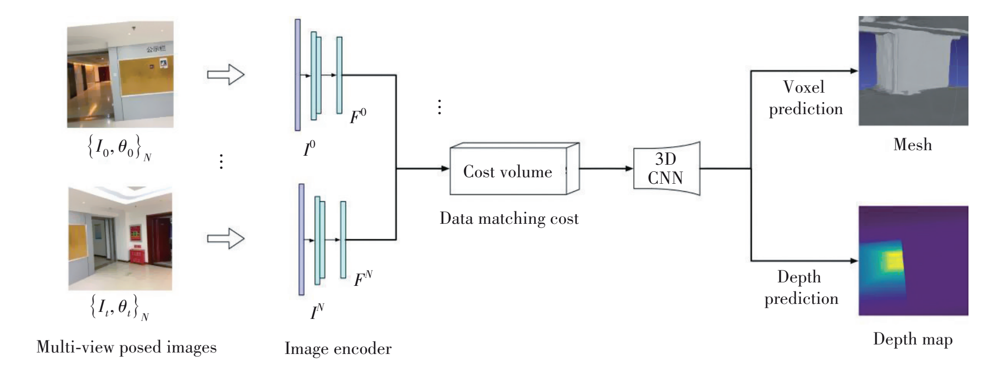
Fig. 2 shows the workflow of convolution-based 3D reconstruction techniques, which mainly consists of feature extraction, cost volume construction, and cost volume regularization. These steps leverage the powerful capabilities of deep learning and convolutional neural networks to achieve accurate reconstruction and depth estimation from images, with wide application prospects in computer vision, robotics, augmented reality, and other fields.

### 2.2.1 Feature Extraction

Feature extraction is the first step in 3D reconstruction, aiming to extract useful features from input images. CNNs are commonly used as feature extraction networks, with popular network architectures including ResNet[19] and U-Net[20]. These networks can extract local and global features from images and provide more representative feature representations for subsequent steps. In the feature extraction process, techniques such as fusing prior information and hierarchical convolution can be used to integrate features from different modalities and scales. For example, SimpleRecon[11] combines pose, geometry features, and depth image features through convolutional feature extraction, while NeuralRecon[12] utilizes distance priors obtained from the MVS process to ensure accuracy in texture-rich and edge regions, and normal priors to preserve completeness in texture-lacking regions.

### 2.2.2 Cost Volume Construction

The cost volume is a key concept in 3D reconstruction and is used to represent the similarity of image matching under different depth hypotheses. The basic idea of cost volume construction is to use a plane sweeping algorithm to project the source images onto parallel planes of a reference camera frustum and compute the similarity among the projected images. This process can be achieved through pairwise image matching and view aggregation. The construction of the cost volume can effectively filter out reliable depth hypotheses and provide a basis for subsequent depth estimation. Since the disparity values are in pixel units, this task becomes a classification problem, where



CNN: convolutional neural network

▲Figure 2. Overall structure of learning-based methods each class represents a discretized disparity value. Generally, CNNs can produce more reliable results. For MVS, the methods for generating the cost volume are mainly divided into two categories. For example, MVSNet[21] applies variance to all feature vectors to construct the cost volume, while DPSNet[22] concatenates features pair by pair and averages all N-1 volumes to obtain the final cost volume.

### 2.2.3 Cost Volume Regularization

The purpose of cost volume regularization is to predict relatively accurate depth values based on aggregated features and smooth and refine the cost volume to generate high-quality depth maps. Common methods for cost volume regularization include 3D CNN-based neural networks, such as those used in Atlas[23], recurrent neural networks (RNNs) as in DHC-RMVSNet[24], and a coarse-to-fine aggregation strategy used in NeuralRecon[12]. Among them, 3D CNN can aggregate local and global features across all dimensions but requires higher computational cost; RNN reduces memory consumption by sequentially processing each depth hypothesis; the coarse-to-fine strategy improves the accuracy and details of the depth map through multiple stages of prediction and refinement.

In summary, the workflow of convolution-based 3D reconstruction includes three key steps: feature extraction, cost volume construction, and cost volume regularization. Through these steps, features can be extracted from input images, cost volumes can be constructed, and depth estimation and reconstruction can be performed using the cost volumes, thereby achieving a complete 3D reconstruction process.

## 2.3 Neural Implicit Surface-Based 3D Reconstruction Methods

Neural implicit surface-based 3D reconstruction methods (Fig. 3), such as NeuS and nvdiffrec, represent cutting-edge technology in the field of 3D reconstruction. These methods, which involve learning the implicit representation of object surfaces through neural networks, have made significant advancements. For instance, NeuS introduces a novel volumetric rendering method by training neural signed distance function (SDF) representations, achieving high-quality 3D reconstructions. Nvdiffrec adopts the marching tetrahedra algorithm to generate higher-quality mesh models. However, neural implicit surface-based methods still face some challenges, including accurately capturing boundary information in complex scenes, handling dynamic objects and camera motion and enhancing their ability to process details and textures. In the future, the development of these methods may focus on improving the robustness of models, expanding their applicability to a wider range of scenarios and enhancing their ability to handle details and boundary information. These methods have opened up new possibilities in the field of 3D reconstruction and hold potential for future research and applications.

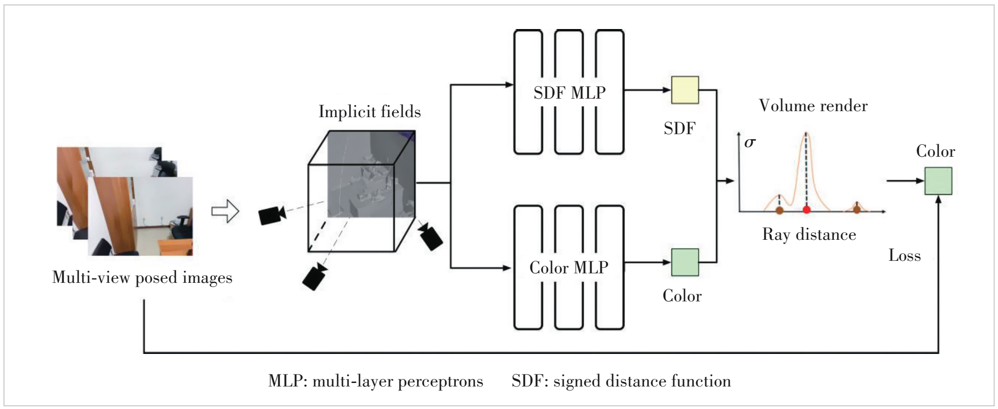
Inspired by the density-based volume rendering algorithm in NeRF[8], significant progress has been made in combining neural implicit surface representation with volume rendering in 3D reconstruction. Neural implicit surface-based 3D reconstruction methods learn the implicit representation function of objects through neural networks and project the reconstructed models to pixel space through volume rendering for training optimization, ultimately achieving high-quality 3D reconstruction.

### 2.3.1 Neural Implicit Surface Representation

Most neural implicit surface representation methods[13–14] model the surface of the target object or scene using two functions. The first function $f: R^3 \to R$ converts spatial coordinates into signed distances from the point to the object surface, where the object's surface is represented by the zero level set of SDF, as shown in Eq. (1).

$$S = \left\{ x \in R^3 | f(x) = 0 \right\}. \tag{1}$$

The other function $c: R^3 \times S^2 \to R^3$ encodes pixel colors related to spatial coordinates and viewing directions. Two multilayer perceptrons (MLPs) are used to approximate these two functions. A new volume rendering method is developed in NeuS[14] to train the neural SDF representation. YARIV et al.[13] improved the geometric representation and reconstruction in neural volume rendering by modeling the volume density as a function of geometry, as arbitrary level set extraction of the density function may lead to low-fidelity reconstruction.



▲Figure 3. Overflow of neural implicit surface-based methods

To address the presence of numerous large planar surfaces and weakly textured areas in indoor scenes, some methods also add constraints to the loss function to constrain the training results and produce smoother surface representations. For example, MonoSDF[25] uses depth and normal maps to constrain the reconstruction results and eliminate noise and discontinuities in the reconstructed surface through normal consistency.

### 2.3.2 Surface Reconstruction

The generated neural implicit field describes the density and color information of each spatial point in the scene. However, to further analyze and visualize the reconstruction results, it is necessary to extract the surface geometry information of objects from the model and convert the continuous object geometry into a discrete voxel representation. Voxel sampling divides the 3D space into a set of small cubic units. By inputting the coordinates of each voxel vertex into an MLP network, the implicit function value of the voxel can be obtained, which records whether it is inside or outside the object and the distance to the object's surface. Specifically, to begin the surface reconstruction process, a three-dimensional grid is first defined within the 3D space. Typically, this grid takes the form of a cube or voxel grid. This grid serves as the basis for point sampling, facilitating the computation of the distance from each point to the object's surface. For each point within this grid, its coordinates are fed into a pre-trained neural network. The neural network then produces an output representing the distance from the point to the object's surface, which is the output of the neural implicit surface function. This distance value is used to determine whether each point is located on the object's surface, with points on the surface generally having a distance of zero or very close to zero. Consequently, the sign of the distance value can be employed to detect points situated on the surface. After identifying the surface points, triangles or other polygons can be generated by connecting these points, thus reconstructing the geometric shape of the object's surface. By traversing the voxel grid and using the implicit function values for interpolation, a continuous geometric surface can be generated. For example, the Marching Cubes[26] algorithm, based on the idea of isosurface extraction, converts the continuous density field into a discrete 3D grid representation to obtain the surface geometry of the object. Additionally, many studies have employed optimized voxelization methods to generate smoother surface representations, such as using the marching tetrahedra algorithm[27] instead of the Marching Cubes[26] algorithm to generate high-quality mesh models in Nvdiffrec[28].

### 2.3.3 Texture Rendering

To present more realistic object details and achieve more convincing visual effects, existing neural implicit surface-based reconstruction techniques use various methods to model the color, texture, material, and lighting information on the object's surface. For example, in Nvdiffrec[28], a coordinate-based network is used to achieve a compact representation of volume textures, and environmental lighting segmentation and an approximate differential formula are introduced to efficiently recover full-frequency lighting. The output triangle mesh, along with spatially varying materials and environmental lighting, can be directly viewed in any traditional graphics engine. Another example is BakedSDF[29], which bakes the implicit scene representation into a high-quality triangle mesh and then designs a view-dependent appearance model based on spherical Gaussians. This approach generates models that can be used for real-time view synthesis using accelerated polygon rasterization pipelines on commodity hardware.

The implicit texture generation process using xAtlas[30] involves several steps as follows:

1) Surface data generation: Initially, the 3D model's surface data, including geometric information but excluding texture information, is generated based on the implicit surface neural radiance field.

2) UV mapping: xAtlas[30] is then used to perform UV mapping, which associates texture coordinates with the surface of the 3D model. UV mapping is a 2D coordinate system commonly used to map texture images onto the surface of 3D objects. In this step, xAtlas calculates UV coordinates for each vertex of the triangular mesh, ensuring proper texture mapping onto the model.

3) Texture Atlas packing: This step involves packing multiple texture images into a single large texture map, which reduces the number of texture switches during rendering and enhances rendering performance. To efficiently allocate texture space on the texture image, xAtlas[30] subdivides the triangular mesh into multiple regions, each having its own UV space. The size and shape of these regions are determined based on surface characteristics to ensure even texture allocation.

4) Optimizing texture layout: Once the UV subdivision is completed, xAtlas[30] determines the texture layout for each region based on their shapes and sizes. This optimization aims to minimize empty areas and wasted texture space.

5) Combining texture images: Finally, xAtlas[30] combines the texture images from these regions into a single large texture map and generates a new UV mapping that correctly maps each triangle on the 3D model's surface to the appropriate texture region. This new UV mapping and the merged texture images are fine-tuned through trainable parameters to produce the final texture map.

In summary, the process involves generating a UV mapping that links the 3D model's surface to a set of texture regions, packing multiple texture images into a single large texture map, and optimizing the layout of these texture regions to minimize waste. This results in a final texture map that can be applied to the 3D model for rendering with texture information.

## 3 Existing Problems

Existing indoor 3D reconstruction techniques based on multi-view RGB images have achieved good reconstruction results in certain situations and within a certain range. However, there are still several problems when it comes to reconstructing complex indoor scenes.

Traditional feature-based 3D reconstruction methods face difficulties in dealing with a large number of textureless areas and repetitive texture areas in indoor scenes, resulting in holes and noise in the reconstruction results. This limits the application of traditional methods in large-scale and complex scenes.

Deep learning-based 3D reconstruction methods typically use computationally expensive structures such as 3D CNNs or transformers, requiring processing a large amount of image data. Generating high-resolution voxel grids also consumes significant storage resources, posing challenges in terms of computational resources and storage space. Additionally, complex scenes and shape variations require more complex and robust deep learning models to capture details and handle changes. Furthermore, deep learning-based methods often require a large amount of annotated data and model training time, which demands substantial computational resources and time.

3D reconstruction methods based on neural implicit surface representation demonstrate better performance in handling complex scenes and shape variations. However, in the absence of boundary information, such as in large scenes or under low lighting conditions, neural implicit surface representation methods may struggle to accurately capture the boundary information of the scene, leading to blurry or incomplete reconstruction results. Moreover, when there are dynamic objects or camera motion in the scene, methods based on neural implicit surface representation may be affected by motion blur. The movement of dynamic objects can result in the discontinuity of point cloud or voxel grid data, which in turn affects the learning of implicit surface functions and the accuracy of reconstruction results. Furthermore, complex indoor scenes typically contain rich details and structures, such as furniture, decorations, and complex textures. Methods based on neural implicit surface representation may struggle to capture these details when dealing with complex scenes, leading to a decrease in the level of reconstruction detail.

## 4 Future Directions

With the rapid development of deep learning and neural networks, 3D reconstruction methods based on neural implicit surface representation are continuously evolving and improving. The future research directions include:

1) Fusion of stereovision and deep learning: Combining traditional stereovision methods with deep learning techniques implements deep learning models for feature representation and matching, improving the reconstruction results for textureless and repetitive texture areas. For example, CNNs can be used to extract features, followed by the integration of traditional stereovision matching methods for geometric constraint optimization.

2) Adaptive reconstruction algorithms: Adaptive reconstruction algorithms that leverage reinforcement learning methods are developed to learn the optimal reconstruction strategies through interaction with the environment. The algorithm's parameters and strategies can be adjusted based on the complexity and characteristics of the scene. This aims to enhance the robustness and effectiveness of reconstruction, improving reconstruction efficiency for simple scenes while maintaining high-quality reconstruction results for complex scenes.

3) Fusion of cross-modal data: RGB images are combined with other data sources such as depth maps, normal maps and metadata to provide a more comprehensive and diverse information source. Consideration can also be given to incorporating semantic information into the 3D reconstruction process to improve the semantic consistency and accuracy of the reconstruction results. Techniques such as semantic segmentation and object detection can guide the model to better understand and model different objects and scenes during the reconstruction process.

4) Generalization capability of models: Current neural implicit surface reconstruction methods typically require a large amount of training data and need to be retrained for different objects and scenes. The future research direction is to improve the generalization capability of models, enabling them to learn and reconstruct different objects and scenes from limited data and handle complex situations such as different lighting conditions, dynamic scenes, and camera motion. It is also necessary to construct larger and more diverse indoor scenes.

## 5 Conclusions

This paper presents three major categories of methods for indoor 3D reconstruction using multi-view RGB images: traditional methods based on feature matching, deep learning-based methods, and methods based on neural implicit surfaces. The specific workflows and development status of each method are described, and the existing issues of current methods are analyzed. Future directions are proposed to guide the future development of 3D reconstruction in indoor scenes.

### References

[1] FISHER A, CANNIZZARO R, COCHRANE M, et al. ColMap: a memory-efficient occupancy grid mapping framework [J]. Robotics and autonomous systems, 2021, 142: 103755. DOI: 10.1016/j.robot.2021.103755

[2] CERNEA D. OpenMVS: multi-view stereo reconstruction library [EB/OL]. [2023-05-20]. https://cdcseacave. github. io/openMVS

[3] YANG J Y, MAO W, ALVAREZ J M, et al. Cost volume pyramid based depth inference for multi-view stereo [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 4876 – 4885. DOI: 10.1109/CVPR42600.2020.00493

[4] WEI Z Z, ZHU Q T, MIN C, et al. AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network [C]//Proc. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 6167 – 6176. DOI: 10.1109/ICCV48922.2021.00613

[5] LU P, SHENG B, SHI W Z. Scene visual perception and AR navigation applications [J]. ZTE communications, 2023, 21(1): 81 – 88. DOI: 10.12142/ZTECOM.202301010

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017: 30

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL]. (2020-01-

22)[2023-05-20]. http://arxiv.org/abs/2010.11929

[8] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 405 – 421. DOI: 10.1007/978-3-030-58452-8_24

[9] MATURANA D, SCHERER S. VoxNet: a 3D Convolutional Neural Network for real-time object recognition [C]//Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015: 922 – 928. DOI: 10.1109/IROS.2015.7353481

[10] WANG N Y, ZHANG Y D, LI Z W, et al. Pixel2Mesh: generating 3D mesh models from single RGB images [C]//European Conference on Computer Vision. Cham: Springer, 2018: 55 – 71. DOI: 10.1007/978-3-030-01252-6_4

[11] SAYED M, GIBSON J, WATSON J, et al. SimpleRecon: 3D reconstruction without 3D convolutions [M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 1 – 19. DOI: 10.1007/978-3-031-19827-4_1

[12] SUN J M, XIE Y M, CHEN L H, et al. NeuralRecon: real-time coherent 3D reconstruction from monocular video [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 15593 – 15602. DOI: 10.1109/CVPR46437.2021.01534

[13] YARIV L, GU J T, KASTEN Y, et al. Volume rendering of neural implicit surfaces [C]//Proc. 35th International Conference on Neural Information Processing System. NIPS, 2021: 4805 – 4815

[14] WANG P, LIU L J, LIU Y, et al. NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction [EB/OL]. (2020-06-20) [2023-05-20]. http://arxiv.org/abs/2106.10689

[15] NG P C, HENIKOFF S. SIFT: predicting amino acid changes that affect protein function [J]. Nucleic acids research, 2003, 31(13): 3812 – 3814. DOI: 10.1093/nar/gkg509

[16] BAY H, ESS A, TUYTELAARS T, et al. Speeded-up robust features (SURF) [J]. Computer vision and image understanding, 2008, 110(3): 346 – 359. DOI: 10.1016/j.cviu.2007.09.014

[17] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF [C]//Proc. International Conference on Computer Vision. IEEE, 2011: 2564 – 2571. DOI: 10.1109/ICCV.2011.6126544

[18] MOULON P, MONASSE P, PERROT R, et al. OpenMVG: open multiple view geometry [M]//KERAUTRET B, COLOM M, MONASSE P, eds. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017: 60 – 74. DOI: 10.1007/978-3-319-56414-2_5

[19] WU Z F, SHEN C H, VAN DEN HENGEL A. Wider or deeper: revisiting the ResNet model for visual recognition [J]. Pattern recognition, 2019, 90: 119 – 133. DOI: 10.1016/j.patcog.2019.01.006

[20] BARKAU R L. UNET: one-dimensional unsteady flow through a full network of open channels. User's manual [R]. Hydrologic Engineering Center Davis CA, 1996

[21] YAO Y, LUO Z X, LI S W, et al. MVSNet: depth inference for unstructured multi-view stereo [C]//European Conference on Computer Vision. Springer, 2018: 785-801. DOI: 10.1007/978-3-030-01237-3_47

[22] IM S, JEON H G, LIN S, et al. DPSNet: end-to-end deep plane sweep stereo [EB/OL]. (2019-05-02)[2023-05-06]. http://arxiv.org/abs/1905.00538

[23] MUREZ Z, VAN AS T, BARTOLOZZI J, et al. Atlas: end-to-end 3D scene reconstruction from posed images [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 414 – 431. DOI: 10.1007/978-3-030-58571-6_25

[24] YAN J F, WEI Z Z, YI H W, et al. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking [C]//16th European Conference on Computer Vision. Cham: Springer, 2020: 674-689.10.1007/978-3-030-58548-8_39

[25] YU Z H, PENG S Y, NIEMEYER M, et al. MonoSDF: exploring monocular geometric cues for neural implicit surface reconstruction [EB/OL]. (2022-06-01)[2023-05-20]. http://arxiv.org/abs/2206.00665

[26] LORENSEN W E, CLINE H E. Marching cubes: a high resolution 3D surface construction algorithm [J]. ACM SIGGRAPH computer graphics, 1987, 21(4): 163 – 169. DOI: 10.1145/37402.37422

[27] SHEN T, GAO J, YIN K, et al. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis [J]. Advances in Neural Information Processing Systems, 2021, 34: 6087-6101.

[28] MUNKBERG J, CHEN W Z, HASSELGREN J, et al. Extracting triangular 3D models, materials, and lighting from images [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 8270 – 8280. DOI: 10.1109/CVPR52688.2022.00810

[29] YARIV L, HEDMAN P, REISER C, et al. BakedSDF: meshing neural SDFs for real-time view synthesis [EB/OL]. (2022-06-01)[2023-05-20]. http://arxiv.org/abs/2302.14859

[30] FAREK J, HUGHES D, SALERNO W, et al. xAtlas: Scalable small variant calling across heterogeneous next-generation sequencing experiments [J]. GigaScience, 2023, 12: giac125. DOI: 10.1093/gigascience/giac125

## Biographies

**LU Ping** (lu.ping@zte.com.cn) is the deputy president of ZTE Corporation, where he is also the general manager of the Industrial Digitalization Solution Dept., and the executive deputy director of State Key Laboratory of Mobile Network and Mobile Multimedia Technology. His research interests include cloud computing, big data, augmented reality, and multimedia service-based technologies. He has supported and participated in multiple major national science and technology projects and national science and technology support projects. He has published multiple papers and authored two books.

**SHI Wenzhe** (shi.wenzhe@zte.com.cn) is a strategy planner of ZTE Corporation where he is also an engineer for XRExplore Platform Product Planning and a member of the National Key Laboratory for Mobile Network and Mobile Multimedia Technology. His research interests include indoor visual AR navigation, SFM 3D reconstruction, visual SLAM, real-time cloud rendering, VR, and spatial perception.

**QIAO Xiuquan** is currently a full professor with Beijing University of Posts and Telecommunications, China, where he is also the deputy director of the Network Service Foundation Research Center, State Key Laboratory of Networking and Switching Technology. He has authored or co-authored over 60 technical papers in international journals and at conferences, including the *IEEE Communications Magazine*, *Proceedings of IEEE*, *Computer Networks*, *IEEE Internet Computing*, *IEEE Transactions on Automation Science and Engineering*, and *ACM SIGCOMM Computer Communication Review*. His current research interests include the future Internet, services computing, computer vision, distributed deep learning, augmented reality, virtual reality, and 5G networks. Dr. QIAO was a recipient of the Beijing Nova Program in 2008 and the First Prize of the 13th Beijing Youth Outstanding Science and Technology Paper Award in 2016. He served as the associate editor for *Computing* (Springer) and the editor board of *China Communications*.