J. Eur. Opt. Society-Rapid Publ. 2024, **20**, 18
© The Author(s), published by EDP Sciences, 2024
https://doi.org/10.1051/jeos/2024018
Available online at: https://jeos.edpsciences.org

Journal of the European Optical Society-Rapid Publications

REVIEW ARTICLE OPEN ACCESS

Overview of image-based 3D reconstruction technology

Yuandong Niu[1], Limin Liu[1,*], Fuyu Huang[1], Siyuan Huang[2], and Shuangyou Chen[1]

[1] Shijiazhuang Campus, Army Engineering University of PLA, Shijiazhuang 050003, PR China
[2] 32398 units of PLA, Beijing 100192, PR China

Received 9 December 2023 / Accepted 8 April 2024

Abstract. Three-dimensional (3D) reconstruction technology is the key technology to establish and express the objective world by using computer, and it is widely used in real 3D, automatic driving, aerospace, navigation and industrial robot applications. According to different principles, it is mainly divided into methods based on traditional multi-view geometry and methods based on deep learning. This paper introduces the above methods from the perspective of three-dimensional space representation. The feature extraction and stereo matching theory of traditional 3D reconstruction methods are the theoretical basis of 3D reconstruction methods based on deep learning, so the paper focuses on them. With the development of traditional 3D reconstruction methods and the development of deep learning related theories, the explicit deep learning 3D reconstruction method represented by MVSNet and the implicit 3D reconstruction method represented by NeRF have been gradually developed. At the same time, the dataset and evaluation indicators for 3D reconstruction were introduced. Finally, a summary of image based 3D reconstruction was provided.

Keywords: 3D reconstruction, Multi-view geometry, Deep learning.

1 Introduction

3D reconstruction technology mainly obtains data information such as laser point cloud and images through laser scanning or cameras, then analyzes and processes the obtained data information, and uses 3D reconstruction related methods to model and reproduce scenes in the real world. It has been widely used in the fields of 3D real scene [1–3], digital twin [4–6], virtual reality [7, 8], artificial intelligence [9–11], automatic driving [12–14], Indoor Robots [15–18], Outdoor Robots [19–23], UAV Applications [24–27], 3D printing [28] and so on.

At present, there are many 3D reconstruction technologies, which can be divided into contact technology [28] and non-contact technology [1–3, 7–14] based on whether the measuring device is in direct contact with the actual target during the mapping process. The non-contact 3D reconstruction method can be divided into active vision method and passive vision method according to whether the light source is projected to obtain 3D information during the measurement process. The specific classification is shown in Figure 1. In the practical application process, non-contact methods represented by laser and image technology are the most widely used due to their excellent reconstruction effect [29–32], high reconstruction efficiency and simple operation. According to different reconstruction principles, image-based 3D reconstruction technologys can be divided into 3D reconstruction algorithms based on traditional multi-view geometry and 3D reconstruction algorithms based on deep learning. The traditional 3D reconstruction algorithm of Multi-View geometry, represented by sparse point cloud reconstruction with structure from motion (SFM) and dense point cloud reconstruction with multi-view stereo (MVS), integrates information from multiple images. It has great advantages in the measurement of 3D objects and the reconstruction of highly realistic 3D models, and is the most widely used. However, due to the removal of some points with low confidence in the process of filtering and fusion, the surface holes of the reconstructed model appear, which affects the integrity of the surface reconstruction. With the development of relevant theories in the field of deep learning, more and more researchers begin to use convolutional neural network (CNN) to carry out 3D reconstruction research. Since the neural radiation field (NeRF) was proposed in 2020, the 3D reconstruction algorithm based on deep learning has replaced the steps of MVS, surface reconstruction and texture reconstruction in traditional methods by building a neural network, and can directly generate a real 3D model without holes after input of multi-view images and internal and external parameters of the camera. Therefore, in the field of small scenes such as indoor scenes, its reconstruction accuracy, speed and integrity are gradually ahead of traditional methods. When

* Corresponding author: liulimin0807@aeu.edu.cn

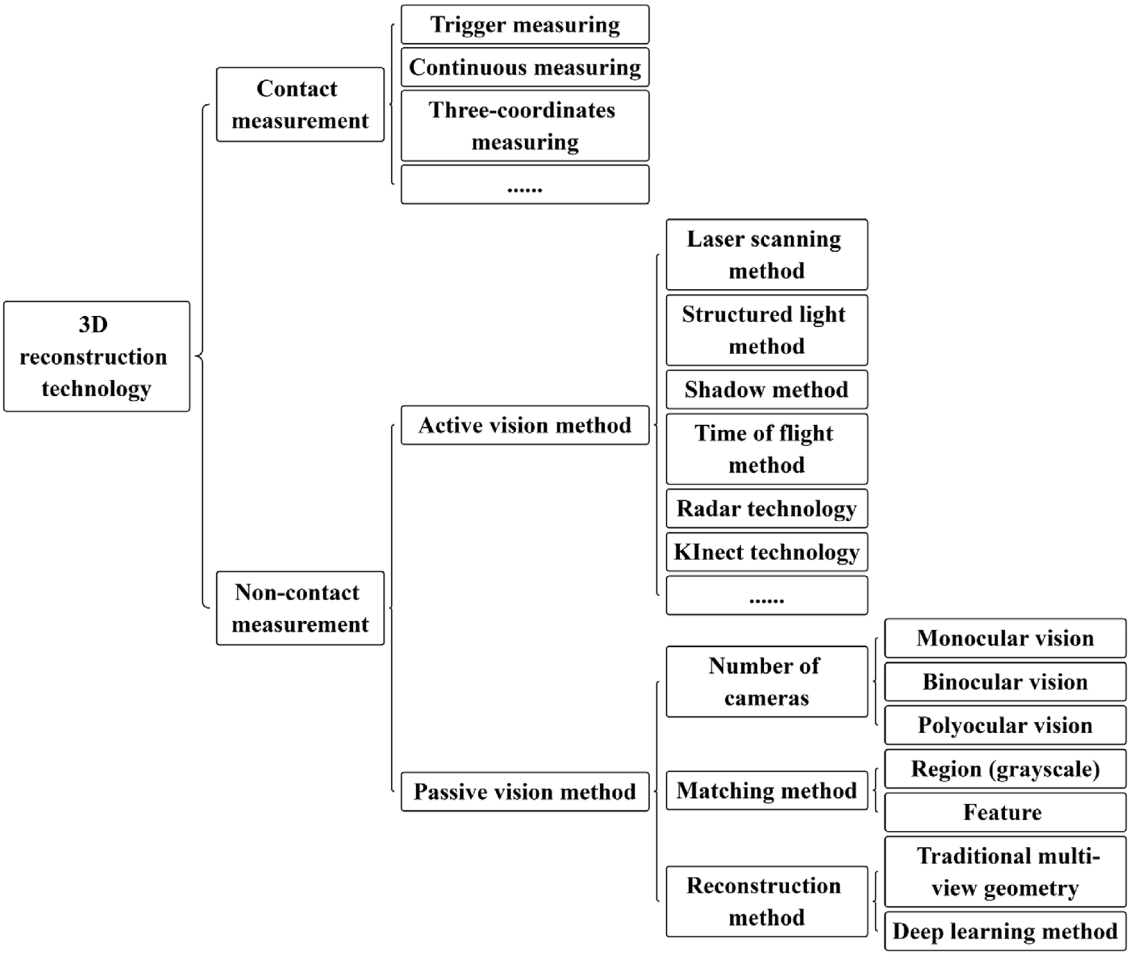

Fig. 1. 3D reconstruction technology.

it is applied to the 3D reconstruction of large scenes, due to the influence of factors such as illumination, excessive amount of network learning and image deformities, the method based on deep learning has no advantage compared with the traditional method in terms of reconstruction accuracy by improving the rendering quality.

In this paper, the development status and main technical routes of image-based 3D reconstruction technology are introduced in detail. Firstly, explicit 3D representation and implicit 3D representation are introduced in detail. Then, it focuses on the technical routes, important theories, data sets and evaluation indexes of traditional multi-view 3D reconstruction algorithm and deep learning 3D reconstruction algorithm, and analyzes the advantages and disadvantages of related technical routes in detail. Finally, according to the development status of 3D reconstruction, the relevant technical directions in the field of 3D reconstruction are summarized and prospected.

2 Common three-dimensional expression

As a key technology in computer vision, 3D representation can directly reflect 3D scenes by adding depth information compared with traditional 2D representation. It is widely used in 3D scene restoration, Simultaneous Localization And Mapping (SLAM) [33, 34], Augmented Reality (AR) [35] and other fields. Common 3D expressions can be divided into explicit and implicit:

2.1 Explicit 3D representation

The explicit 3D expression includes point cloud, Volume Pixel (voxel) and Mesh. By directly modeling the 3D scene, it is possible to directly generate the 3D scene.

2.1.1 Point cloud

The point cloud is a collection of massive points in a certain coordinate system, which can be obtained through Light Detection and Ranging (Lidar), RGB D cameras, SFM and other methods. It includes massive information such as 3D coordinates, reflection intensity, and flight time. Different from 2D images which need to be calculated through multi view geometry, point clouds directly provide 3D depth information of the target scene, preserving complete 3D spatial geometric information, but cannot provide the connection relationship between points.

In 2017, Charles et al. proposed PointNet on CVPR2017 [36], which pioneered the use of deep learning for feature learning of point clouds without changing the invariant characteristics of point cloud point arrangement, and applied the learned features to 3D point cloud classification and segmentation tasks. In 2017, Charles et al. proposed PointNet++ as an improved version of PointNet [37], which is a multi-layer neural network that mimics CNN and designs a network structure that iteratively extracts local features and combines multi-level encoders to solve the problem of PointNet not extracting layer by layer features. In 2017, Fan et al. proposed a network for generating 3D object reconstruction point sets from a single image on CVPR2017 [38], which solved the problem of 3D reconstruction from a single image and generated a direct form of output point cloud coordinates, but also shows a strong 3D shape completion performance and a good variety of credible prediction capabilities. In 2021, Nie et al. proposed the RfD-Net network in CVPR2021 [39], which is a method of directly generating 3D scenes from 3D point clouds. This method proposes the idea of first detecting and then reconstructing. Due to the RfD-Net network's support for implicit learning, it can effectively complete the shape completion task. In 2019, Lu et al. proposed a two-stage training-intensive point cloud-generated network [40]. After combining these two stages and fine-tuning, an end-to-end network was obtained that can generate dense point clouds from a single image. In 2021, Luo et al. proposed a diffusion probability model of 3D point cloud generation in CVPR2021 [41], which was inspired by the non-equilibrium thermodynamic diffusion process and transformed the point cloud generation process into a reverse diffusion process using noise distribution to generate the required shape distribution.

2.1.2 Voxel

Voxel is the abbreviation for Volume Pixel, which is a regular data structure in three-dimensional space. It uses a fixed volume cube as the smallest unit to represent an object as an N3 three-dimensional grid, and sets the grid state to 1 or 0 based on the occupied or idle state of the grid.

In 2015, Wu et al. proposed a convolutional deep belief network called 3D ShapeNets in CVPR2015 [42], which represents geometric three-dimensional shapes as the probability distribution of binary variables on a three-dimensional voxel grid, and jointly recognizes and reconstructs objects from 2.5D depth maps of a single view (such as popular RGB-D camera). In 2016, Choy et al. proposed a new Recurrent neural network architecture called 3D cyclic reconstruction neural network (3D-r2n2) in ECCV2016 [43]. This network learns the mapping of object images to their underlying 3D shapes from a large amount of synthesized data without requiring any image labels or object class labels for training or testing. It obtains one or more images of object instances from any viewpoint and outputs the reconstruction of the 3D scene in the form of a 3D occupied grid. In 2018, Wu et al. proposed an algorithm called ShapeHD in ECCV2018 that combines depth convolution network with shape prior of adversarial learning [44]. It uses a single image to complete the 3D reconstruction task by combining depth Generative model with shape prior of adversarial learning, experimental verification of the ShapeHD single view 3D shape completion and reconstruction tasks has shown that the algorithm performs well.

2.1.3 Mesh

A mesh is composed of vertices, edges, and faces, which can be triangles or polygons. It is a form used to represent the surface of irregular 3D objects in computer vision. Since triangle mesh is the smallest unit in the mesh, so any polygon mesh can be represented by multiple triangle mesh.

In 2018, Kanazawa et al. proposed a single view prediction framework for learning texture 3D meshes using image sets as supervision in ECCV2018 [45]. This method allows for training using annotated image sets, where learning deformable models and 3D prediction mechanisms do not rely on 3D truth or multi view image supervision, and are not only applicable to single view scenes, but may yield better results for multi view scenes. In 2018, Wang et al. proposed a 3D mesh model called Pixel2Mesh in ECCV2018 [46], which is an end-to-end deep learning architecture that can generate 3D shapes in triangular meshes from monochromatic images. Different from the existing methods, Pixel2Mesh network represents the 3D mesh in the graph based Convolutional neural network, and uses the perceptual features extracted from the input image to generate the correct geometric shape by gradually deforming the ellipsoid. In 2019, Wen et al. proposed in ICCV2019 that a 3D mesh model called Pixel2Mesh++ be generated from multi-view images [47]. Compared with the shape generated directly from prior information before, this paper further improves the shape quality by using the cross-view information of the graph convolutional network, but also demonstrate good generalization ability across different semantic categories, number of input images, and mesh initialization quality.

2.2 Implicit 3D representation

The implicit 3D representation uses functions to express the 3D scene. By modeling the space occupation of the 3D object, the explicit 3D scene can be obtained by post-rendering and other processing, but no intermediate 3D scene reconstruction process is required. Because the implicit 3D representation uses the neural network method to solve the problem of many holes in the traditional 3D reconstruction, the difference between the implicit 3D representation and the traditional 3D reconstruction method is that the expression is continuous. Common implicit 3D expression functions include Occupancy Function, Signed Distance Field (SDF) and NeRF. In 2019, Mescheder et al. proposes a new learning based 3D reconstruction method called Occupancy Networks on CVPR2019 [48], which implicit represents 3D surfaces as continuous decision boundaries of deep neural network classifiers. Compared to existing methods, this method encodes a description of the 3D output at infinite resolution without taking up too much memory. In 2019, Park et al. proposed DeepSDF on CVPR2019 [49], which utilizes a learned continuous signed distance function (SDF) to represent a class of shapes in

partially noisy 3D input data, achieving high-quality shape representation, interpolation, and completion. In 2020, Millenhall et al. proposed a 3D reconstruction method based on neural radiation files on ECCV2020 [50], which queries the 5D coordinates along the camera ray to synthesize the view, uses full connected (non convolutional) depth network to represent the scene, and uses classic volume rendering technology to project the output color and density into the image.

3 3D reconstruction algorithm based on traditional multi-view geometry

The 3D reconstruction technology based on traditional multi-view geometry mainly goes through the steps of image depth data acquisition, feature extraction, stereo matching, 3D image reconstruction, etc., and finally transforms the real target environment into a 3D mathematical model that can be processed and expressed by computer. According to the acquisition methods of image depth information, it can be divided into passive vision method and active vision method. The key and difficult point of algorithm implementation is to obtain the depth information of target scene, and then carry out stereoscopic matching and fusion of the data, so as to realize the three-dimensional reconstruction of the target environment. The specific idea is to restore the depth of the object by calculating the three-dimensional spatial position of the object image taken from different angles, find the corresponding feature matching relationship from the image through geometric constraints and feature matching relationship to restore the spatial coordinate relationship between the object and the camera, and then carry out dense reconstruction to determine the position and orientation of each face. This process fuses information from multiple images and has great advantages in 3D object measurement and highly realistic 3D model reconstruction. Highly realistic 3D model reconstruction.

3.1 Image depth data acquisition

According to the different information acquisition mechanism, stereo vision technology can be divided into monocular vision, binocular vision and multiocular vision. In monocular vision, sparse point clouds and dense point clouds are obtained through SFM and MVS to achieve 3D scene reconstruction. Binocular vision uses binocular cameras with well-calibrated internal and external parameters to obtain stereoscopic correction and stereoscopic matching to generate parallax map, and calculate depth map to generate point cloud map. The acquisition of depth information in multiocular vision is similar to the working principle of binocular vision, which can be captured by images taken by multiocular cameras.

Referring to Figure 2, when the left and right cameras observe the same three-dimensional point at the same time, the difference between the projection points projected on the left and right image planes is called parallax. The reason why humans can perceive the distance of space objects is because human eyes are a binocular stereo vision system.

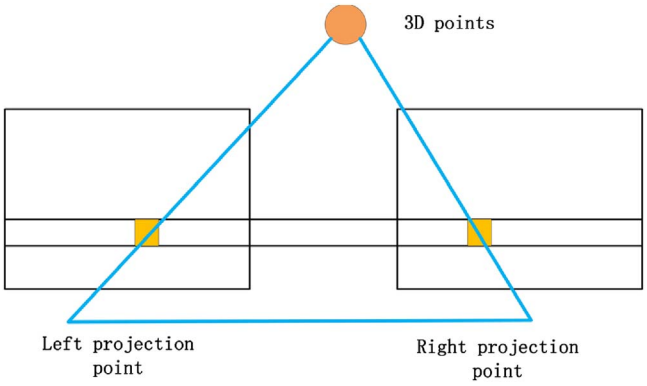

Fig. 2. Parallax schematic diagram.

Compared to monocular imaging, binocular imaging acquisition is more complex, but obtaining depth information is simpler. According to the relative position of the camera, binocular imaging can also be divided into binocular horizontal mode, binocular convergence horizontal mode, and binocular axial mode. In practical applications, two monocular cameras can be used to work simultaneously, or one monocular camera can be used to work separately at different angles to complete image acquisition.

3.2 Image feature extraction

Image features are abstract expressions and descriptions of pixels and pixel sets, which are crucial for stereo matching of images and determining the corresponding relationships of the same scene in different images. Common features include: point features, line features, surface features and body features.

In 1981, Moravec first applied the corner detector to image matching [51]. In 1988, Harris and Stephens improved the corner detector proposed by Moravec and proposed the Harris corner detection operator [52]. Moving the window grayscale in any direction at a corner will result in significant changes, as the image gradient has two or more main directions in the area near the corner, which can be used to detect corners. In 1993, Harris applied the Harris corner detection operator to the SFM [53], and demonstrated its effect in 3D reconstruction and motion tracking. Since then, Harris corner detection operator has not only been used to detect corner points, but also to detect image positions with large gradients in any direction on a specific scale. However, the Harris corner detection algorithm is very sensitive to changes in image scale due to its inability to change the size of the template. It has poor adaptability for corner detection at different scales and cannot achieve good matching results. In 1999, Lowe proposed a new image feature generation method called Scale Invariant Feature Transform (SIFT) based on the behavioral model of complex cells in the mammalian visual cortex [54]. This method effectively recognizes scale invariant features by using a phased filtering method to transform the image into a large set of local feature vectors, describing the local image region sampled relative to its scale spatial coordinate frame. The partial invariance of local changes

is achieved by blurring the gradient position of the image. Each local feature vector is not affected by Image scaling, translation and rotation, and is less affected by lighting changes, affine and 3D projection. In 2001, Mikolajczyk et al. proposed the Harris-Laplacian operator on ICCV2001 [55], which used the normalized Laplacian operator to convolver images on multiple scales, and found the extreme point of the normalized Laplacian response value in the scale space for each pixel. In 2002, Brown and Lowe proposed a new method [56]. First, Hough transform was used to eliminate most error matches, and then RANSAC and polar constraints were used to deal with the remaining Outlier, which can achieve accurate positioning of key points and solve the problem that the paper in 1999 could not accurately locate key points. In 2004, Lowe finally improved the SIFT descriptor on the basis of the 1999 paper and formally proposed the SIFT operator [57], which was widely used in the field of computer vision. In this paper, the SIFT operator represents different scale Spaces by using Gaussian ambiguity of different parameters, and approximates the Laplacian operator by using the Difference of Gaussian (DOG). The features extracted by this method are invariant to the scaling and rotation of images. Moreover, it can match the large scale affine transformation, 3D viewpoint change, noise addition and illumination change. In 2006, Bay et al. proposes a new scale and rotation invariant detector and descriptor called SURF (Speeded-Up Robust Features) [58]. Compared with SIFT, SURF utilizes first-order Harr wavelet response distributions in the x and y directions instead of gradients to obtain the main direction of feature points, and uses 64D to calculate the integral image, reducing feature calculation and matching time and enhancing robustness. In 2006, Rosten et al. proposed a corner detection method based on machine learning called FAST [59], which can quickly extract corner points but cannot effectively describe diagonal points, so the scale invariance of layout and rotation invariance. In 2011, Rublee et al. proposed a BRIEF based fast binary descriptor called ORB [60], which is two orders of magnitude faster than SIFT and has rotation invariance and anti-noise properties. In 2012, Cruz-Mota et al. proposed a spherical coordinate SIFT algorithm for omnidirectional images [61], which can generate two types of local descriptors: local spherical descriptor and local plane descriptor. In addition, this algorithm introduces a plane-to-sphere mapping and gives its estimation algorithm, which allows the object to be extracted from an omnidirectional image given SIFT descriptor in a planar image. In 2016, Lakshmi et al. proposed a Image registration algorithm based on image local features [62], which is invariant to image size, illumination, rotation and viewpoint change by using the features extracted by SIFT operator. Compared with previous registration algorithms, it is more robust. In 2017, Al-khafaji et al. proposed spectral spatial scale invariant feature transform (SS-SIFT) [63]. As a new method to extract significant invariant features from hyperspectral images, it can simultaneously explore spectral and spatial dimensions to extract spectral and Geometric transformation invariant features for hyperspectral image registration under different spectral conditions. In 2021, Li et al. proposed a feature detector called FD-TR applied to digital image watermarking [64]. It based on scale-invariant feature transformation and bidirectional feature regionalization, the detector extracts key points using SIFT operators, and proposes edge and neighbor filtering methods to generate candidate feature points. The comparison with existing methods shows that the proposed method has better performance in terms of robustness and watermarking quality. The advantages and disadvantages of the above feature point detection algorithms are shown in Table 1 below.

3.3 Image stereo matching

Image stereo matching is the process of finding feature points with the same name after feature extraction, establishing the corresponding relationship between the same 3D point in different images, and calculating the depth image corresponding to the disparity [65–68]. At present, the commonly used stereo matching methods include: gray-based stereo matching method, feature-based stereo matching method and deep learning-based stereo matching algorithm. Among them, gray-scale based stereo matching method is a region based method, which needs to consider the relevant neighborhood properties of matching points. Feature based image stereo matching mainly utilizes the coordinates of inflection points and corner points, edge line segments, and target contours in the image. The steps are as follows:

1. Feature detection;
2. Feature matching;
3. Matching feature point pairs, using the least square method and other methods to calculate the matching parameters;
4. According to the transformation model, the image to be matched is matched to the reference image to realize the matching between images;
5. The depth information of the image is obtained by calculating parallax, and the dense depth map and dense parallax field are obtained by data interpolation.

Because the gray-based stereo matching method depends on the statistical features of the image gray level, it is more sensitive to the target surface structure and lighting conditions, so it is not effective in the case of insufficient texture information and large distortion on the surface of spatial objects. However, feature-based stereo matching methods have lower sensitivity to changes in ambient lighting and more stable performance, so it is the most widely used.

In 2011, Mei et al. proposed a GPU-based stereo matching algorithm named ADCensus in ICCV2011 [69], which performed well in both accuracy and speed. For the four data sets (Tsukuba, Venus, Teddy and Cones), the CPU implementation requires 2.5 s, 4.5 s, 15 s and 15 s respectively, while the GPU implementation requires only 0.016 s, 0.032 s, 0.095 s and 0.094 s respectively. The GPU-friendly system design brings an impressive 140× speedup in the processing speed. In 2011, Bleyer et al. proposed a local algorithm [70], which calculated a

Table 1. Advantages and disadvantages of feature point detection algorithms.

Feature point detection algorithm	References	Advantages	Disadvantages
Harris	Moravec [51], Harris and Stephens [52], Harris [53]	Rotational invariance, luminance invariance	Lack of scale invariant properties and affine invariance properties
SIFT	Lowe [54, 57], Brown and Lowe [56], Cruz-Mota et al. [61], Al-khafaji et al. [63], Li and Yuan [64]	Rotational invariance, scale invariance, luminance invariance, good robustness	The calculated dimension is too large and the operation speed is slow
Harris-Laplacian	Mikolajczyk and Schmid [55]	Rotational invariance, scale invariance	Redundancy point, anti-interference ability is not strong, real-time performance is poor
SURF	Bay et al. [58]	Compared with SIFT, the calculation is smaller and the speed is faster, Rotational invariance, scale invariance, good robustness	The computational speed is an order of magnitude faster than SIFT and an order of magnitude slower than ORB
FAST	Rosten and Drummond [59]	Fast operation speed	Easy to be affected by noise, poor robustness, Lack of rotation invariance
ORB	Rublee et al. [60]	Fast operation speed, Rotational invariance	Lack of scale invariant properties

3D plane on each pixel by combining PatchMatch algorithm (spatial propagation), graph propagation and time propagation. By calculating plane parameters, sub pixel precision disparity values and optimal planes can be obtained through matching. In 2015, Han et al. proposed an image stereo matching algorithm called MatchNet on CVPR2015 [71], which combined the learning feature representation and the learning feature comparison function. This algorithm is composed of convolutional neural network and three fully connected layer networks, which are used to extract features and calculate the similarity between features respectively. In 2015, Barron et al. proposed a Stereo matching algorithm called Fast Bilateral Space Stereo algorithm in CVPR2015 [72], which uses the idea of bilateral filtering to obtain the minimum global matching cost and the speed is 10–100 times faster than other matching algorithms on the basis of obtaining higher quality defocusing effects. The solution method of Fast Bilateral Space Stereo cannot be applied to the deep learning process, because its cost function cannot be derived for back propagation. In 2016, Barron et al. proposed a new edge-perception smoothing algorithm called Fast Bilateral Solver (FBS) algorithm based on the idea of Fast Bilateral Space Stereo algorithm [73]. It's not only improves the problem that the Fast Bilateral Space Stereo algorithm cannot be applied in deep learning, but also combines the flexibility and speed of simple filtering methods, as well as the accuracy of domain specific optimization algorithms. In 2015, Žbontar and LeCun proposed a convolutional neural network algorithm called MC-CNN algorithm to calculate stereo matching cost in CVPR2015 [74]. This algorithm transforms the cost calculation of stereo matching into classification problem in deep learning, and applies deep learning theory to stereo matching for the first time. In 2015, Chen et al. proposed a data-driven stereo matching algorithm in ICCV2015 [75]. This algorithm draws on the idea of MC-CNN, uses convolutional neural network to learn the visual similarity relationship between corresponding image blocks, and directly maps the intensity value to the embedded feature space to measure pixel dissimilarity. After testing on KITTI and Middlebury datasets, it is proved that the pixel similarity measurement proposed by this algorithm is superior to the traditional matching method. In 2016, Žbontar and LeCun further expanded MC-CNN algorithm by using deep learning to calculate stereo matching cost, and proposed a fast and accurate architecture [76]. The above two network architectures are tested on KITTI 2012, KITTI 2015 and Middlebury dataset, which shows that convolutional neural networks are feasible for real-time stereo matching cost calculation and parallax estimation. In 2017, Ye et al. proposed an efficient stereo matching algorithm based on convolutional neural network [77], which mainly consists of two parts: one is based on a multi-scale feature fusion architecture, which can learn rich local information; The other is to combine optimal disparity and suboptimal disparity, and then use different basic learners to learn the end-to-end disparity optimization model of contextual information. In 2019, Zhang et al. proposed a cost aggregation network for stereo matching algorithm called GA-Net on CVPR2019 [78]. This algorithm consists of two new neural network layers, aiming to capture the cost dependencies of local and entire images, respectively. In 2022, Zhang et al. proposed stereo selective whitening (SSW) loss and stereo constrained feature (SCF) loss [79], which improved

Table 2. Advantages and disadvantages of Stereo image matching algorithms.

Stereo image matching algorithms	References	Advantages	Disadvantages
ADCensus	Mei et al. [69]	High matching speed, high accuracy	The matching fuzziness in duplicate area and similar texture area is easy to cause mismatching
PatchMatch	Bleyer et al. [70]	Global matching is realized in the inclined plane and sub-pixel matching accuracy is obtained	Many operations need to be processed by a single pixel one by one, resulting in slow running speed and need to be carried out in parallel
MatchNet	Han et al. [71]	A new deep learning network structure with fewer descriptors is proposed, which significantly improves the patch-matching effect	Image blocks can only be processed after sampling, and it is impossible to find and match the whole image
Fast Bilateral Solver	Barron et al. [72], Barron and Poole [73]	Matching speed is very fast	The result is easily affected by the reference image
MC-CNN	Žbontar and LeCun [74, 76], Ye et al. [77]	The deep learning theory is applied to stereo matching for the first time, and the matching accuracy is improved	The matching effect is not good in occluded areas, untextured areas and repetitive pattern areas
GA-Net	Zhang et al. [78]	The matching accuracy of occluded area, untextured area and reflection area is improved	Memory usage are too high

the generalization ability of stereo matching networks by ensuring feature consistency among matched pixels. SSW and SCF can be expressed as:

$$\mathcal{L}^{\text{scf}} = \frac{1}{\sum_{(u,v)\in C} M_{u,v}} \sum_{(u,v)\in C} L^{f}(u,v) \odot M_{u,v} \tag{1}$$

where $L^{f}(u,v)$ stands for pixel-by-pixel contrast loss. $M_{u,v}$ represents the reserved region after the non-matching region is removed.

$$\mathcal{L}^{\text{ssw}} = \frac{1}{\Gamma} \sum_{\gamma=1} \left\| \sum\nolimits_{\gamma} (\widehat{\mathrm{X}}^{l}) \odot \widetilde{M} \odot \widehat{M} \right\|_{1} \tag{2}$$

where $\hat{M}$ is a strict upper triangular matrix as the covariance matrix is symmetric, Γ is the number of layers to which the SSW loss is applied, and γ indexes the corresponding layer (i.e. conv1, conv2 $\times$ in PSMNet).

The advantages and disadvantages of the above stereo image matching algorithms are shown in Table 2.

3.4 Common visual 3D reconstruction algorithms

Among the traditional 3D reconstruction algorithms, SFM [80–85] and MVS [86–90] are the most widely used. In 2013, Moulon et al. proposed a global SFM that utilizes unordered image sequences for large-scale 3D reconstruction, which ensures the robustness and accuracy of the algorithm while ensuring scalability [80]. In 2015, Heller et al. developed an online service platform that used SFM to restore 3D scenes from images [81]. In 2016, Schonberger et al. proposed an incremental SFM called COLMAP in CPR2016 [82], which greatly improved the accuracy, robustness and integrity of the SFM algorithm by improving the steps of triangulation and Bundle Adjustment (BA) in incremental SFM. The BA algorithm takes the camera attitude and the three-dimensional coordinates of the measuring points as unknown parameters, takes the coordinates of the characteristic points detected on the image for the front intersection as the observation data, and uses the least square method to adjust to obtain the optimal camera parameters and the world coordinate system. In 2017, Cui et al. proposed a hybrid SFM called HSFM in CVPR2017 [83], which takes into account efficiency, accuracy and robustness under a unified framework, and solves the problems of low efficiency of incremental SFM and poor robustness of global SFM. In 2020, Yin and Yu proposed a monocular 3D reconstruction method based on incremental SFM [84], which first combined SIFT and ORB features matching as the input of sparse reconstruction, then used the incremental SFM algorithm to obtain sparse 3D points from the image set, and finally combined optical flow and ORB features to reconstruct the image intensively. In 2021, Wang et al. proposed a deep neural network-based SFM technology to solve the problem of reconstructing 3D faces from multi view facial images [85]. This algorithm utilizes a new unsupervised 3D face reconstruction architecture and achieves accurate training of facial pose and depth maps through multi view geometric constraints. In 2006,

Table 3. Advantages and disadvantages of common visual 3D reconstruction algorithms.

Common visual 3D reconstruction algorithms	References	Advantages	Disadvantages
Incremental SFM	Schönberger and Frahm [82], Yin and Yu [84]	The performance is robust and the reconstruction precision is high	Affected by the initial image on the selection and camera add order, the cumulative error is large and the efficiency is not high in the reconstruction of large scenes
Global SFM	Moulon et al. [80]	Not affected by the initial image pair and the order of camera addition, the cumulative error is small, and the reconstruction efficiency is high	The robustness is not good, and the completeness of scene reconstruction is insufficient
Hybrid SFM	Cui et al. [83]	The cumulative error is small and the robustness is good	The efficiency is not high
Voxel based MVS	Sinha et al. [87]	The generated point cloud is regular and mesh is easy to extract	Reconstruction accuracy is related to voxel particle size, and it is difficult to deal with large scenes
Feature point growing based MVS	Lin et al. [88]	The point cloud has high precision and uniform distribution	Areas with weak textures are prone to holes and require reading all images at once
Depth-map merging based MVS	Seitz et al. [86], Lindenberger et al. [89], Zhou et al. [90]	It can be used in parallel computation for 3D reconstruction of large scenes, and the number of point clouds obtained is large	Too dependent on neighborhood image group selection

Seitz et al. published an article in CVPR2006 [86], which systematically introduced MVS algorithm from classification, multi-view data set and evaluation of MVS algorithm, and quantitatively compared several multi-view stereo image reconstruction algorithms. In 2007, Sinha et al. proposed a voxel based MVS method that utilizes photometric consistency to partition voxels, which to some extent solves the problem of low voxel resolution [87]. In 2020, Lin et al. proposed a binocular stereo vision 3D reconstruction method based on feature point matching [88], which utilized binocular stereo vision, feature matching and other traditional technologies to achieve 3D scene reconstruction. In 2021, Lindenberger et al. proposed an algorithm for improving SFM accuracy using depth feature measurement in ICCV2021 [89]. This algorithm improves motion structure by directly aligning low-level image information from multiple views, and optimizes feature measurement errors based on dense features predicted by neural networks. In 2021, Zhou et al. proposed a high-precision 3D reconstruction system with good robustness to solve the problem of insufficient model details and low accuracy [90]. This algorithm starts from the incremental SFM structure and adopts the idea of deep fusion, achieving accurate restoration of depth map details while significantly reducing memory consumption. The advantages and disadvantages of common visual 3D reconstruction algorithms are shown in Table 3.

Because SFM can obtain camera internal and external parameters through feature point matching, but the sparse feature matching points make SFM can only obtain sparse point clouds. By matching each pixel of the calibrated image one by one, MVS can obtain the three-dimensional coordinates of each pixel to the maximum extent and generate the dense point cloud pair. Each specific step is shown in Figure 3.

1. Obtain sequence images through multi view shooting and use them as input to the system;
2. Feature extraction and matching: In this process, feature points are extracted according to texture features to estimate the internal and external parameters of the camera, and the matching relationship between image pixels is established;
3. Sparse reconstruction: The process of SFM is mainly to extract sparse feature points (sparse point cloud) from 3D scene sequence images to obtain basic geometric information required for 3D reconstruction;
4. Dense reconstruction: MVS uses the information extracted from SFM and the information in the 2D image that has not been fully utilized to match pixels in the image one by one and generate dense point clouds to make the 3D model information more complete. The process of MVS is to first estimate the

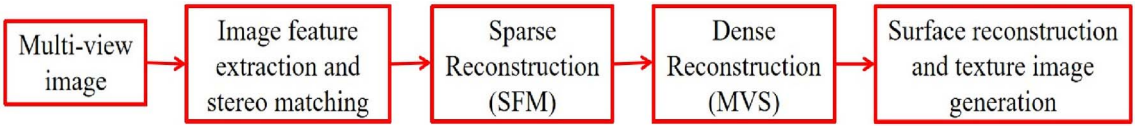

Fig. 3. 3D reconstruction process.

depth of each image, then fuse the depth of all perspectives, and finally obtain the complete point cloud and model [90]. Among them, the method based on block matching is the most common, and it can be applied to multi-view 3D reconstruction of large scenes with the help of parallel operation of GPU;

5. Finally, the obtained point cloud is used to reconstruct the surface of the object and generate texture images, restoring the original 3D scene.

3.5 Brief summary

The traditional 3D reconstruction technology based on multi-view geometry first collects the target image data through the camera, then uses the image registration to obtain the parallax image, further calculates the depth image, and uses the traditional multi-view geometry 3D reconstruction method or deep learning method to complete the restoration and reconstruction of the 3D spatial information in the target scene. The depth information obtained by calculating parallax using stereo vision is not accurate compared to the depth information obtained from laser point clouds, so the reconstruction accuracy of pure visual 3D reconstruction methods is not high. Due to the different types of cameras used in 3D visual reconstruction methods, the reconstruction effect is also different. For example, using a color camera to obtain information, the reconstruction effect can accurately express the color information of the target environment. However, due to the passive acquisition method used in this method, it is susceptible to the influence of environmental light intensity; If the infrared camera is used to collect the information of the target environment, it is not affected by the light intensity of the target environment, and can achieve all-weather work, but it cannot obtain the color information of the target environment.

4 3D reconstruction algorithm based on deep learning

The traditional reconstruction method uses luminosity consistency to calculate dense 3D information, which is highly accurate in ideal environment, but in some environments such as weak texture, high reflection and repeated texture, it is easy to have reconstruction difficulties or holes. The 3D reconstruction method based on deep learning uses the prior information to transform the 3D reconstruction problem into the process of encoding and decoding, and can realize the reconstruction of the 3D scene without complex calibration and mathematical processes [80–86].

In 2014, Eigen et al. proposed a CNN neural network consisting of two networks to make global prediction and local prediction for the depth of monocular images, and used CNN neural network for 3D reconstruction for the first time [91]. In 2015, Eigen and Fergus proposed a CNN neural network consisting of three networks based on the previous paper [92], which completed three tasks: monocular image depth estimation, normal vector estimation and image semantic segmentation. In 2017, Crispell and Bazik proposed a 3D face reconstruction method called Pix2Face from 2D images in ICCV2017, which uses an improved U-Net neural network architecture to estimate dense 3D coordinates and 3D geometry [93]. In 2018, Yao et al. proposed a depth prediction network based on multi-view images called MVSNet on ECCV2018 [94]. It opens the way for 3D reconstruction with multiple views using depth. In 2019, Yao et al. proposed R-MVSNet on CVPR2019 [95]. This algorithm improves MVSNet by using GRU for cost aggregation, reducing the model size without much reduction in algorithm accuracy. In 2019, Chen et al. made improvements on the basis of MVSNet and proposed an algorithm called Point-MVSNet on ICCV2019 [96]. This algorithm carried out scene processing by operating point cloud, fully combined geometric prior information and 2D texture information to enhance point cloud features and improve efficiency. In 2020, Zhang et al. proposed Vis-MVSNet on BMVC2020 [97]. This algorithm improved MVSNet and clearly inferred and integrated pixel-by-pixel occlusion information in MVS network by matching uncertainty estimation, significantly improving the depth accuracy in severe occlusion scenes. In 2021, Wei et al. proposed a novel recurrent multi-view stereo network based on long short-term memory (LSTM) with adaptive aggregation on 2021 [98]. Different from traditional 3D CNNS, this algorithm uses a hybrid network with a cyclic structure for cost-volume regularization, achieving high-resolution reconstruction and finer hypothetical plane scanning. AA-RMVSNet used a hybrid structure DHU-LSTM, which absorbed the advantages of LSTM and U-Net, and regularized the 3D matching body into a prediction depth map containing different levels of information. After the predicted depth map was input into LSTM, the reconstruction accuracy was maintained while the memory cost was reduced.

In 2022, Peng et al. proposed a coarse-to-fine framework called UniMVSNet on CVPR2022 [99]. This framework combines the advantages of regression and classification problems applied in depth estimation, and redefines depth estimation as a multi-label classification task, so that it can achieve accurate prediction of depth while maintaining robustness. In 2020, Millenhall et al. proposed a 3D reconstruction method based on NeRF on ECCV2020 [50]. The input of this method is a function including 5D vector,

and the radiation field approximated by fully connected neural network is used to implicitly express the 3D scene. The proposal of NeRF greatly promoted the development of 3D scene implicit representation technology. In 2021, Yen-Chen et al. proposed an anti-radiation attitude estimation method called iNeRF based on NeRF. This method can use NeRF to realize 6DOF attitude estimation with complex geometric shapes and objects in the case of mesh-free and RGB-only [100]. Ma et al. proposed a neural network framework based on NeRF called Deblur-NeRF [101]. By using Deformable Sparse Kernel, the fuzzy kernel with spatial variation was modeled by deformable sparse kernel deformed at each spatial position, so that clear NeRF could be reconstructed even under fuzzy input. In 2022, Xu et al. proposed a Point-based Neural Radiance Fields 3D reconstruction algorithm called Point-NeRF on CVPR2022 [102], which integrates NeRF with deep MVS. This method uses MVS to generate point clouds, and then uses NeRF to quickly generate three-dimensional scenes on the basis of the generated point clouds, and the reconstruction effect exceeds the visual quality of NeRF. In 2023, Jiang et al. put forward a 3D reconstruction method called AligNeRF on CVPR2023 [103]. This method uses high-resolution data to train NeRF, adding and recovering more high-frequency detail than the most advanced NeRF models without significantly increasing the cost of training and testing. In 2023, Xu et al. proposed a grid-guided neural radiation field for large-scale scenes on CVPR2023 [104]. By combining NeRF and grid, this method can effectively encode local and global scene information, and finally achieve high visual fidelity rendering of ultra-large urban scenes. In 2020, Stucker and Schindler proposed a method for dense 3D reconstruction of scenes using deep learning in CVPR2020 [105]. This method uses traditional stereo matching algorithms for approximate 3D reconstruction, and trains deep neural networks through residual learning to enhance the reconstruction effect. Among them, ResNet based on residual learning was proposed by He Kaiming et al in 2015 [106]. Using the assumption that the optimal function is similar to the linear function, the residual function modeled by adding the input (identity function) to the output of the network can effectively solve the problem of gradient disappearance or negative optimization, and greatly accelerate the training speed of the network. In 2021, Peng et al. proposed a new view synthesis technology named Neural Body in CVPR2021 [107], which can better capture human actions with fewer input viewpoints and solve the problem that NeRF cannot process dynamic scenes. In 2021, Choe et al. proposed a deep fusion network called VolumeFusion for 3D scene reconstruction in ICCV2021 [108], which refers to traditional 3D reconstruction technology and has advantages compared with traditional 3D reconstruction algorithms and deep learning algorithms. In 2021, Wang et al. proposed a transformers based multi-view 3D reconstruction algorithm in ICCV2021 [109], which integrates feature extraction and view fusion into a Transformer network and studies the relationship between images by using self-attention among multiple unordered input images. In 2022, Huang et al. proposed a 2D convolutional network and 3D neural radiation field mutual learning method on CVPR2022 [110]. This algorithm utilizes neural radiation fields to express the continuous and dense features of 3D scenes, resulting in high-quality 3D consistent stylization effects for the reconstructed scenes.

Specific research ideas include the following three:

1. The depth learning method is introduced into the traditional 3D reconstruction algorithm for improvement;
2. Deep learning 3D reconstruction algorithm and traditional 3D reconstruction algorithm are integrated to complement each other's advantages;
3. Imitate animal vision and directly use depth learning algorithm for 3D reconstruction.

Traditional 3D reconstruction methods usually include SFM, MVS, surface reconstruction, and texture mapping. However, obtaining multiple images of the same object through an accurately calibrated camera is not practical in some cases, and sometimes the problem of surface emptiness of the reconstructed model may occur due to the absence of the viewing angle image. The 3D reconstruction method based on deep learning uses prior knowledge to build a neural network to replace the three steps of MVS, surface reconstruction and texture mapping in the traditional method, overcoming the problem that the traditional method is prone to surface voids.

5 Dataset introduction

This section introduces datasets that are widely used in image-based 3D reconstruction, and the relevant data sets are shown in Table 4. According to different data sources, data sets can be divided into real acquisition and synthesis.

In 2012, Geiger et al. proposed a dataset called KITTI dataset based on autonomous driving scenarios on CVPR2012 [111–113], which can be applied to tasks such as stereo vision, optical flow, visual odometry, SLAM, and 3D object detection. In 2014, Jensen et al. proposed a dataset containing 80 large variability scenes [114, 115]. Each scene in this dataset is obtained by a 6-axis industrial robot, and consists of 49 or 64 precise camera positions and reference structured light scanning. In 2015, Chang et al. proposed a dataset consisting of 3D CAD models of objects called ShapeNet [116, 117], which includes ShapeNetCore and ShapeNetSem sub datasets. The dataset contains over 3,000,000 models, of which 220,000 are divided into 3135 categories and provide many semantic labels for each 3D model. In 2017, Dai et al. proposed a richly annotated RGB-D video dataset called ScanNet on CVPR2017 [118]. This dataset contains 2.5 million images from 1513 scenes with estimated calibration parameters, 3D camera pose, surface reconstruction, semantic segmentation, texture mesh, dense object level semantic segmentation, and labels for aligning CAD models. In 2017, Knapitsch et al. proposed a 3D reconstruction dataset based on video images called Tanks and Temples dataset [119], which includes a total of 14 scenes, including objects such as tanks

Table 4. Common datasets.

Dataset	Release time	Data source	Number of models	Label category
KITTI	2012	Camera, LiDAR, GPS	389 stereo images and optical flow pairs, labels for over 200k 3D targets	2D/3D
DTU	2016	Camera, structured light scanner	80 scenes with a total of 3920 views	3D
ShapeNet	2015	CAD synthesis	Over 3 million models	3D
ScanNet	2017	RGB-D camera	1513 scenes with a total of 2.5 million images	3D
Tanks and Temples	2017	Cameras, industrial laser scanners	14 scenes, including objects such as tanks and artillery	3D
ETH3D	2017	Professional DSLR camera, multi camera shooting platform, laser scanner	13 training sets and 12 testing scenarios, 5 training and 5 testing videos, 27 training and 20 testing frames	3D
ApolloScape	2018	Camera, LiDAR	110K frames	2D/3D
Semantic KITTI	2019	LiDAR, GPS	28 labeling categories	3D
BlendedMVS	2020	Camera	113 3D models, totaling 17,818 images	3D
BDD100K	2020	Camera, GPS	100,000 HD videos	2D
nuScenes	2020	Cameras, Radar, LiDAR	1000 scenes tagged with 23 object categories totaling 14 million 3D tag boxes	2D/3D

and artillery, as well as large indoor environments such as auditoriums and museums. In 2017, Schöps et al. proposed the ETH3D dataset in CVPR2017 [120], which was used to evaluate binocular stereo vision and multi-view stereo vision methods, providing the first handheld multi-view stereo vision benchmark using consumer-grade cameras (professional Nikon D3X DSLR camera), as well as in line evaluation and algorithm comparison. In 2018, Huang et al. proposed the ApolloScape dataset in CVPR2018 [121, 122], which consists of simulation dataset, demonstration dataset, and labeling dataset, and can be applied to tasks such as target recognition and segmentation, stereo vision, semantic segmentation, etc. In 2019, Behley et al. proposed the SemanticKITTI dataset in ICCV2019 [123, 124], annotating all sequences in the KITTI dataset. In 2020, Yao et al. proposed the BlendedMVS dataset for MVS network training in CVPR2020 [125], which includes unstructured camera poses and provides training images and ground truth depth maps. In 2020, Yu et al. proposed the BDD100K dataset in CVPR2020 [126], which has 100K videos and 10 tasks including image labeling and semantic segmentation, among which 100K videos are divided into training set (70K), verification set (10K) and test set (20K). In 2020, Caesar et al. proposed the nuScenes dataset on CVPR2020 [127], which fully annotated the 3D bounding boxes and 8 attributes of 23 types of objects, with 7 times the number of labels and 100 times the number of images compared to the KITTI dataset. The datasets described above are shown in Table 1 below. As a very important part of the evaluation system of 3D reconstruction algorithm, the datasets also brings some problems to the development of 3D reconstruction algorithm. Firstly, the current use of datasets to evaluate the performance of 3D reconstruction algorithms has led to many algorithms only performing well on images in the dataset or similar images, while performing poorly on image categories that are not present in the dataset, resulting in poor generalization ability of existing 3D reconstruction methods. In addition, the images in the data set are generally single images without background, but they are more complex in the real environment. Therefore, it is necessary to combine image segmentation and recognition technology with 3D reconstruction to improve the generalization ability of 3D reconstruction technology.

6 Evaluation index

6.1 Mean square error (MSE)

In the field of 3D reconstruction, the mean square error (MSE) represents the symmetric surface distance between the 3D reconstructed shape and the real shape, and represents the difference between the reconstructed result and the real shape, which is defined as follows:

$$\mathrm{MSE} = \frac{1}{m \times n} \sum_{i}^{m*n} (I_i - K_i)^2, \tag{3}$$

where the image size is $m \times n$, I_i and K_i represent the predicted and true value of each pixel, respectively.

6.2 Peak signal-to-noise ratio (PSNR)

PSNR is the ratio of the maximum possible power of a signal to the destructive noise power that affects its accuracy. It can be used to quantify image reconstruction quality in

the field of image processing, and can be defined by mean squared error. The expression is as follows:

$$\begin{aligned}\mathrm{PSNR} &= 10 \times \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \\ &= 20 \times \log_{10}\left(\frac{\mathrm{MAX}_I}{\sqrt{\mathrm{MSE}}}\right),\end{aligned} \tag{2}$$

where MAX_I represents the maximum pixel value in the image (if 8 bit represents the image pixel value, MAX_I is 255).

6.3 Structural similarity (SSIM)

SSIM is an indicator for measuring the similarity between two images, which can be calculated using mean, standard deviation, and covariance to represent the similarity of luminance, contrast, and structure respectively. The expression is as follows:

$$S(x,y) = l(x,y)^{\alpha} * c(x,y)^{\beta} * s(x,y)^{\gamma}, \tag{5}$$

where $l(x,y)$, $c(x,y)$ and $s(x,y)$ represent luminance, contrast and structural features respectively; α, β and γ represent the proportion of the above three features respectively.

Luminance $l(x,y)$ is measured by the gray mean value μ_x and μ_y of the reconstructed model and the real model. The comparison function is expressed as:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}. \tag{6}$$

In the above formulas, μ_x and μ_y are expressed as:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{7}$$

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{8}$$

In the above formulas, x_i and y_i respectively represent the gray values of the reconstructed model and the real model at i.

$c(x,y)$ is the contrast measured by the gray standard deviation σ_x and σ_y between the reconstructed model and the real model. The contrast function is expressed as follows:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}. \tag{9}$$

In the above formulas, σ_x and σ_y are expressed as:

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}}, \tag{10}$$

$$\sigma_y = \left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2\right)^{\frac{1}{2}}. \tag{11}$$

The structural similarity $S(x,y)$ can be obtained by normalizing $(x-\mu_x)/\sigma_x$ and $(y-\mu_y)/\sigma_y$. Using correlation coefficient measurement, $S(x,y)$ can be expressed as follows:

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{12}$$

where the covariance is $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$, $C_3 = C_2/2$.

C_1 and C_2 are defined in formulas (11) and (12) as positive constants that prevent the formula from having a zero-division exception setting, which can be expressed as:

$$C_1 = (K_1 L)^2, \tag{13}$$

$$C_2 = (K_2 L)^2. \tag{14}$$

Among them, the default values of K_1 and K_2 are 0.01 and 0.03 respectively, and $L = 2^B - 1$ is the range of dynamic values of pixels.

Let α, β, γ equal to 1, put the formulas (4), (7), (10) into (3) to get the following formula:

$$S(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)}. \tag{15}$$

6.4 Mean structural similarity (MSSIM)

The sliding window is used to divide the image into N image blocks with $H \times W$ size. After weighted calculation of the mean, variance and covariance of each image block, the SSIM of each image block is calculated and its average value is taken as an index to measure the similarity between the two images, which is called MSSIM.

The mean is calculated as follows:

$$\mu_x = \sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij} X(i,j), \tag{16}$$

$$\mu_y = \sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij} Y(i,j). \tag{17}$$

In the above formulas, $X(i,j)$ and $Y(i,j)$ respectively represent the gray values of the reconstructed model and the real model at (i,j), $W(i,j)$ represents the weight of the image block at (i,j) when calculating the MSSIM.

The variance is calculated as follows:

$$\sigma_x = \left(\sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij}(X(i,j) - \mu_x)\right)^{\frac{1}{2}}, \tag{18}$$

$$\sigma_y = \left(\sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij}(Y(i,j) - \mu_y)\right)^{\frac{1}{2}}. \tag{19}$$

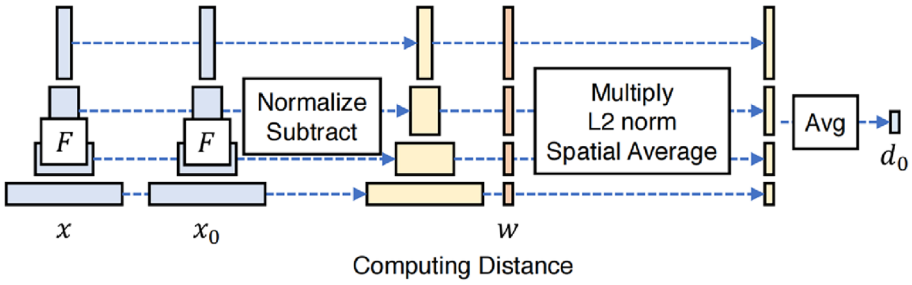

Fig. 4. Computing distance from a network.

The covariance is calculated as follows:

$$\sigma_{xy} = \sum_{i=1}^{H}\sum_{j=1}^{W} w_{ij}(X(i,j) - \mu_x)\big(Y(i,j) - \mu_y\big). \tag{20}$$

Formulas (13)–(18), MSSIM can be expressed as:

$$\begin{aligned}\mathrm{MSSIM} &= \frac{1}{N}\sum_{k=1}^{N} S(x_k, y_k) \\ &= \frac{1}{N}\sum_{k=1}^{N} \frac{(2\mu_{x_k}\mu_{y_k} + C_1)(2\sigma_{x_k y_k} + C_2)}{(\mu_{x_k}{}^2 + \mu_{y_k}{}^2 + C_1)(\sigma_{x_k}{}^2 + \sigma_{y_k}{}^2 + C_2)}.\end{aligned} \tag{21}$$

6.5 Learned perceptual image patch similarity (LPIPS)

LPIPS is an index to measure image similarity [128]. The specific process is as follows: Firstly, deep neural networks (VGG, Alexnet, Squeezenet, etc.) are used to extract features from two inputs (x, x_0). The output of each layer is activated and normalized, denoted as ($\hat{y}^l$, $\hat{y}_0^l$). Then, the weight is assigned by multiplying the vector w points to calculate $L2$ distance. Finally, the average is taken and the sum is calculated layer by layer. The specific calculation process is shown in Figure 4 and formula (20):

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l}\sum_{h,w} \left\| w_l \odot \big(\hat{y}^l_{hw} - \hat{y}^l_{0hw}\big)\right\|_2^2 \tag{22}$$

The LPIPS uses deep convolutional neural network learning (unsupervised, self-supervised and supervised models) to extract features. By calculating the differences of extracted features, LPIPS can obtain image block similarity, which can better simulate the measurement of image similarity by human visual perception system.

6.6 Chamfer distance (CD)

CD can be used as an evaluation index of 3D reconstruction model in 3D space, which is obtained by calculating the distance between two target point sets. It can be used as a loss function for a 3D reconstruction network and is defined as follows:

$$d_{\mathrm{CD}}(S_1, S_2) = \frac{1}{S_1}\sum_{x \epsilon s_1} \min_{y \epsilon s_2} \|x - y\|_2^2 + \frac{1}{S_1}\sum_{y \epsilon s_2} \min_{x \epsilon s_1} \|y - x\|_2^2. \tag{23}$$

S_1 and S_2 respectively represent two groups of point clouds, and the above two items respectively represent the sum of the minimum distance between any point in a point cloud and another point cloud. The smaller the distance, the better the three-dimensional reconstruction effect.

6.7 *F*-score

F-score is an evaluation index that can be used as a classification method in machine learning, and can also be used as an evaluation index of 3D reconstruction algorithm. It is related to accuracy rate and recall rate. The calculation formula is as follows:

$$F = \frac{(\alpha^2 + 1)P \times R}{\alpha^2 \times P + R}. \tag{24}$$

When $\alpha = 1$, F is F_1, also known as the balanced F fraction. $\alpha = 1$ means that P and R have equal weights in the weighted harmonic average, so it is impossible to compare when the distance thresholds used by different methods are different. F_1 is represented by:

$$F_1 = \frac{2 \times P \times R}{P + R}, \tag{25}$$

where P is the accuracy rate, representing the proportion of the number of positive cases with correct classification to the number of predicted positive cases; R is the recall rate, representing the proportion of the number of correctly classified positive cases to the number of all positive cases, defined as follows:

$$P = \frac{TP}{TP + FP}, \tag{26}$$

$$R = \frac{TP}{TP + FN}, \tag{27}$$

where TP, FP and FN are defined by the confusion matrix, as shown in Table 5.

6.8 Intersection over union (IoU)

IoU refers to the proportion of intersection and union of predicted and actual frames, which is usually used in the evaluation of voxel models in the field of 3D reconstruction. In the field of 3D reconstruction, after voxelizing the 3D

Table 5. Confusion matrix.

	Predicted as a positive example	Predicted as negative example
Positive example	TP	FN
Negative example	FP	TN

model, the IoU of reconstructed volume A and real volume B is calculated as the evaluation index of the 3D reconstruction algorithm, and the calculation formula is as follows:

$$\mathrm{IoU} = \frac{A \bigcap B}{A \bigcup B}. \tag{28}$$

6.9 Cross entropy (CE)

The average value of cross entropy loss is defined as follows:

$$\mathrm{CE} = -\frac{1}{N}\sum_{i=1}^{N}\left(x_i \log y_i + (1-x_i)\log\left(1-y_i\right)\right), \tag{29}$$

where N represents the number of voxels in the 3D reconstruction process or the number of points in the point cloud; x_i and y_i represent the true value and predicted value at i respectively. If the cross entropy is lower, the reconstruction effect is better.

6.10 Earth mover's distance (EMD)

In 2000, Rubner et al. proposed a similarity measurement method for image retrieval [129–133], which transformed similarity measurement into a transportation problem. They proposed a histogram similarity measurement method that converts the minimum cost of one normalized distribution into another as an indicator of similarity between two distributions. If the cost is smaller, the similarity is better. In the field of image processing, the EMD idea can be used to evaluate the similarity between two images. Its idea is to calculate the minimum cost of converting from one image to another as the EMD, which represents the similarity between the two images. The smaller the EMD value, the greater the degree of similarity between the two images. Applying EMD to the field of 3D reconstruction [94], the similarity between the reconstruction model and the true value can be obtained by calculating the EMD value. EMD is defined as follows:

$$d_{\mathrm{EMD}}(S_1, S_2) = \min_{\emptyset: S_1 \to S_2} \sum_{x \epsilon S_1} \|x - \emptyset(x)\|_2. \tag{30}$$

6.11 Brief summary

With the further development of 3D reconstruction technology, the evaluation system of 3D reconstruction algorithm needs to be further improved. First of all, some evaluation indexes of 3D reconstruction are only applicable to some specific tasks, such as loU is only applicable to voxel models, and F_1 scores cannot be compared when different methods use different distance thresholds. Moreover, most of the current 3D reconstruction evaluation indexes only focus on the evaluation of 3D reconstructed shapes and ignore the evaluation of texture information. This evaluation index system limits the development of improved 3D reconstruction technology of texture information. Therefore, it is necessary to improve the generalization ability of the existing 3D reconstruction evaluation system.

7 Summary and outlook

Starting with the application of 3D reconstruction technology, this paper systematically introduces the data acquisition mechanism and expression mode of 3D reconstruction. Then, for image-based 3D reconstruction technology, the technical roout, important theories, data sets and evaluation indicators of traditional multi view 3D reconstruction algorithm and deep learning 3D reconstruction algorithm are systematically discussed, and the advantages and disadvantages of relevant Technology roadmap are analyzed in detail. Finally, based on the development status of 3D reconstruction, the relevant technical directions in the field of 3D reconstruction are summarized and prospected.

In this paper, a lot of papers about 3D reconstruction methods are introduced from the perspective of three-dimensional space expression. The traditional methods are becoming more and more mature, but the reconstruction model is prone to holes, texture aliasing and low resolution. The deep learning method represented by NeRF can realize photo-level scene synthesis, and has obvious advantages in detail restoration and no holes, etc. Although it has shortcomings such as generalization ability and real-time performance, real-time integration of virtual and reality can be gradually realized with the development of graphics processor technology and related theories.

Funding

This research was funded by the National Natural Science Foundation of China, number 62171467.

Conflicts of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability statement

This article has no associated data generated.

Author contributions statement

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Yuandong Niu, Siyuan Huang and Shuangyou Chen. The first draft of the manuscript was written by Yuandong Niu. The format and content of drafts are regulated by Limin Liu and Fuyu Huang. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1 Guo J.W., He Y.S., Qi X.Z., Wu G., Hu Y., Li B., Zhang J.W. (2019) Real-time measurement and estimation of the 3D geometry and motion parameters for spatially unknown moving targets, *Aerosp. Sci. Technol.* **97**, 105619.

2 Xu D.F., Zhu Y.K., Choy C.B., Li F.F. (2017) Scene graph generation by iterative message passing, in: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21–26 July.

3 Liu A., Makadia A., Tucker R., Snavely N., Jampani V., Kanazawa V. (2021) Infinite nature: Perpetual view generation of natural scenes from a single image, in: International Conference on Computer Vision, Montreal, Canada, 10–17 October.

4 Fuller A., Fan Z., Day C., Barlow C. (2020) Digital twin: Enabling technologies, challenges and open research, *IEEE Access* **8**, 108952–108971.

5 Tao F., Zhang H., Liu H., Nee A.Y.C. (2019) Digital twin in industry: State-of-the-art, *IEEE Tran. Ind. Inform.* **15**, 2405–2415.

6 Vuković M., Mazzei D., Chessa S., Fantoni G. (2021) Digital twins in industrial IoT: A survey of the state of the art and of relevant standards, in: IEEE International Conference on Communications Workshops, Montreal, Canada, 14–23 June.

7 Weidlich D., Zickner H., Riedel T., Böhm A. (2009) Real 3D geometry and motion data as a basis for virtual design and testing, in: CIRP Design Conference, Cranfield University, 30–31 March.

8 Richter S.R., Alhaija H.A., Koltun V. (2023) Enhancing photorealism enhancement, *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1700–1715.

9 Xue Y., Li Y., Singh K.K., Lee Y.J. (2022) GIRAFFE HD: A high-resolution 3D-aware generative model, in: IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 18–24 June.

10 Tan S., Wong K., Wang S., Manivasagam S., Ren M., Urtasun R. (2021) SceneGen: Learning to generate realistic traffic scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, Nashville, USA, 20–25 June.

11 Fan Y., Lin Z., Saito J., Wang W., Komura T. (2022) FaceFormer: Speech-driven 3D facial animation with transformers, in: IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 18–24 June.

12 Wang J.K., Pun A., Tu J., Manivasagam S., Sadat A., Casas S., Ren M. (2021) AdvSim: Generating safety-critical scenarios for self-driving vehicles, in: IEEE Conference on Computer Vision and Pattern Recognition, Nashville, USA, 20–25 June.

13 Mi L., Zhao H., Nash C., Jin X.H., Gao J.Y., Sun C., Schmid C. (2021) HDMapGen: A hierarchical graph generative model of high definition maps, in: IEEE Conference on Computer Vision and Pattern Recognition, Nashville, USA, 20–25 June.

14 Luo C.Y., Yang X.D., Yuille A. (2021) Self-supervised pillar motion learning for autonomous driving, in: IEEE Conference on Computer Vision and Pattern Recognition, Nashville, USA, 20–25 June.

15 Iwashita S., Murase Y., Yasukawa Y., Kanda S., Sawasaki N., Asada T. (2005) Developing a service robot, in: IEEE International Conference Mechatronics and Automation, Niagara Falls, Canada, 29 July 2005–01 August.

16 Luo Z., Xue W., Chae J., Fu G. (2022) Skp: Semantic 3d keypoint detection for category-level robotic manipulation, *IEEE Robot. Automat. Lett.* **7**, 5437–5444.

17 Zhou Z., Li L., Fürsterling A., Durocher H.J., Mouridsen J., Zhang X. (2022) Learning-based object detection and localization for a mobile robot manipulator in SME production, *Robot. Comput.-Integr. Manuf.* **73**, 102229.

18 Jiang S., Yao W., Wong M.S., Hang M., Hong Z., Kim E.J., Joo S. H., Kuc T.Y. (2019) Automatic elevator button localization using a combined detecting and tracking framework for multi-story navigation, *IEEE Access* **8**, 1118–1134.

19 Xiang L., Gai J., Bao Y., Yu J., Schnable P.S., Tang L. (2023) Field-based robotic leaf angle detection and characterization of maize plants using stereo vision and deep convolutional neural networks, *J. Field Robot.* **40**, 1034–1053.

20 Montoya Angulo A., Pari Pinto L., Sulla Espinoza E., Silva Vidal Y., Supo Colquehuanca E. (2022) Assisted operation of a robotic arm based on stereo vision for positioning near an explosive device, *Robotics* **11**, 100.

21 Vizzo I., Mersch B., Marcuzzi R., Wiesmann L., Behley J., Stachniss C. (2022) Make it dense: Self-supervised geometric scan completion of sparse 3D lidar scans in large outdoor environments, *IEEE Robot. Autom. Lett.* **7**, 8534–8541.

22 Jiang S., Hong Z. (2023) Unexpected Dynamic Obstacle Monocular Detection in the Driver View, *IEEE Intell. Transp. Syst. Mag.* **15**, 68–81.

23 Weerakoon K., Sathyamoorthy A.J., Patel U., Manocha D. (2022) Terp: Reliable planning in uneven outdoor environments using deep reinforcement learning, in: 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, USA, 23–27 May.

24 Duan R., Paudel D.P., Fu C., Lu P. (2022) Stereo orientation prior for UAV robust and accurate visual odometry, *IEEE/ASME Trans. Mechatron.* **27**, 3440–3450.

25 Ding C., Dai Y., Feng X., Zhou Y., Li Q. (2023) Stereo vision SLAM-based 3D reconstruction on UAV development platforms, *J. Electron. Imaging* **32**, 013041.

26 Sumetheeprasit B., Rosales Martinez R., Paul H., Ladig R., Shimonomura K. (2023) Variable baseline and flexible configuration stereo vision using two aerial robots, *Sensors* **23**, 1134.

27 Petrakis G., Antonopoulos A., Tripolitsiotis A., Trigkakis D., Partsinevelos P. (2023) Precision mapping through the stereo vision and geometric transformations in unknown environments, *Earth Sci. Inform.* **16**, 1849–1865.

28 Xie J.Y., You X.Q., Huang Y.Q., Ni Z.R., Wang X.C., Li X.R., Yang C.Y. (2020) 3D-printed integrative probeheads for magnetic resonance, *Nat. Commun.* **11**, 5793.

29 Pang S., Morris D., Radha H. (2022) Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection, in: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, 03–08 January.

30 Downs L., Francis A., Koenig N., Kinman B., Hickman R., Reymann K., McHugh T.B., Vanhoucke V. (2022) Google scanned objects: A high-quality dataset of 3d scanned household items, in: International Conference on Robotics and Automation (ICRA), Philadelphia, USA, 23–27 May.

31 Pirone D., Sirico D., Miccio L., Bianco V., Mugnano M., Ferraro P., Memmolo P. (2022) Speeding up reconstruction of 3D tomograms in holographic flow cytometry via deep learning, *Lab Chip* **22**, 793–804.

32 Jiang S., Tarabalka Y., Yao W., Hong Z., Feng G. (2023) Space-to-speed architecture supporting acceleration on VHR image processing, *ISPRS J. Photogramm. Remote Sens.* **198**, 30–44.

33 Mur-Artal R., Montiel J., Tardós J. (2015) ORB-SLAM: A versatile and accurate monocular SLAM system, *IEEE Trans. Robot.* **31**, 5, 1147–1163.

34 Rosinol R., Leonard J., Carlone L. (2023) NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, USA, 01–05 October.

35 Luo K., Yang G., Xian W., Haraldsson H., Hariharan B., Belongie S. Stay, Positive, (2021) Non-negative image synthesis for augmented reality, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 20–25 June.

36 Charles R.Q., Su H., Kaichun M., Guibas L.J. (2017) PointNet: Deep learning on point sets for 3D classification and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21–26 July.

37 Charles R.Q., Li Y., Hao S., Leonidas J.G. (2017) PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: International Conference on Neural Information Processing Systems, Long Beach, USA, 4–9 December 2017.

38 Fan H., Su H., Guibas L. (2017) A point set generation network for 3D object reconstruction from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21–26 July.

39 Nie Y., Hou J., Han X.G., Nießner M. (2021) RfD-Net: Point scene understanding by semantic instance reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition, Nashville, USA, 20–25 June.

40 Lu Q., Xiao M., Lu Y., Yuan X.H., Yu Y. (2019) Attention-based dense point cloud reconstruction from a single image, *IEEE Access* **7**, 137420–137431.

41 Luo S., Hu W. (2021) Diffusion probabilistic models for 3D point cloud generation, in: IEEE Conference on Computer Vision and Pattern Recognition, Nashville, USA, 20–25 June.

42 Wu Z.R., Song S.R., Khosla A., Yu F., Zhang L.G., Tang X.O., Xiao J.X. (2015) 3D ShapeNets: A deep representation for volumetric shapes, in: IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 07–12 June.

43 Choy C.B., Xu D.F., Gwak J.Y., Chen K., Savarese S. (2016) 3D–R2N2: A unified approach for single and multi-view 3D object reconstruction, in: European Conference on Computer Vision, Amsterdam, Netherlands, 11–14 October.

44 Wu J.J., Zhang C.K., Zhang X.M., Zhang Z.T., Freeman W.T., Tenenbaum J.B. (2018) Learning shape priors for single-view 3D completion and reconstruction, in: European Conference on Computer Vision, Munich, Germany, 8–14 September.

45 Kanazawa A., Tulsiani S., Efros A.A., Malik J. (2018) Learning category-specific mesh reconstruction from image collections, in: European Conference on Computer Vision, Munich, Germany, 8–14 September.

46 Wang N.Y., Zhang Y.D., Li Z.W., Fu Y.W., Liu W., Jiang Y.G. (2018) Pixel2Mesh: Generating 3D mesh models from single RGB images, in: European Conference on Computer Vision, Munich, Germany, 8–14 September.

47 Wen C., Zhang Y.D., Li Z.W., Fu Y.W. (2019) Pixel2Mesh++: Multi-view 3D mesh generation via deformation, in: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October 2019–02 November 2019.

48 Mescheder L., Oechsle M., Niemeyer M., Nowozin S., Geiger A. (2019) Occupancy networks: Learning 3D reconstruction in function space, in: IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 16–20 June.

49 Park J.J., Florence P., Straub J., Newcombe R., Lovegrove S. (2019) DeepSDF: Learning continuous signed distance functions for shape representation, in: IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 16–20 June.

50 Mildenhall B., Srinivasan P.P., Tancik M., Barron J.T., Ramamoorthi R., Ng R. (2020) NeRF: Representing scenes as neural radiance fields for view synthesis, in: European Conference on Computer Vision, Glasgow, UK, 23–28 August.

51 Moravec H.P. (1981) Rover visual obstacle avoidance, in: International Joint Conference on Artificial Intelligence, Vancouver, Canada, 24–28 August.

52 Harris C., Stephens M. (1988) A combined corner and edge detector, in: Alvey Vision Conference, Manchester, UK, 31 August–2 September.

53 Harris C. (1993) Geometry from visual motion, in: *Active vision* **5**, 263–284.

54 Lowe D.G. (1999) Object recognition from local scale-invariant features, in: IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 Sept.

55 Mikolajczyk K., Schmid C. (2001) Indexing based on scale invariant interest points, in: IEEE International Conference on Computer Vision, Vancouver, Canada, 7–14 July.

56 Brown M., Lowe D. (2002) Invariant features from interest point groups, in: British Machine Vision Conference, Cardiff, UK, 2–5 September.

57 Lowe D.G. (2004) Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* **60**, 91–110.

58 Bay H., Ess A., Tuytelaars T., Van Gool L. (2006) SURF: Speeded up robust features, in: European Conference on Computer Vision, Graz, Austria, 7–13 May.

59 Rosten E., Drummond T. (2006) Machine learning for high-speed corner detection, in: European Conference on Computer Vision, Graz, Austria, 7–13 May.

60 Rublee E., Rabaud V., Konolige K., Bradski G. (2011) ORB: An efficient alternative to SIFT or SURF, in: International Conference on Computer Vision, Barcelona, Spain, 06–13 November.

61 Cruz-Mota J., Bogdanova I., Paquier B., Bierlaire., M., Thiran, J. (2012) Scale invariant feature transform on the sphere: Theory and applications, *Int. J. Comput. Vis.* **98**, 217–241.

62 Lakshmi K.D., Vaithiyanathan V. (2016) Image registration techniques based on the scale invariant feature transform, *IETE Tech. Rev.* **34**, 22–29.

63 Al-khafaji S.L., Zhou J., Zia A., Liew A.W. (2018) Spectral-spatial scale invariant feature transform for hyperspectral images, *IEEE Trans. Image Process.* **27**, 837–850.

64 Li M.J., Yuan X.C. (2021) FD-TR: Feature detector based on scale invariant feature transform and bidirectional feature regionalization for digital image watermarking, *Multimed. Tools Appl.* **80**, 32197–32217.

65 Andrade N., Faria F., Cappabianco F. (2018) A practical review on medical image registration: From rigid to deep learning based approaches, in: SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 29 October 2018–01 November 2018.

66 Sedghi A., O'Donnell L.J., Kapur T., Learned-Miller E., Mousavi P., Wells W.M. III (2021) Image registration: Maximum likelihood, minimum entropy and deep learning, *Med. Image Anal.* **69**, 101939.

67 Yu K., Ma J., Hu F.Y., Ma T., Quan S.W., Fang B. (2019) A grayscale weight with window algorithm for infrared and visible image registration, *Infrared Phys. Technol.* **99**, 178–186.

68 Ruppert G.S.R., Favretto F., Falcão A.X., Yasuda C. (2010) Fast and accurate image registration using the multiscale parametric space and grayscale watershed transform, in: International Conference on Systems, Signals and Image Processing, Rio de Janeiro, Brazil, 17–19 June 2010.

69 Mei X., Sun X., Zhou M., Jiao S., Wang H., Zhang X.P. (2011) On building an accurate stereo matching system on graphics hardware, in: *IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, 6–13 November 2011.

70 Bleyer M., Rhemann C., Rother C. (2011) PatchMatch stereo-stereo matching with slanted support windows, in: British Machine Vision Conference, Dundee, UK, 29 August–2 September.

71 Han X.F., Leung T., Jia Y.Q., Sukthankar R., Berg A.C. (2015) MatchNet: Unifying feature and metric learning for patch-based matching, in: IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 07–12 June.

72 Barron J.T., Adams A., Shih Y., Hernández C. (2015) Fast bilateral-space stereo for synthetic defocus, in: IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 07–12 June.

73 Barron J.T., Poole B. (2016) The fast bilateral solver, in: European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October.

74 Žbontar J., LeCun Y. (2015) Computing the stereo matching cost with a convolutional neural network, in: IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 07–12 June.

75 Chen Z.Y., Sun X., Wang Y., Yu Y.N., Huang C. (2015) A deep visual correspondence embedding model for stereo matching costs, in: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 Dec.

76 Žbontar J., LeCun Y. (2016) Stereo matching by training a convolutional neural network to compare image patches, *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 2287–2318.

77 Ye X.Q., Li J.M., Wang H., Huang H.X., Zhang X.L. (2017) Efficient stereo matching leveraging deep local and context information, *IEEE Access* **5**, 18745–18755.

78 Zhang F.H., Prisacariu V., Yang R.G., Torr P.H.S. (2019) GA-Net: Guided aggregation net for end-to-end stereo matching, in: IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 16–20 June.

79 Zhang J.W., Wang X., Bai X., Wang C., Huang L., Chen Y.M., Gu L. (2022) Revisiting domain generalized stereo matching networks from a feature consistency perspective, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 18–24 June.

80 Moulon P., Monasse P., Marlet R. (2013) Global fusion of relative motions for robust, accurate and scalable structure from motion, in: IEEE International Conference on Computer Vision, Sydney, Australia, 01–08 December.

81 Heller J., Havlena M., Jancosek M., Torii A., Pajdla T. (2015) 3D reconstruction from photographs by CMP SfM web service, in: IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May.

82 Schönberger J.L., Frahm J.L. (2016) Structure-from-motion revisited, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 27–30 June.

83 Cui H., Gao X., Shen S.Hu Z. (2017) HSfM: Hybrid structure-from-motion, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 21–26 July.

84 Yin H.Y., Yu H.Y. (2020) Incremental SFM 3D reconstruction based on monocular, in: International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 12–13 December.

85 Wang Y.X., Lu Y.W., Xie Z.H., Lu G.Y. (2021) Deep unsupervised 3D SfM face reconstruction based on massive landmark bundle adjustment, in: Deep Unsupervised 3D SfM Face Reconstruction Based on Massive Landmark Bundle Adjustment. ACM International Conference on Multimedia, New York, United States, 20–24 October.

86 Seitz S.M., Curless B., Diebel J., Scharstein D., Szeliski R. (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 17–22 June.

87 Sinha S., Mordohai P., Pollefeys M. (2007) Multi-View Stereo via Graph Cuts on the Dual of an Adaptive Tetrahedral Mesh, in: 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October.

88 Lin X.B., Wang J.X., Lin C. (2020) Research on 3d reconstruction in binocular stereo vision based on feature point matching method, in: International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 27–29 September.

89 Lindenberger P., Sarlin P.E., Larsson V., Pollefeys M. (2021) Pixel-perfect structure-from-motion with featuremetric refinement, in: IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10–17 Oct.

90 Zhou L., Zhang Z., Jiang H., Sun H., Bao H., Zhang G. (2021) DP-MVS: Detail preserving multi-view surface reconstruction of large-scale scenes, *Remote Sens.* **13**, 4569.

91 Eigen D., Puhrsch C., Fergus R. (2014) Depth map prediction from a single image using a multi-scale deep network, in: International Conference on Neural Information Processing Systems, Cambridge, United States, December 8–13.

92 Eigen D., Fergus R. (2015) Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 Dec.

93 Crispell D., Bazik M. (2017) Pix2face: Direct 3D face model estimation, in: IEEE International Conference on Computer Vision, Venice, Italy, 22–29 Oct.

94 Yao Y., Luo Z., Li S., Fang T., Quan L. (2018) MVSNet: Depth inference for unstructured multi-view stereo, in: European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September.

95 Yao Y., Luo Z., Li S., Shen T., Fang T.Quan L. (2019) Recurrent MVSNet for high-resolution multi-view stereo depth inference, in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 15–20 June.

96 Chen R., Han S., Xu J., Su H. (2019) Point-Based Multi-View Stereo Network, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October 2019–02 November 2019.

97 Zhang J., Yao Y., Li S., Luo Z., Fang T. (2020) Visibility-aware multi-view stereo network, in: The 31st British Machine Vision Virtual Conference, Virtual Conference, 7–10 September.

98 Wei Z., Zhu Q., Min M., Chen Y., Wang G. (2021) AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network, in: The IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10–17 Oct.

99 Peng P., Wang R., Wang Z., Lai Y., Wang R. (2022) Rethinking depth estimation for multi-view stereo: A unified representation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, June 2022, pp. 18–24.

100 Yen-Chen L., Florence P., Barron J., Rodriguez A., Isola P., Lin T. (2021) iNeRF: Inverting neural radiance fields for pose estimation, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September 2021–01 October 2021.

101 Ma L., Li X., Liao J., Zhang Q., Wang X., Wang J., Sander P. (2022) Deblur-NeRF: Neural radiance fields from blurry images, in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, 18–24 June.

102 Xu Qiangeng, Xu Z., Philip J., Bi S., Shu Z., Sunkavalli K., Neumann Ulrich (18–24 June 2022.) *Point-NeRF: Point-based Neural Radiance Fields*, New Orleans, USA.

103 Jiang Y., Hedman P., Mildenhall B., Xu D., Barron J., Wang Z., Xue T. (2023) AligNeRF: High-fidelity neural radiance fields via alignment-aware training, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 18–22 June.

104 Xu L., Xiangli Y., Peng S., Pan X., Zhao N., Theobalt C., Dai B., et al. (2023) Grid-guided neural radiance fields for large urban scenes, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 18–22 June.

105 Stucker C., Schindler K. (2020) ResDepth: Learned residual stereo reconstruction, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, USA, 14–19 June.

106 He K, Zhang X., Ren S, Sun J (2016) Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 27–30 June.

107 Peng S.D., Zhang Y.Q., Xu Y.H., Wang Q.Q., Shuai Q., Bao H.J., Zhou X.W. (2021) Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, Nashville, USA, 19–25 June.

108 Choe J., Im S., Rameau F., Kang M., Kweon I.S. (2021) VolumeFusion: Deep depth fusion for 3d scene reconstruction, in: IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10–17 Oct.

109 Wang D., Cui X.R., Chen X., Zou Z.X., Shi T.Y., Salcudean S., Wang Z.J. (2021) Multi-view 3D reconstruction with transformers, in: IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10–17 Oct.

110 Huang Y.H., He Y., Yuan Y.J., Lai Y.K., Gao L. (2022) StylizedNeRF: Consistent 3D scene stylization as stylized NeRF via 2D–3D mutual learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 18–24 June.

111 Geiger A., Lenz P., Urtasun R. (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 16–21 June.

112 Geiger A., Lenz P., Stiller C., Urtasun R. (2013) Vision meets robotics: The KITTI dataset, *Int. J. Robot. Res.* **32**, 1231–1237.

113 Menze M., Geiger A. (2015) Object scene flow for autonomous vehicles, in: IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 07–12 June.

114 Jensen R.R., Dahl A., Vogiatzis G., Tola E., Aanæs H. (2014) Large scale multi-view stereopsis evaluation, in: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 23–28 June.

115 Aanæs H., Jensen R.R., Vogiatzis G., Tola E., Dahl A.B. (2016) Large-scale data for multiple-view stereopsis, *Int. J. Comput. Vision* **120**, 153–168.

116 Chang A.X., Funkhouser T., Guibas L., Hanrahan P., Huang Q.X., Li Z.M., Savarese S. (2015) *ShapeNet: An information-rich 3d model repository*, pp. 1–11. ArXiv preprint available at https://doi.org/10.48550/arXiv.1512.03012

117 Yi L., Kim V.G., Ceylan D., Shen I., Yan M.Y., Su H., Lu C. (2016) A scalable active framework for region annotation in 3D shape collections, *ACM Trans. Graph.* **35**, 1–12.

118 Dai A., Chang A.X., Savva M., Halber M., Funkhouser T., Nießner M. (2017) ScanNet: Richly-annotated 3d reconstructions of indoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21–26 July.

119 Knapitsch A., Park J., Zhou Q.Y., Koltun V. (2017) Tanks and temples: Benchmarking large-scale scene reconstruction, *ACM Trans. Graph.* **36**, 1–13.

120 Schöps T., Schönberger J.L., Galliani S., Sattler T., Schindler K., Pollefeys M., Geiger A. (2017) A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21–26 July.

121 Huang X.Y., Cheng X.J., Geng Q.C., Cao B.B., Zhou D.F., Wang P., Lin Y.Q. (2018) The apolloscape dataset for autonomous driving, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, USA, 18–22 June.

122 Huang X.Y., Wang P., Cheng X.J., Zhou D.F., Geng Q.C., Yang R. G. (2020) The apolloscape open dataset for autonomous driving and its application, *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 2702–2719.

123 Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. : SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October 2019–02 November 2019.

124 Behley J., Garbade M., Milioto A., Quenzel J., Behnke S., Gall J., Stachniss C. (2021) Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI dataset, *Int. J. Robot. Res.* **40**, 959–967.

125 Yao Y., Luo Z.X., Li S.W., Zhang J.Y., Ren Y.F., Zhou L., Fang T. (2020) BlendedMVS: A large-scale dataset for generalized multi-view stereo networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 13–19 June.

126 Yu F., Chen H.F., Wang X., Xian W.Q., Chen Y.Y., Liu F.C., Madhavan V. (2020) BDD100K: A diverse driving dataset for heterogeneous multitask learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 13–19 June.

127 Caesar H., Bankiti V., Lang A.H., Vora S., Liong V.E., Xu Q., Krishnan A., Pan Y., Baldan G., Beijbom O. (2020) nuScenes: A multimodal dataset for autonomous driving, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 13–19 June.

128 Zhang R., Isola P., Efros A.A., Shechtman E., Wang Q. (2018) The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 18–23 June.

129 Rubner Y., Tomasi C., Guibas L.J. (2000) The earth mover's distance as a metric for image retrieval, *Int. J. Comput. Vision* **40**, 99–121.

130 Zhang C., Cai Y.J., Lin G.S., Shen C.H. (2020) DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers, in: IEEE/CVF conference on computer vision and pattern recognition, Seattle, USA, 13–19 June.

131 Achlioptas P., Diamanti O., Mitliagkas I., Guibas L. (2018) Learning representations and generative models for 3d point clouds, in: International Conference on Machine Learning, Stockholm, Sweden, 10–15 July.

132 Wen C., Yu B.S., Tao D.C. (2021) Learning progressive point embeddings for 3d point cloud generation, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, Nashville, USA, 19–25 June.

133 Zhang C., Cai Y.J., Lin G.S., Shen C.H. (2023) DeepEMD: Differentiable earth mover's distance for few-shot learning, *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 5632–5648.