Computational Visual Media
https://doi.org/10.1007/s41095-021-0250-8 Vol. 8, No. 3, September 2022, 369–393

Review Article

High-quality indoor scene 3D reconstruction with RGB-D cameras: A brief review

Jianwei Li[1], Wei Gao[2,3], Yihong Wu[2,3], Yangdong Liu[4], and Yanfei Shen[1] (✉)

Abstract High-quality 3D reconstruction is an important topic in computer graphics and computer vision with many applications, such as robotics and augmented reality. The advent of consumer RGB-D cameras has made a profound advance in indoor scene reconstruction. For the past few years, researchers have spent significant effort to develop algorithms to capture 3D models with RGB-D cameras. As depth images produced by consumer RGB-D cameras are noisy and incomplete when surfaces are shiny, bright, transparent, or far from the camera, obtaining high-quality 3D scene models is still a challenge for existing systems. We here review high-quality 3D indoor scene reconstruction methods using consumer RGB-D cameras. In this paper, we make comparisons and analyses from the following aspects: (i) depth processing methods in 3D reconstruction are reviewed in terms of enhancement and completion, (ii) ICP-based, feature-based, and hybrid methods of camera pose estimation methods are reviewed, and (iii) surface reconstruction methods are reviewed in terms of surface fusion, optimization, and completion. The performance of state-of-the-art methods is also compared and analyzed. This survey will be useful for researchers who want to follow best practices in designing new high-quality 3D reconstruction methods.

Keywords 3D reconstruction; image processing; camera pose estimation; surface fusion

1 School of Sports Engineering, Beijing Sports University, Beijing 100084, China. E-mail: J. Li, jianwei@bsu.edu.cn; Y. Shen, syf@bsu.edu.cn (✉).
2 National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China. E-mail: wgao@nlpr.ia.ac.cn.
3 University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: yhwu@nlpr.ia.ac.cn.
4 Huawei Technologies Co., Ltd., Beijing 100085, China. E-mail: liuyangdong@huawei.com.
Manuscript received: 2021-05-20; accepted: 2021-07-30

1 Introduction

Real-world 3D reconstruction is a longstanding goal in computer vision. Many tools have been applied to accurately perceive the 3D world, including stereo cameras, laser range finders, monocular cameras, and RGB-D cameras. Advances in consumer RGB-D cameras, such as the Microsoft Kinect, Asus Xtion Live, Intel RealSense, Google Tango, and Occiptial's Structure Sensor, facilitate numerous new and exciting applications, e.g., in augmented reality (AR) to fuse supplementary elements with the real-world environment (e.g., Holoportation [1]), in virtual reality (VR) to provide users with reliable environment perception [2], in digital cultural heritage protection for realistic modeling [3], and in simultaneous localization and mapping (SLAM) for automatic robot navigation. This led to various research into 3D reconstruction with consumer RGB-D cameras. A typical pipeline for RGB-D based 3D scene reconstruction is summarized in Fig. 1, and consists of three modules: image processing, camera pose estimation, and surface reconstruction. Camera pose estimation finds the transformation between two RGB-D images, while surface reconstruction takes RGB-D data as input and fuses the dense overlapping depth frames into one reconstructed model using some specific representation. A complete scene is reconstructed from views acquired along the camera trajectory, while each view covers only a small part of the environment. The pre-processed RGB-D images with estimated camera poses are integrated into a complete 3D scene model.

The advent of affordable consumer grade RGB-D cameras has brought about profound advances in visual scene reconstruction methods. Researchers in the field of computer graphics and computer vision

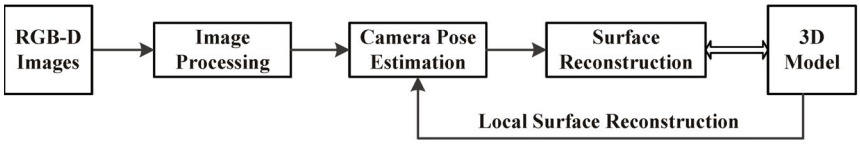


Fig. 1 Pipeline of 3D scene reconstruction with a consumer RGB-D camera.

have expended significant effort to develop entirely new algorithms to capture comprehensive shape models of real-world scenes with RGB-D cameras. Figure 2 gives a brief history of research into indoor scene 3D reconstruction with RGB-D cameras, and indicates some representative methods in the past decade. KinectFusion [4] is a seminal RGB-D based real-time indoor scene 3D reconstruction system. It uses a volumetric representation based on truncated signed distance function (TSDF) [5], in conjunction with fast iterative closest point (ICP) [6] pose estimation to provide a real-time fused dense model. A major limitation of KinectFusion is that camera pose estimation is performed by frame-to-model registration using an ICP algorithm, which is only reliable for RGB-D data with small shifts between consecutive frames acquired by high-frame-rate RGB-D cameras. Since then, improved variants of systems and methods have been proposed. We classify tasks as below and give representative methods:

- Large-scale fusion, e.g., Kintinuous [7], LSD-RGBD SLAM [8], large-scale 3D reconstruction [9, 10], and SG-NN [11].
- Semantic fusion, e.g., SLAM++ [12], automatic semantic modeling [13, 14], SemanticFusion [15], 3D-SIS [16], and SISNet [17].
- Dynamic fusion, e.g., DynamicFusion [18], Fusion4D [19] , FusionMLS [20], and PIFu [21].
- Efficient fusion, e.g., VoxelHashing [22], FastFusion [23], and InfiniTAM [24, 25].
- High-quality fusion, e.g., Redwood [3], BundleFusion [26], Intrinsic3D [27], and UncertaintyAware [28].

Kintinuous [7] extends the work of KinectFusion and creates highly detailed maps of extended scale environments in real time. SLAM++ [12] is the first work on semantic scene reconstruction. It focuses on an implementation of joint 3D object recognition and RGB-D SLAM, and creates semantically meaningful maps by combining geometric and semantic information. Figure 3 shows an example of semantic reconstruction from SemanticFusion [15], which is a real-time visual SLAM system capable of semantically annotating a dense 3D scene using

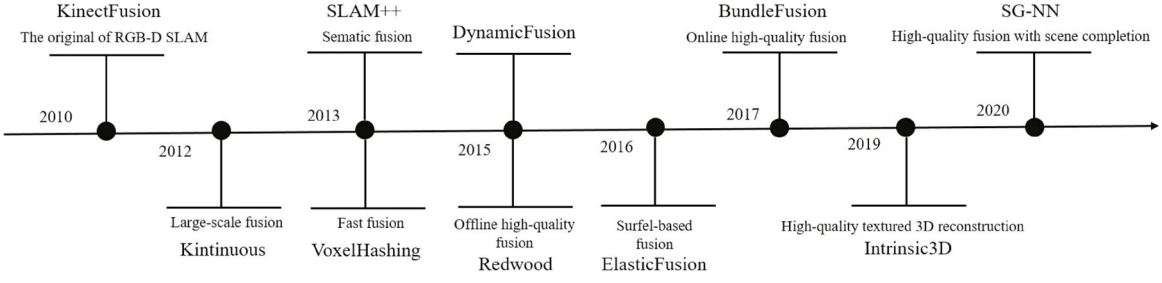


Fig. 2 History of research into 3D scene reconstruction with RGB-D cameras.

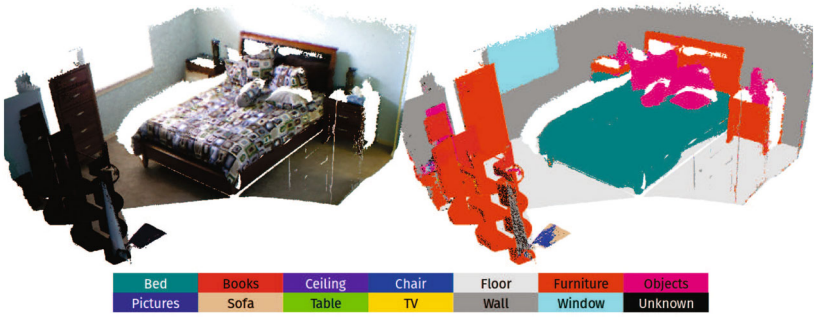


Fig. 3 Semantic reconstruction by semantically annotating a dense 3D scene. Reproduced with permission from Ref. [15], © IEEE 2017.

CNNs. VoxelHashing [22] uses a simple spatial hashing scheme that compresses space, and allows for real-time access and updates of implicit surface data efficiently. For scene reconstruction, an inherent problem is dealing with the tracking drift due to accumulated pose estimation errors. Redwood [3] deals with the accumulated pose estimation errors by reconstructing locally smooth scene fragments and deforming these fragments to align them with each other, obtaining high-quality 3D scene models offline. DynamicFusion [18] presents the first system capable of reconstructing non-rigidly deforming scenes in real time. Figure 4 shows an example of dynamic reconstruction from Fusion4D [19], which is a real-time human volumetric capture system with consumer RGB-D cameras. ElasticFusion [29] proposes surfel-based fusion coupled with frequent model refinement through non-rigid surface deformations. BundleFusion [26] uses additional color features for registration and global bundle adjustment to obtain precise scene geometry in real time. Intrinsic3D [27] obtains high-quality 3D reconstructions by simultaneously optimizing for reconstructed geometry, surface albedo, camera pose, and scene lighting. SG-NN [11] converts partial and noisy RGB-D scans into high-quality 3D scene reconstructions by inferring unobserved scene geometry through self-supervised learning.

In this paper, we focus on high-quality 3D reconstruction of indoor scenes with consumer RGB-D cameras, and review the methods in terms of depth image processing, camera pose estimation, and surface reconstruction. The cited methods focus on articles published in leading conferences and journals in recent years. This review will be useful for researchers who want to follow best practices in designing new high-quality 3D reconstruction methods. The main contributions of our paper are as follows:

1. depth image processing methods in 3D scene reconstruction are analyzed and discussed in terms of depth enhancement and depth completion,
2. camera pose estimation methods are analyzed and discussed in terms of ICP-based, feature-based, and hybrid methods,
3. surface reconstruction methods are analyzed and discussed in terms of surface fusion, surface optimization, and surface completion, and
4. evaluation methods are compared and performance of state-of-the-art systems is analyzed.

The structure of this survey is organized as follows. Section 2 discusses related work on indoor scene 3D reconstruction and gives the motivation for our review. Sections 3–5 review the methods used in 3D reconstruction in terms of image processing, camera pose estimation, and surface reconstruction respectively. Performance of state-of-the-art methods is compared and analyzed in Section 6, while Sections 7 and 8 present a summary and concluding remarks, and consider future developments.

2 Related work

In this section, we provide a brief review of related work in high-quality 3D scene reconstruction methods, RGB-D datasets, benchmarks for 3D scene reconstruction, and related surveys on 3D reconstruction.

2.1 High-quality 3D scene reconstruction

High-quality 3D scene reconstruction aims to obtain complete 3D models with highly-detailed geometry or high-quality surface textures. Existing indoor scene reconstruction methods can be classified as online, i.e., dense SLAM or dynamic reconstruction, or offline, with higher accuracy. To ensure accuracy, low-level geometric and texture information, as well as high-level semantic information, can be used in the reconstruction algorithm. High-quality 3D reconstruction of the real-world is a key component in AR/VR and digital cultural heritage protection. With semantic information, indoor scene

Fig. 4 Dynamic reconstruction of challenging nonrigid sequences. Reproduced with permission from Ref. [19], © Owner/Author 2016.

reconstruction systems have potential applications to intelligent systems like autonomous robot navigation and human–computer interaction. Figure 5 shows examples of high-quality indoor scene reconstructions by accurate geometric registration [3], joint appearance and geometry optimization [30], and semantic segmentation [31], respectively. Due to noisy depth data, inaccurate registration, camera tracking drift, and the lack of accurate surface details, 3D models reconstructed from consumer RGB-D cameras are not yet popularly used in applications. The purpose of this state-of-the-art report is to review current approaches that try to solve this problem.

2.2 Datasets and benchmarks

There are many RGB-D datasets for evaluating real-world and synthetic scene reconstruction methods. We collect and analyze state-of-the-art RGB-D datasets for 3D reconstruction in Table 1, which gives their magnitude, availability of ground truth of camera pose and surface, and semantic annotation. Real-world scenes are scanned by hand-held cameras or robots equipped with RGB-D cameras, while synthetic scenes are obtained by technologies such as rendering and ray tracing. The synthetic ICL-NUIM dataset [35] and real-world TUM RGB-D dataset [32] are two benchmarks widely used to compare and analyze 3D scene reconstruction systems in terms of camera pose estimation and surface reconstruction. Choi et al. [3] provided code and executables to evaluate global registration algorithms for 3D scene reconstruction system, and proposed the augmented (Aug) ICL-NUIM dataset. As can be seen from Table 1, with new applications in robotics and HCI, a trend in RGB-D datasets is towards large-scale scenes with dynamic objects. The newly proposed

Fig. 5 High-quality indoor scene reconstruction with consumer RGB-D cameras by accuracy geometric registration (left) [3], joint appearance and geometry optimization (middle) [30], and semantic segmentation (right) [31]. Reproduced with permission from Ref. [3], © IEEE 2015; Ref. [30], © Springer-Verlag Berlin Heidelberg 2016; Ref. [31], © IEEE 2017.

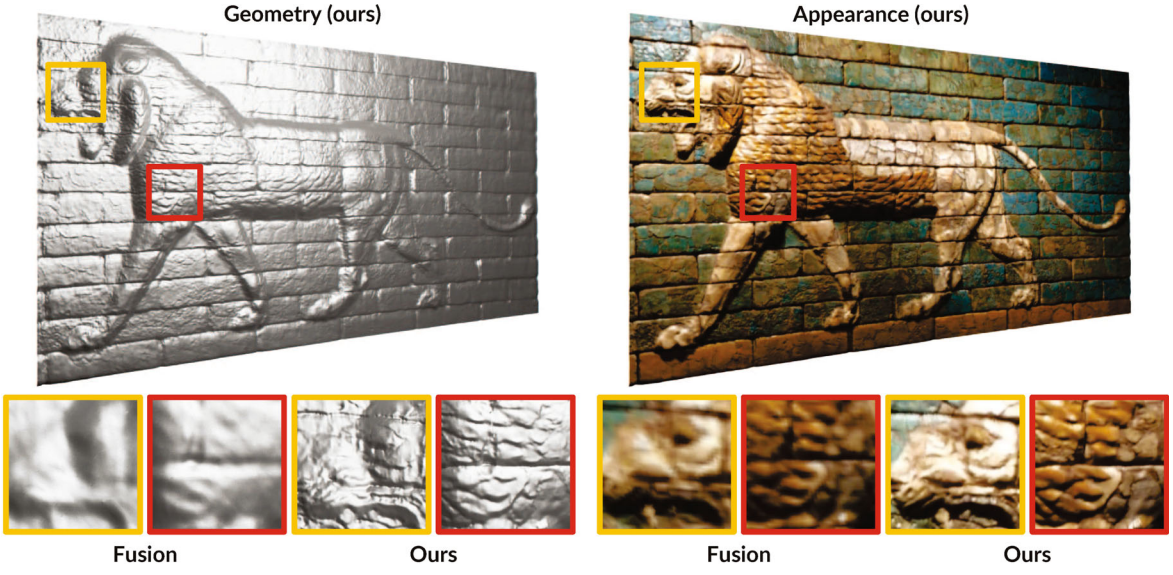


Fig. 6 High-quality 3D reconstruction by joint appearance and geometry optimization. Models have fine-detail geometry (left) and compelling visual appearance (right); close-up views below. Reproduced with permission from Ref. [27], © IEEE 2017.

Table 1 State-of-the-art RGB-D datasets used in 3D reconstruction. √ denotes conforming description, while × denotes nonconforming description; ⋆ denotes partly conforming description

Datasets	Number	Real-world	Pose	GT surface	Semantic	Dynamic	Year
TUM benchmark [32]	47 sequences	√	√	×	×	⋆	2012
Stanford 3D Scene [33]	7 sequences	√	√	√	×	×	2013
SUN 3D [34]	415 scenes	√	√	×	12 categories	×	2014
ICL-NUIM [35]	8 sequences	×	√	√	×	×	2014
Aug ICL-NUIM [3]	4 sequences	×	√	√	×	×	2015
SceneNN [36]	100 scenes	√	√	√	50 categories	×	2016
CoRBS [37]	20 sequences	√	√	√	×	×	2016
Matterport3D [38]	90 scenes	√	√	√	40 categories	×	2017
Scannet [31]	1513 scenes	√	√	×	21 categories	×	2017
SceneNet RGB-D [39]	57 scenes	×	√	√	√	×	2017
Bonn RGB-D [40]	24 sequences	√	√	×	×	√	2019
InteriorNet [41]	15k sequences	×	√	×	40 categories	×	2019
Replica [42]	18 scenes	×	√	√	88 categories	√	2019
OpenLORIS-Scene [43]	22 sequences	√	√	×	√	√	2020

replica dataset [42] contains 18 highly photo-realistic 3D indoor scene reconstructions at room and building scale, in which each scene consists of a dense mesh, high-dynamic-range (HDR) textures, semantic class, instance information, and so on. The corresponding benchmarks need to be further standardized to push the development of high-quality 3D reconstruction.

2.3 Other surveys

There are several surveys related to our work. Berger et al. [45] surveyed the field of surface reconstruction, and provided a categorization with respect to priors, data imperfections, and reconstruction output. Chen et al. [46] provided an overview of recent advances in indoor scene modeling techniques, as well as public datasets and code libraries which can facilitate experiments and evaluation. Stotko [47] reviewed several registration algorithms developed in recent years and compared their performance. Xu et al. [48] gave an overview of the main concepts and components of data-driven shape analysis and processing techniques. Recently, Zollhöfer et al. [49] presented a survey of the state-of-the-art in 3D reconstruction with RGB-D cameras, and reviewed the recent developments in RGB-D scene reconstruction for static and dynamic scenes. Han et al. [50] reviewed the state-of-the-art and trends in 3D object reconstruction in the deep learning era. Roldão et al. [51] identified, compared, and analyzed techniques of semantic scene completion (SSC) for both methods and datasets. Recently, Liu et al. [52] covered SLAM related datasets, including an overview and comparison of existing datasets, review of evaluation criteria, and discussions of current limitations and future directions. The above surveys do not specifically analyze the influencing factors in high-quality 3D scene reconstruction methods. Thus, we give the first comprehensive and critical review of high-quality indoor scene 3D reconstruction with RGB-D cameras, focusing on image processing, camera pose estimation, surface reconstruction, and performance comparison, providing a summary and discussion, and looking ahead to future trends.

3 Depth image processing

As Fig. 1 shows, depth image processing is the first stage of 3D scene reconstruction. Consumer RGB-D cameras employ one of two main approaches to depth sensing, triangulation and time-of-flight (ToF). Triangulation is realized by structured light, an active system which projects an infrared light pattern onto the scene and estimates the disparity given by the perspective distortion of the pattern due to variations in the object's depth. ToF cameras measure the time that light emitted by an illumination unit requires to travel to an object and back to a detector. Consumer grade RGB-D cameras relying on these methods often suffer from significant noise and distortion, and cannot capture subtle details. The raw depth images have to be taken into account in algorithm development for high-quality 3D reconstruction.

3.1 Depth enhancement

To enhance the quality of depth images used in 3D reconstruction, many approaches focus on depth denoising and depth super-resolution. Researchers also have exploited techniques of shape from shading (SfS) and shape from polarization (SfP) for depth images.

3.1.1 Depth denoising

The noise of depth images captured by consumer RGB-D cameras depends on a variety of parameters, such as the distance to the acquired object, and pixel position in the depth image. Many researchers have evaluated and analyzed the accuracy of depth images [53–55]. A commonly used method is bilateral filtering [56] which can effectively smooth depth images and is widely used in RGB-D based 3D reconstruction systems. Li et al. [57] processed depth images with a depth adaptive bilateral filter to effectively improve the accuracy of 3D scene models. In recent years, deep depth denoising techniques (e.g., Ref. [58]) which can better capture the global context of each scene have attracted more attention.

3.1.2 Depth super-resolution

Consumer RGB-D cameras can capture high resolution (HR) RGB images (e.g., 1280×1024), but only low resolution (LR) depth images (e.g., 640×480). In order to facilitate reconstruction, both RGB images and depth images are used at low resolution in most 3D reconstruction. Super-resolution techniques improve the observed low resolution images to corresponding high resolution images: high-resolution depth maps can be inferred from low-resolution depth measurements and an additional high-resolution intensity image of the same scene. Although there are depth super-resolution techniques without color information (e.g., example-based methods [59]), most existing methods improve the resolution of depth images using high-resolution color images [60–62]. Some deep learning depth super-resolution techniques [63–65] also exist to improve the resolution of depth images. Hui et al. [63] used two CNNs to downsample an HR image concurrently with upsampling the LR depth image: after the generation of RGB features from the downsampling CNN, these features were used to fine-tune the upsampling of the depth images. Riegler et al. [65] used an energy minimization model to guide the model for generating HR depth images without the need for reference images.

3.1.3 Shading-based methods

Shape from shading [66] deals with the recovery of shape from the gradual variation of shading in the images. This method is capable of capturing high quality shape details of a dynamic object under natural illumination, and is widely used to enhance depth images from consumer grade RGB-D cameras [67–70]. For instance, Han et al. [67] estimated detailed shape of diffuse objects with uniform albedo from a single RGB-D image. Yu et al. [68] presented a shading-based shape refinement algorithm which uses a noisy, incomplete depth image from Kinect to help resolve ambiguities in SfS. Wu et al. [69] presented the first real-time method for refinement of depth images using SfS in general uncontrolled scenes with consumer RGB-D cameras. RGBD-fusion [70] uses a lighting model to handle natural scene illumination, and enhances the depth image by fusing intensity and depth information to create more detailed range profiles. Nevertheless, the robustness of SfS methods is limited due to use of an illumination model.

3.1.4 Polarization-based methods

Shape from polarization is an application of polarization imaging and aims to digitize the shape of the observed object. Polarization reveals surface normal information, and is thus helpful to propagate depth to featureless regions. Researchers have exploited polarization techniques [44, 71–73] to enhance depth images. Polarized 3D [44] enhances coarse depth images by using shape information from polarization cues; an experimental result of this method is shown in Fig. 7, which compares shading enhancement and polarization enhancement. Cui et al. [71] combined per-pixel photometric information from polarization and obtained good reconstruction performance especially on featureless 3D objects. Deep SfP [72] makes the first attempt to bring the SfP problem to the realm of deep learning and performs well. Since then, many methods (e.g., Ref. [73]) have tried to combine deep learning with polarization techniques for depth enhancement. The equipment used in polarization-based methods is expensive due to use of polarization technology, limiting its wider application to in 3D reconstruction.

As can be seen from the above, depth enhancement approaches have been applied in RGB-D reconstruc-

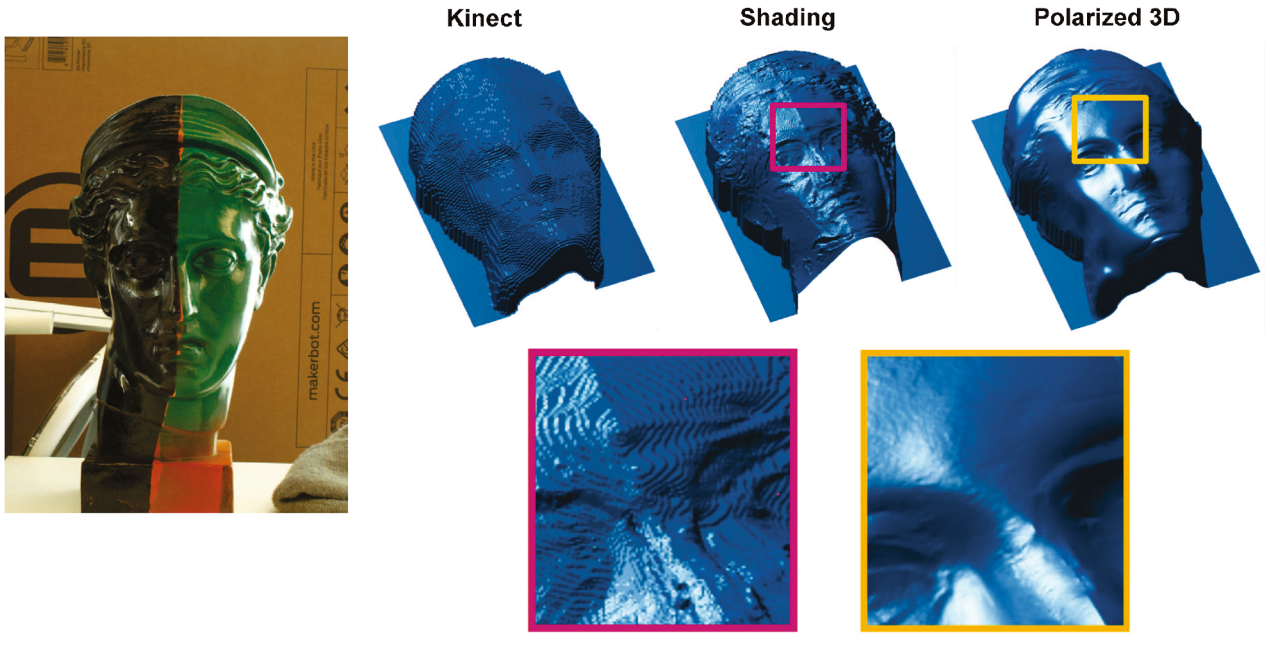


Fig. 7 High-quality 3D reconstruction by depth enhancement. The enlarged views show the results of shading enhancement (red box) and polarization enhancement (yellow box). Reproduced with permission from Ref. [44], © IEEE 2015.

tion; most techniques belong to traditional image processing and have been widely applied in practice.

3.2 Depth completion

Raw depth images produced from consumer grade RGB-D cameras are often incomplete when surfaces are shiny, bright, transparent, or far from the camera. To addressing this problem, various approaches have emerged which try to complete the sparse depth measurements into a dense depth image. Figure 8 shows an example of depth completion [74] for an indoor scene. Techniques to complete depth data of RGB-D images can be divided into traditional and data-driven methods.

3.2.1 Traditional methods

A few early works addressed depth completion through image filtering or optimization. To fill in holes in a raw depth image, NYU v2 [75] uses cross-bilateral filtering to produce a visually pleasing depth map, but introduces artifacts. Xiao et al. [34] improved the depth image by using TSDF to voxelize the space, accumulating the depth map from nearby frames using camera poses, and then used ray casting to get a reliable depth image. Chen and Koltun [76] developed a global high-resolution MRF optimization approach to improve the accuracy of depth images. These methods use traditional image processing algorithms for depth completion, but their prediction ability is limited given large data loss.

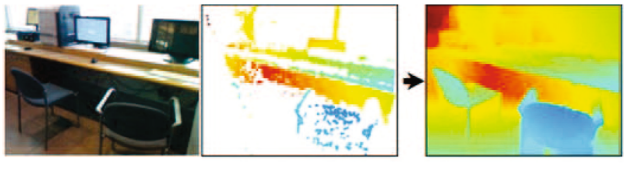


Fig. 8 Depth completion. The inputs are the RGB image and corresponding sparse depth image; the output is the completed depth image. Reproduced with permission from Ref. [74].

3.2.2 Data-driven methods

With recent advances in deep learning and the availability of various RGB-D datasets, researchers have started to look at data-driven approaches to depth estimation. Most algorithms [74, 77–82] utilize the RGB image and additional information that can be inferred from the depth map, such as surface normal, to give geometrical guidance to the training process. Sparse-to-dense [77] first introduced a robust and accurate depth estimation method from RGB images with additional sparse depth samples acquired from a low-resolution depth sensor; it was used in a SLAM system. Later, Chen et al. [78] presented a deep model that can accurately produce dense depth images given an RGB image with known depths at a very sparse set of pixels. Zhang and Funkhouser [74] trained a deep network to predict dense surface normals and occlusion boundaries, and combined

those predictions with raw depth observations to solve for depths for all pixels, including those missing in the original observation. To address depth smearing between objects, Imran et al. [83] proposed a depth coefficient representation which enables convolutions to more easily avoid inter-object depth mixing. In recent work, Zhu et al. [84] introduced a local implicit neural representation built on ray-voxel pairs that allows generalization to unseen transparent objects and provides fast inferencing.

Research into depth completion has developed with the advent of consumer RGB-D cameras, and most existing approaches focus on deep learning. To train deep networks, a large corpus of training data with accurate ground-truth is required. This is limits the application of data-driven depth completion methods to 3D reconstruction. Li et al. [85] attempted to obtain high-quality 3D reconstruction with depth super-resolution and completion, and evaluated its feasibility on the synthetic ICL-NUIM dataset, but application to real-world scenes remains to be studied.

4 Camera pose estimation

In 3D reconstruction, the goal of camera pose estimation is to find the transformation $\boldsymbol{T}$ between two images. To obtain accurate camera poses, a complete estimation pipeline often contains two phases: (i) front-end camera tracking (e.g., frame-to-frame tracking [86] or frame-to-model tracking [4]), and (ii) back-end optimization (e.g., loop closure and global optimization [3, 87, 88]). According to the tracking characteristic, camera pose estimation methods can be divided into ICP-based and feature-based frameworks. In the following, we discuss both, and hybrid methods.

4.1 ICP-based methods

ICP-based methods estimate camera pose by maximizing the consistency of geometric information as well as color information between pairs of adjacent frames. The ICP algorithm introduced by Besl and Mckay [89] is a popular method for 3D reconstruction with RGB-D cameras. It aligns two partially overlapping point clouds given an initial guess for the relative transform. Each point in one data set is paired with the closest point in the other data set to form correspondence pairs. Given two scene scans $\boldsymbol{P}$ and $\boldsymbol{Q}$, the transformation $\boldsymbol{T} = [\boldsymbol{R} \mid \boldsymbol{t}]$ between them is estimated by minimising:

$$E_{\text{icp}} = \sum_i \|\boldsymbol{p}_i - \boldsymbol{R}\boldsymbol{q}_i - \boldsymbol{t}\|^2 \tag{1}$$

where $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ are the points from $\boldsymbol{P}$ and $\boldsymbol{Q}$ respectively. This error metric is the sum of the squared distances between points in each correspondence pair. Figure 9 illustrates the ICP algorithm using point-to-plane errors: $\arg\min_{\boldsymbol{T}} \Sigma_i[(\boldsymbol{T}\cdot\boldsymbol{s}_i - \boldsymbol{d}_i)\cdot\boldsymbol{n}_i]^2$, where $\boldsymbol{s}_i$ is a source point, $\boldsymbol{d}_i$ is the corresponding destination point, and $\boldsymbol{n}_i$ is the unit normal vector at $\boldsymbol{d}_i$.

This process is iterated until the error becomes smaller than a threshold or it stops changing. In scenes containing textureless regions (e.g., walls and floors), depth information alone is insufficient to compute the camera pose. The direct method reported in dense visual odometry (DVO) SLAM [91] uses color information to overcome this issue. The goal of the direct method is to estimate the camera motion such that the warped second image matches the first image based on the photo-consistency assumption; Fig. 10 shows this process for two images. The photometric error E_{rgb} is defined as

$$E_{\text{rgb}} = \sum_i (\boldsymbol{I}_1(w(\xi, \boldsymbol{p}_i)) - \boldsymbol{I}_2(\boldsymbol{p}_i))^2 \tag{2}$$

where ξ is the camera motion, $\boldsymbol{p}_i$ is a pixel point, and w is the warping function that matches the current

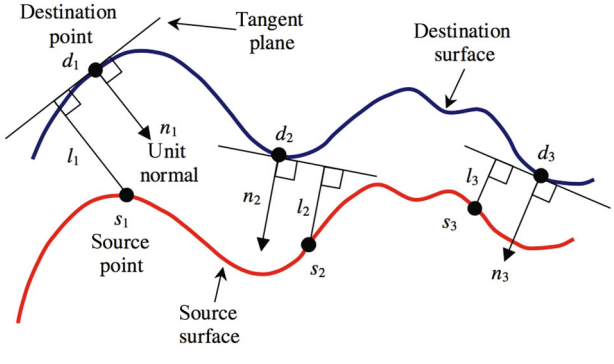


Fig. 9 Point-to-plane error between two surfaces in an iterative closest point algorithm. Reproduced with permission from Ref. [90].

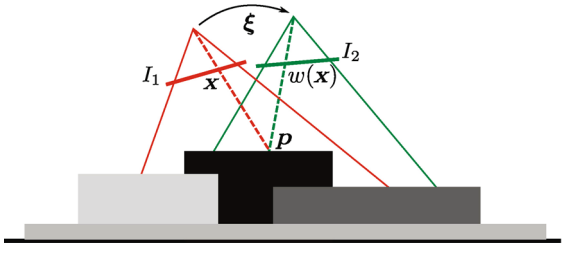


Fig. 10 Photo-consistency assumption in the direct method. Reproduced with permission from Ref. [91], © IEEE 2013.

image $\boldsymbol{I}_2$ to the previous image $\boldsymbol{I}_1$. DVO SLAM estimates the camera pose combining geometric error and photometric error, in what it calls combined ICP [92]. The combined error E_{combined} is given by Eq. (3):

$$E_{\text{combined}} = E_{\text{icp}} + \lambda E_{\text{rgb}} \tag{3}$$

where λ is a weight. Both error functions use the same correspondences and their limitations do not affect each other. This idea is further used in Kintinuous and ElasticFusion.

In addition to combined ICP, variants of ICP methods (e.g., Color-ICP [93], efficient ICP [6], non-rigid ICP [94], generalized-ICP [95], NICP [96]) have been proposed. For instance, non-rigid ICP is capable of modeling nonrigid objects. Generalized-ICP constructs point-to-point, point-to-plane, and plane-to-plane error metrics. In recent years, ICP-based methods combined with deep learning also have been proposed. Deep closest point (DCP) [97] replaces the Euclidean nearest point step of ICP by a learnable per-point embedding network, followed by a high-dimensional feature-matching. Following DCP, many iterative methods [98, 99] extend the feature matching idea, where the general scheme is to learn the mapping, apply the inferred transformation to the source point cloud, and learn a new alignment map, until convergence. Recently, deep weighted consensus [100] presents a new paradigm for rigid alignment based on a learnable weighted consensus which is robust to noise.

4.2 Feature-based methods

Feature-based methods introduce RGB features into camera pose estimation by maximizing the 3D position consistency of corresponding feature points between frames, to improve the robustness of camera tracking. Existing 3D reconstruction systems within an SLAM framework often use sparse features to establish 2D–3D matches between features in a query image and points in a 3D map. In general, the transformation $\boldsymbol{T}$ is estimated using feature re-projection error:

$$E_{\text{feature}} = \boldsymbol{x}' - \boldsymbol{K}\boldsymbol{T}\boldsymbol{x} \tag{4}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ denote the position of a 3D feature and the matched feature respectively, and $\boldsymbol{K}$ is the camera intrinsic matrix.

Point features are a popular choice for feature extraction and matching in 3D reconstruction, such as SIFT [34], FPFH [3], ORB [101], and some learned features (e.g., 3DMatch [87] and PixLoc [102]). Lines and planes are the most common structures used in indoor scenes and are less sensitive to lighting variation than points. Researchers are increasingly studying methods [103–109] to use them for high-quality 3D reconstruction. Such methods can generally achieve good performance under both constant and varying lighting conditions.

Based on *line features*, Choi et al. [103] presented a 3D edge detection approach for RGB-D point clouds, which can exploit the organized structure of the RGB-D image to efficiently detect edges, and make use of both 3D shape information and photometric texture information. Lu and Song [104] fused point and line features to form a robust RGB-D visual odometry algorithm, which extracts 3D points and lines from RGB-D images, analyzes their measurement uncertainties, and computes camera motion using maximum likelihood estimation. Zhou and Koltun [105] proposed a depth camera tracking method with contour cues, which can be used to establish correspondence constraints that carry information about scene geometry and constrain pose estimation. The contour constraints reliably improve camera tracking accuracy.

Based on *plane features*, Taguchi et al. [106] presented a point–plane 3D reconstruction system, which uses the minimal set of features in an RANSAC framework to robustly compute correspondences and estimate camera pose. Dense planar SLAM [107] densely maps the environment using bounded planes and surfels extracted from depth images. It takes advantage directly of the planarity of many parts of the scenes via a data-driven process to directly regularize planar regions and represent their accurate extent efficiently using an occupancy approach with on-line compression. CPA-SLAM [108] consistently integrates frame-to-keyframe and frame-to-plane alignment, and models the environment with a global plane model. It makes use of the dense image information available in keyframes for accurate short-term camera tracking and uses the global model to reduce drift. PlaneMatch [109] densely models the environment with plane information through a CNN that takes in RGB, depth, and normal information of a planar patch in an image, and outputs a descriptor to find coplanar patches in other images for scene reconstruction.

4.3 Hybrid methods

In practical applications, feature-based methods often combine multiple features (e.g., points, edges, lines, and planes) to improve camera tracking stability. For instance, Manhattan SLAM [110] makes use of point, line, and plane features for robust tracking in challenging scenes, allowing for accurate camera tracking and efficient dense mapping. Generally speaking, feature-based methods are better than ICP-based ones at handling RGB-D data with large shifts, since they simply run a quadratic minimization problem to directly compute the relative transformation between two consecutive frames.

To obtain high-quality 3D scene models robustly, some systems estimate the camera pose combining ICP-based methods and feature-based methods in hybrid methods. SDF-2-SDF [86] proposes an implicit-to-implicit surface registration scheme, and can be utilized both for frame-to-frame camera tracking and global optimization. BundleFusion employs correspondences based on sparse features and dense geometric and photometric matching, and obtains a highly accurate camera pose. Kehl et al. [111] formulated a joint contour and ICP tracking approach. 3D match [87] presents a data-driven model that learns a local volumetric patch descriptor for establishing correspondences between partial 3D data. Semantic information [112] and optical flow [113] also have been used in camera pose estimation. Schönberger et al. [112] proposed the first semantic visual localization method, which is robust to missing observations where previous approaches failed. GeoNet [113] estimates dense depths, optical flow, and camera pose using unsupervised learning. Recently, Tang et al. [114] estimated camera pose using dense scene matching (DSM), where a cost volume is constructed between a query image and a scene. The cost volume and the corresponding coordinates are processed by a CNN to predict dense coordinates.

5 Surface reconstruction

Surface reconstruction fuses RGB-D images from different camera views into a complete 3D model. In this section, we consider surface reconstruction methods in terms of surface fusion, surface optimization, and surface completion.

5.1 Surface fusion

The basic surface fusion approaches for dense 3D reconstruction are volume-based or surfel-based. Existing high-quality 3D reconstruction systems [3, 25, 26, 115] are mainly based on these or their improvements.

5.1.1 Volume-based fusion

Volume-based fusion provides efficient and simple ways of integrating multiple RGB-D images into a complete 3D model. In a volume-based framework, TSDF is discretized into a voxel grid to represent a physical volume of space: see Fig. 11. On the left is a two-dimensional example of signed distance values stored at voxels within the truncation distance of the observed surface, with rays cast from the observing sensor, and on the right is the voxel grid underlying the reconstruction volume. Each voxel contains a signed distance function (SDF) indicating the distance from the cell to a surface and a weight representing confidence in the accuracy of the distance.

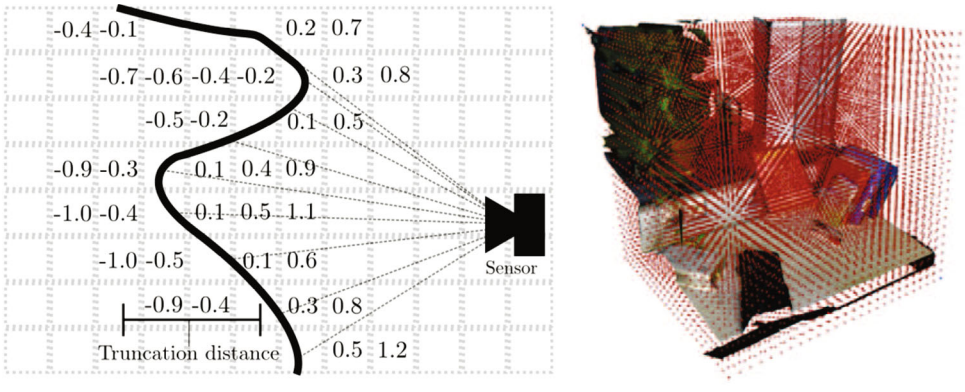


Fig. 11 Two-dimensional example of TSDF representation and the voxel grid underlying the reconstruction volume. Reproduced with permission from Ref. [8], © The Author(s) 2014.

For a given voxel $\boldsymbol{v}$ in the fused scene model F, the update of signed distance value $F(\boldsymbol{v})$ is defined by

$$\begin{cases} F'(\boldsymbol{v}) = \frac{F(\boldsymbol{v})W(\boldsymbol{v}) \pm f_i(\boldsymbol{v})w_i(\boldsymbol{v})}{W(\boldsymbol{v}) \pm w_i(\boldsymbol{v})} \\ W'(\boldsymbol{v}) = W(\boldsymbol{v}) \pm w_i(\boldsymbol{v}) \end{cases} \tag{5}$$

where addition is used for integration, and subtraction for de-integration. The signed distance function $f_i(\boldsymbol{v})$ is the projective distance between a voxel and the ith depth frame, and weighting function $w_i(\boldsymbol{v})$ represents the confidence in the accuracy of the distance.

The original idea of volumetric 3D reconstruction from depth images dates back to Ref. [5]. Later, the advent of consumer RGB-D cameras and massively parallel processors in GPUs led to the seminal KinectFusion system, which inspired a wide range of further work. Volume-based fusion is at the core of many state-of-the-art RGB-D reconstruction frameworks [3, 4, 22, 26, 116]. One disadvantage of volume-based fusion is its large memory footprint, as the required memory grows linearly with the overall volume that is represented rather than with its surface area. This issue has been addressed by sparse volumetric representations, such as multi-scale octrees [23, 117] and hierarchical structures [24]. For non-rigid fusion, embedded deformation for the shape manipulation algorithm [118] has been introduced in some volume-based dynamic fusion systems (e.g., DynamicFusion [18] and DeepHuman [119]). 3D deep learning of volumetric methods, such as deep implicit representation (DIR) [120–122], have also been proposed and studied. For instance, DeepSDF [120] introduces a learned continuous SDF representation of a class of shapes that enables high-quality shape representation, interpolation, and completion from partial and noisy 3D input data. Scene representation networks (SRNs) [121] represent scenes as continuous functions that map world coordinates to a feature representation by encoding both geometry and appearance. Neural sparse voxel fields (NSVF) [122] defined a set of voxel-bounded implicit fields organized in a sparse voxel octree to model local properties, and can successfully represent complex 3D scenes.

5.1.2 *Surfel-based fusion*

Surfel-based fusion is a powerful paradigm to efficiently render complex geometric objects. Figure 12 shows a surfel (surface element) representation [123] in object space and texture space. The maximum

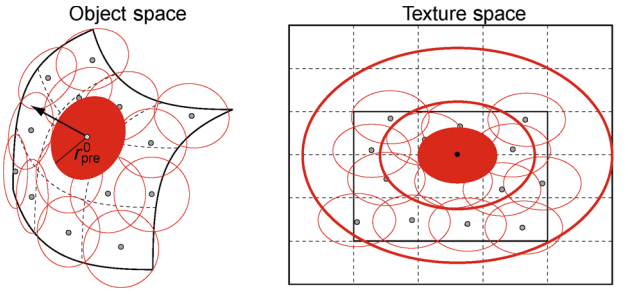


Fig. 12 Surfels in object space and texture space. The tangent disks in object space are mapped to ellipses in texture space using the predefined texture parameterization of the surface. Reproduced with permission from Ref. [123], © ACM 2000.

distance between adjacent surfels in object space is the radius (r^0_{pre}) of the tangent disk. A surfel is a point sample of an object's surface that includes geometric attributes such as position and normal as well as photometric attributes such as a diffuse color.

During surface fusion, for a given surfel M^{s} with a position $p \in \mathbb{R}^3$, normal $n \in \mathbb{R}^3$, colour $c \in \mathbb{R}^3$, weight $w \in \mathbb{R}$, and radius $r \in \mathbb{R}$, the update rules for each component are

$$\begin{cases} \hat{p} = \dfrac{wp + w'p'}{w + w'} \\ \hat{n} = \dfrac{wn + w'n'}{w + w'} \\ \hat{r} = \dfrac{wr + w'c'}{w + w'} \\ \hat{w} = w + w' \end{cases} \tag{6}$$

where the prime superscript (e.g., p') and hat operator (e.g., $\hat{p}$) denote the newly associated measurement and new updated value for a given surfel respectively.

Andersen et al. [124] proposed a surfel-based geometry reconstruction method for determining a piecewise smooth surface from noisy data. Surfels are well suited to modeling dynamic geometry, because there is no need to compute topological information such as adjacency lists. This surfel-based fusion strategy is used by several reconstruction systems, such as dynamic scenes reconstruction [125], dense planar SLAM [107], ElasticFusion, and InfiniTAM v3 [25]. Based on ElasticFusion, SemanticFusion allows semantic predictions from multiple view points to be probabilistically fused into a semantic map. GravityFusion [115] incorporates gravity measurements into the surfels to avoid the typical curving of 3D maps in long hallways. DeepSurfels [126] combines explicit and neural building blocks to jointly encode geometry and appearance information,

and has better scalability to larger scenes than existing methods. Through combination with prior information and deep learning, these methods have improved reconstruction performance for indoor scenes. Further, point-based representation [125] is also simple but efficient in terms of memory requirements. A 3D shape can be represented using an unordered set $S = (x_i; y_i; z_i)_{i=1}^{N}$ with N points. It is well suited for objects with interacting parts and fine details. Therefore, many papers on point-based 3D object reconstruction, such as DensePCR [127], have appeared over recent years.

5.2 Surface optimization

An initial 3D model reconstructed using consumer RGB-D cameras often contains noisy geometry, and blurred surface textures. Surface optimization is a classical task in 3D reconstruction in computer vision. In the following, we describe, analyze, and classify methods that have been proposed for surface optimization during recent years.

5.2.1 Shape denoising

Shape denoising techniques can be applied to points (e.g., Refs. [128, 129]), meshes (e.g., Refs. [130, 131]), and surfaces (e.g., Refs. [132, 133]) to improve the quality of 3D models. Wolff et al. [128] removed noise and geometrically or photometrically inconsistent outliers in a point cloud. Wang et al. [131] presented a data-driven approach for mesh denoising via cascaded normal regression. High-frequency details are added to the coarse base mesh using color and displacement maps. Schertler et al. [132] proposed a field-aligned online surface reconstruction algorithm that sidesteps the signed-distance computation of classical reconstruction techniques in favor of direct filtering, parametrization, and mesh and texture extraction. Tsai et al. [133] proposed a surface optimization framework for non-line-of-sight imaging. Shape denoising algorithms abound in computer vision and computer graphics, but most of them are suitable for object denoising. Scene surface denoising is still challenging and needs to be further explored.

5.2.2 Surface refinement

Methods used in 3D reconstruction mainly include shading-based geometry refinement [136, 137], joint appearance and geometry optimization [27], and deep learning [138]. Representative shading-based work is VSBR [136], which obtains fine-scale detail through volumetric shading-based refinement of a distance field to solve the problem of over-smoothing in RGB-D reconstructions. To obtain high-quality 3D reconstructions, Intrinsic3D [27] introduces a simultaneous optimization method for geometry encoded in an SDF, and textures from automatically-selected key-frames. It dramatically increases the level of detail in the reconstructed scene geometry and contributes highly to consistent surface texture recovery. DECOR-GAN [138] details 3D shapes by conditional refinement through a generative adversarial network (GAN), which can refine a coarse shape into a variety of detailed shapes with different styles.

5.2.3 Color textures

Image-based texture mapping is a common way of producing texture maps for 3D geometric models. Although a high-quality texture map can be easily computed for accurate geometry and calibrated cameras, texture map quality degrades significantly in the presence of inaccuracies. Researchers have explored several methods [30, 135, 139, 140] for high-quality texture maps. The large-scale scene model with texture map shown in Fig. 5(middle) was acquired by optimizing the texture coordinates of the 3D model to maximize photometric consistency among multiple key frames [30]. 3DLite [135] extrapolates high-level scene geometry, and uses image inpainting to generate sharp surface textures. Liu et al. [139] realized high-quality textured 3D shape reconstruction with cascaded fully convolutional networks. Recently, Huang et al. [140] proposed an approach to produce photo-realistic textures for approximate surfaces even from misaligned images by learning an objective function. Reconstructed scene models with realistic color textures are very useful in AR/VR and digitization of cultural heritage; research using consumer RGB-D cameras is of great interest and remains challenging in practice.

5.3 Surface completion

3D models are quite often incomplete due to occlusion between objects. Surface completion is used to recover a complete object or scene model from one or more images. Inferring a dense 3D scene from 2D or sparse 3D inputs is in fact an ill-posed problem since the input data are insufficient to resolve all ambiguities. Most existing works rely on deep learning to learn semantics and geometric priors from large scale

datasets. Figure 13 compares object completion and scene completion approaches. Initial reconstructions are shown on the left while the completed surface models are shown on the right. We next discuss object and scene completion methods in turn.

5.3.1 Object completion

Traditional object completion methods [134, 141–143] fill small holes by detecting structures and regularities in 3D shapes. Davis et al. [141] addressed situations in which the holes are too geometrically and topologically complex to fill using triangulation algorithms, and applied a diffusion process to extend SDF throughout the volume until its zero set bridges whatever holes may be present. Harary et al. [143] introduced a context-based completion algorithm to synthesize missing geometry for a given triangle mesh that has holes. Rock et al. [142] recovered a complete 3D model using an exemplar-based approach, which retrieves similar objects in a database of 3D models using view-based matching and transfers the symmetries and surfaces from retrieved models. Firman et al. [134] hypothesized that objects of dissimilar semantic classes often share similar 3D shape components, and estimate the hidden geometry for a wide range of objects using a limited dataset. ShapeNet [144] was the first work to apply deep learning to learn a 3D representation on a large scale CAD model database and with capability for shape completion. Following the success of Shapenet, various works have emerged that complete 3D shape using data-driven methods [145–151]. VConv-DAE [145] proposes a fully convolutional volumetric auto encoder to learn a volumetric representation from noisy data by estimating voxel occupancy grids. OctNetFusion [147] presents a 3D CNN architecture to predict an implicit surface representation; it outperforms the traditional volumetric fusion approach in terms of noise reduction and outlier suppression. Dai et al. [148] completed partial 3D shapes through a combination of volumetric deep neural networks and 3D shape synthesis. X-Section [149] predicts the end-point of an object along a ray which can be used with volumetric SDF fusion to obtain completed shapes. RevealNet [150] enables a semantically meaningful decomposition of a scanned scene into individual, complete, 3D objects. GAN style approaches are also widely used in object completion. For instance, 3D GAN [146] can generate high-quality 3D objects from a probabilistic space. The recently proposed ShapeInversion [151] introduces GAN inversion to shape completion, and gives robust results for real-world scans and partial inputs of various forms and incompleteness levels.

5.3.2 Scene completion

Scene completion often uses prior information, such as scene structural priors [152–155] and semantic priors [11, 17, 156, 157]. Silberman et al. [152] proposed a method for scene completion that can infer the layout of a complete room and the full extent of partially occluded objects. Sung et al. [153] used a collection of example 3D shapes to build structural part-based priors for shape completion. Song et al. [154] output semantic labels for all voxels in the camera view frustum with a single depth image as input. Dzitsiuk et al. [155] used plane priors to complete 3D reconstructions. ScanComplete [156] applies 3D CNNs in a hierarchical fashion to take an incomplete 3D scene scan as input and predict a complete 3D model along with per-voxel semantic labels. SISNet [17] reconstructs a complete 3D scene with precise voxel-wise semantics and presents a novel scene–instance–scene network, which takes advantages of both instance and scene level semantic information. Recent work, PALNet [157] utilizes a two-stream network to extract both 2D and 3D

Fig. 13 Comparison of object completion and scene completion. Left: initial reconstruction [134]. Right: completed geometry with sharp surface textures [135]. Reproduced with permission from Ref. [134], © IEEE 2016; Ref. [135], © ACM 2017.

features from multiple stages using fine-grained depth information to capture the context in the scene. Following the proliferation of large-scale 3D datasets, SSC has gained significant momentum in the research community because it holds unresolved challenges in recent years.

6 Performance evaluation

For a quantitative evaluation of indoor scene reconstruction systems, there are two widely used indicators: camera tracking accuracy and surface reconstruction accuracy. In this section, we quantitatively compare ten state-of-the-art reconstruction systems: DVO SLAM, RGBD SLAM [158], Kintinuous, VoxelHashing, SUN3D SfM [34], ElasticFusion, InfiniTAM v3, Redwood, BundleFusion, and UncertaintyAware [28].

DVO SLAM and RGBD SLAM apply pose graph optimization to achieve a globally consistent trajectory and then the global scene model is constructed by integrating all depth images in a volumetric representation. Kintinuous and ElasticFusion achieve a globally consistent model in a map-centric manner by deforming the global model according to global or local constraints. VoxelHashing and InfiniTAM v3 use a spatial hashing scheme to compress space, and can quickly realize surface reconstruction. SUN3D SfM takes a data-driven brute-force approach to RGB-D structure from motion (SfM), and can reconstruct big scenes with object labels. Redwood, BundleFusion, and UncertaintyAware divide the global model into submaps and obtain a globally consistent model by optimizing between submaps. To align submaps globally, Redwood uses dense geometric correspondences, while BundleFusion uses sparse as well as dense correspondences. UncertaintyAware exploits sparse features to align submaps.

6.1 Camera tracking accuracy

The accuracy of camera tracking is evaluated by comparing the estimated trajectory with the ground-truth. Two prominent error measures are the absolute trajectory error (ATE) and the relative pose error (RPE). The ATE directly measures the difference between points of the true and the estimated trajectory, and is well-suited for measuring the performance of visual SLAM systems. The metric most commonly used for quantitative evaluation is the root mean square error (RMSE). For evaluating the accuracy of camera tracking, there are two commonly used benchmarks: the ICL-NUIM synthetic benchmark [35] and the TUM benchmark [32].

Table 2 presents camera tracking accuracy (ATE RMSE) for the living rooms kr0–kr3 from the ICL-NUIM synthetic benchmark for the chosen state-of-the-art reconstruction systems; figures are quoted from the corresponding papers.

We also compared these systems using four common sequences from the TUM RGB-D benchmark: fr1_desk, fr2_xyz, fr3_office, and fr3_nst. Real-world scenes were scanned by a robot using Microsoft Kinect for Windows. The data were recorded at full frame rate (30 Hz) with a sensor resolution of (640×480). Table 3 shows the accuracy (ATE RMSE) of camera tracking on the TUM RGB-D benchmark. Note that the ground-truth (GT) trajectories are provided in the corresponding benchmarks; the results are quoted from corresponding papers. Speeds of those methods are estimated using the data provided in the corresponding papers. The computer configurations are also taken from the corresponding papers. It can be seen that BundleFusion and UncertaintyAware outperform other systems with respect to camera tracking. InfiniTAM v3 has the highest speed on the GPU, while DVO SLAM has the highest speed on the CPU.

6.2 Surface reconstruction accuracy

The accuracy of surface reconstruction is measured by comparing the reconstructions produced by the state-of-the-art methods against the ground-truth 3D surface model. There are five standard statistics computed over the distances for all vertices in the

Table 2 ATE RMSE (mm) on the ICL-NUIM benchmark [35]. Best results in bold

Method	kr0	kr1	kr2	kr3
DVO SLAM	104	29	191	152
RGBD SLAM	26	8	18	433
Kintinuous	72	5	10	355
VoxelHashing	14	**4**	18	120
ElasticFusion	9	9	14	106
Redwood	256	30	33	61
BundleFusion	6	**4**	6	11
UncertaintyAware	**5**	**4**	**5**	**10**

Table 3 Accuracy of estimated camera trajectories (ATE RMSE in mm) and mean speed (fps) of data fusion on the TUM RGB-D benchmark [32]. Best results in bold

Method	Camera trajectories (RMSE)					Mean speed (fps)	
	fr1_desk	fr2_xyz	fr3_office	fr3_nst	Average	GPU	CPU
DVO SLAM	21	18	35	18	23	—	**30 Hz**
RGBD SLAM	23	8	32	17	20	—	2.8 Hz
Kintinuous	37	29	30	31	32	15 Hz (GTX 560Ti)	—
VoxelHashing	23	22	23	87	39	46 Hz (GTX Titan)	—
Redwood	27	91	30	1929	519	—	Offline
ElasticFusion	20	11	17	16	12	32 Hz (GTX 780Ti)	—
InfiniTAM v3	18	21	22	20	203	**910 Hz** (GTX Titan)	—
BundleFusion	16	11	22	**12**	15	36 Hz (GTX Titan)	—
UncertaintyAware	**15**	**6**	**9**	14	**11**	43 Hz (GTX 1080Ti)	

reconstruction: mean, median, standard deviation, min, and max. The commonly used quantitative metrics for evaluating the performance of surface reconstruction are the living room sequences (kr0–kr3) of the synthetic ICL-NUIM benchmark [35]. Figure 14 shows the interior of a synthetic living room scene without color information. Each sequence partly covers the room and the average trajectory length is 7 m. Later, Choi et al. [3] augmented the original ICL-NUIM dataset in a number of ways to adapt it for evaluation of complete scene reconstruction pipelines, giving the Aug ICL-NUIM benchmark. The average trajectory length of each sequence is 36 m and the average surface area coverage reaches 88%.

We have compared the state-of-the-art methods both on the ICL-NUIM and Aug ICL-NUIM benchmarks. Table 4 gives surface reconstruction error on the ICL-NUIM benchmark (median distance in mm), while Table 5 gives the surface reconstruction error on the Aug ICL-NUIM benchmark (median distance in mm). The ground-truth 3D surface models are provided in the corresponding benchmarks,

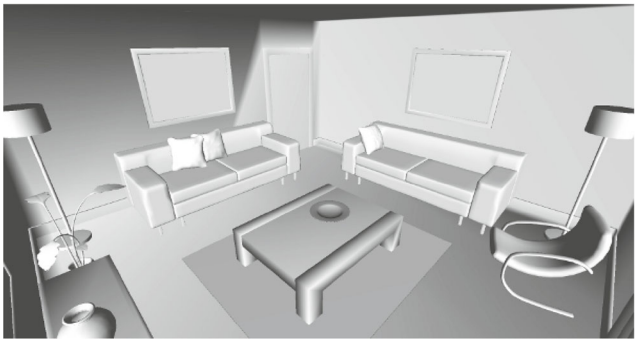


Fig. 14 Interior of a synthetic living room scene without color information. Reproduced with permission from Ref. [35], © IEEE 2014.

Table 4 Surface reconstruction error (mm) on the ICL-NUIM benchmark [35]. Best results in bold

Method	kr0	kr1	kr2	kr3
DVO SLAM	32	61	119	53
RGBD SLAM	44	32	31	167
Kintinuous	11	8	9	150
VoxelHashing	14	**4**	18	120
ElasticFusion	7	7	8	28
Redwood	20	20	13	22
BundleFusion	5	6	7	8
UncertaintyAware	**4**	**5**	**4**	**6**

Table 5 Surface reconstruction accuracy (median distance in mm) on the Aug ICL-NUIM dataset [3]. Best results in bold

Method	LR1	LR2	Off1	Off2	Ave.
DVO SLAM	160	50	80	70	90
Kintinuous	170	100	90	90	113
SUN3D SfM	80	60	110	60	78
Redwood	**30**	**50**	**20**	**30**	**33**
GT trajectory	30	20	10	20	20

and the results are quoted from corresponding papers. It can be seen that UncertaintyAware has the best reconstruction performance on the ICL-NUIM benchmark. The reconstruction accuracies of Redwood are closest to the GT trajectory on the Aug ICL-NUIM benchmark, benefiting from offline optimization.

6.3 Evaluation of pre-processing and post-processing

For high-quality 3D reconstruction, there are two important components in addition to the core 3D reconstruction pipeline: pre-processing and post-processing. The former focuses on handling noise or missing data in RGB-D images, while the latter

focuses on handling noise or missing data in 3D models. To evaluate the performance of depth enhancement and depth completion, experiments commonly use the NYU v2 dataset [75] by downsampling, adding noises, or making holes in the depth image. Furthermore, a quantitative comparison (e.g., RMSE) on the ToFMark dataset [159] can also be used to benchmark depth super-resolution methods. To validate the performance of surface optimization and surface completion, reconstructed models are qualitatively compared through visual observation or perceptual evaluation. Quantitative evaluation is suitable for comparisons on synthetic scenes (e.g., the ICL-NUIM dataset), but is challenging on real-world scenes as there is typically no ground-truth surface model. In particular, the geometrical intersection over union (IoU) and mean intersection over union (mIoU) may be evaluated on input occluded and observed surfaces on the SSC datasets (e.g., synthetic SUNCG-RGBD [160]). The above benchmarks have not been commonly used to evaluate state-of-the-art 3D reconstruction systems and need to be further standardized.

7 Summary and discussions

In this section, we discuss the key techniques and limitations in high-quality scene reconstruction with RGB-D cameras, and summarize application scenarios, challenges, and future directions.

7.1 Key techniques and limitations

Based on the pipeline of 3D scene reconstruction, the key issues are how to reduce errors in camera pose estimation and improve the accuracy of surface reconstruction. During the past decades, most successful RGB-D based reconstruction systems mainly focus on camera localization methods with various features and volume data fusion methods with elastic registration or local–global registration. Introducing deep learning into 3D reconstruction is a direction being explored, but it is hard to make substantial progress in a short time. Efficient methods for depth image processing and 3D model processing can improve the quality of 3D reconstruction with consumer RGB-D cameras. Currently data-driven approaches have obvious advantages in depth completion and surface completion. However they usually need a large amount of RGB-D scene data to support model training, and robust performance is often limited to specific scenarios.

7.2 Applications

As can be seen from the performance comparison in Section 6, the average error of state-of-the-art online and offline systems is just a few millimetres for the ICL-NUIM benchmark. Online scene reconstruction systems (e.g., InfiniTAM v3) with low requirements on computational performance can be applied in mobile devices. For instance, there are some apps (e.g., Polycam) available for mobile phones and tablets. Offline scene reconstruction systems (e.g., Redwood) are usually used for high-quality 3D map creation and digital cultural heritage protection. Real scene models built offline can be used in smart venues and virtual tours. Scene models with semantic information have potential applications in intelligent systems like autonomous robot navigation, HCI, and so on. High-quality dynamic 3D reconstruction can further be used in human action capture for human action analysis applications (e.g., sports performance analysis).

7.3 Challenges and future work

High-quality 3D scene reconstruction is computationally expensive, and the major challenge is how to quickly obtain realistic scene models with convenient devices. In addition to increasing accuracy and efficiency, future work can address the following: (i) task-oriented 3D scene understanding is a key research topic in 3D vision, and different 3D scenes should be reconstructed for different task-oriented purposes, and (ii) quality of reconstruction depends not only on reconstructing the geometry and appearance of the scene, but also exploring invisible information (e.g., purpose and utility) underpinning the scene.

8 Conclusions

The area of high-quality 3D reconstruction with RGB-D cameras has grown from various methods, which can be divided into three phases: image processing, camera pose estimation, and surface reconstruction. Our survey provides insight into this wide array of methods, highlighting strengths and limitations of current approaches. We find the research trends of state-of-the-art methods mainly concentrate on:

(i) combining multiple methods, e.g., BundleFusion and 3DLite, (ii) more use of CNNs and deep learning, e.g., for scan completion and semantic fusion, and (iii) using more information, e.g., object shape priors and scene structural priors.

To inspire researcher to propose new methods, we also suggest directions for future work in high-quality 3D reconstruction. Future directions may move: (i) from static to dynamic, e.g., real-time dynamic fusion, (ii) from local to global, e.g., local-to-global optimization, large-scale scene completion, (iii) from 2D to 3D processing, e.g., occlusion recovery, (iv) from single goal to multiple goals, e.g., scene reconstruction with semantics, geometric reconstruction with color texture, and (v) from low-level to high-level, e.g., 3D reconstruction with scene understanding.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant No. 2018YFC2000600, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 202100009, the National Natural Science Foundation of China under Grant No. 72071018, and the Fundamental Research Funds for Central Universities under Grant No. 2021TD006.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Orts-Escolano, S.; Rhemann, C.; Fanello, S.; Chang, W.; Kowdle, A.; Degtyarev, Y.; Kim, D.; Davidson, P. L.; Khamis, S.; Dou, M.; et al. Holoportation: Virtual 3D teleportation in real-time. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology, 741–754, 2016.

[2] DGene. Available at https://www.dgene.com/tech/model.

[3] Choi, S.; Zhou, Q. Y.; Koltun, V. Robust reconstruction of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5556–5565, 2015.

[4] Newcombe, R. A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A. J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. KinectFusion: Real-time dense surface mapping and tracking. In: Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality, 127–136, 2011.

[5] Curless, B.; Levoy, M. A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 303–312, 1996.

[6] Rusinkiewicz, S.; Levoy, M. Efficient variants of the ICP algorithm. In: Proceedings of the 3rd International Conference on 3-D Digital Imaging and Modeling, 145–152, 2001.

[7] Whelan, T.; Kaess, M.; Fallon, M.; Johannsson, H.; Leonard, J.; Mcdonald, J. Kintinuous: Spatially extended KinectFusion. *Robotics and Autonomous Systems* Vol. 69, No. C, 3–14, 2012.

[8] Whelan, T.; Kaess, M.; Johannsson, H.; Fallon, M.; Leonard, J. J.; McDonald, J. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research* Vol. 34, Nos. 4–5, 598–626, 2015.

[9] Thomas, D.; Sugimoto, A. Modeling large-scale indoor scenes with rigid fragments using RGB-D cameras. *Computer Vision and Image Understanding* Vol. 157, 103–116, 2017.

[10] Golodetz, S.; Cavallari, T.; Lord, N. A.; Prisacariu, V. A.; Murray, D. W.; Torr, P. H. S. Collaborative large-scale dense 3D reconstruction with online inter-agent pose optimisation. *IEEE Transactions on Visualization and Computer Graphics* Vol. 24, No. 11, 2895–2905, 2018.

[11] Dai, A.; Diller, C.; Niessner, M. SG-NN: Sparse generative neural networks for self-supervised scene completion of RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 846–855, 2020.

[12] Salas-Moreno, R. F.; Newcombe, R. A.; Strasdat, H.; Kelly, P. H. J.; Davison, A. J. SLAM++: Simultaneous localisation and mapping at the level of objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1352–1359, 2013.

[13] Shao, T. J.; Xu, W. W.; Zhou, K.; Wang, J. D.; Li, D. P.; Guo, B. N. An interactive approach to semantic modeling of indoor scenes with an RGBD camera. *ACM Transactions on Graphics* Vol. 31, No. 6, Article No. 136, 2012.

[14] Chen, K.; Lai, Y. K.; Wu, Y. X.; Martin, R.; Hu, S. M. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 208, 2014.

[15] McCormac, J.; Handa, A.; Davison, A.; Leutenegger, S. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In: Proceedings of the IEEE International Conference on Robotics and Automation, 4628–4635, 2017.

[16] Hou, J.; Dai, A.; Nießner, M. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4416–4425, 2019.

[17] Cai, Y. J.; Chen, X. S.; Zhang, C.; Lin, K. Y.; Wang, X. G.; Li, H. S. Semantic scene completion via integrating instances and scene in-the-loop. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 324–333, 2021.

[18] Newcombe, R. A.; Fox, D.; Seitz, S. M. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 343–352, 2015.

[19] Dou, M. S.; Khamis, S.; Degtyarev, Y.; Davidson, P.; Fanello, S. R.; Kowdle, A.; Escolano, S. O.; Rhemann, C.; Kim, D.; Taylor, J.; et al. Fusion4D: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics* Vol. 35, No. 4, Article No. 114, 2016.

[20] Meerits, S.; Thomas, D.; Nozick, V.; Saito, H. FusionMLS: Highly dynamic 3D reconstruction with consumer-grade RGB-D cameras. *Computational Visual Media* Vol. 4, No. 4, 287–303, 2018.

[21] Saito, S.; Simon, T.; Saragih, J.; Joo, H. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 81–90, 2020.

[22] Nießner, M.; Zollhöfer, M.; Izadi, S.; Stamminger, M. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics* Vol. 32, No. 6, Article No. 169, 2013.

[23] Steinbrücker, F.; Sturm, J.; Cremers, D. Volumetric 3D mapping in real-time on a CPU. In: Proceedings of the IEEE International Conference on Robotics and Automation, 2021–2028, 2014.

[24] Kähler, O.; Adrian Prisacariu, V.; Yuheng Ren, C.; Sun, X.; Torr, P.; Murray, D. Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Transactions on Visualization and Computer Graphics* Vol. 21, No. 11, 1241–1250, 2015.

[25] Prisacariu, V. A.; Kähler, O.; Golodetz, S.; Sapienza, M.; Cavallari, T.; Torr, P. H.; Murray, D. W. InfiniTAM v3: A framework for large-scale 3D reconstruction with loop closure. *arXiv preprint* arXiv:1708.00783, 2017.

[26] Dai, A.; Nießner, M.; Zollhöfer, M.; Izadi, S.; Theobalt, C. BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 76a, 2017.

[27] Maier, R.; Kim, K.; Cremers, D.; Kautz, J.; Nießner, M. Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In: Proceedings of the IEEE International Conference on Computer Vision, 3133–3141, 2017.

[28] Cao, Y. P.; Kobbelt, L.; Hu, S. M. Real-time high-accuracy three-dimensional reconstruction with consumer RGB-D cameras. *ACM Transactions on Graphics* Vol. 37, No. 5, Article No. 171, 2018.

[29] Whelan, T.; Salas-Moreno, R. F.; Glocker, B.; Davison, A. J.; Leutenegger, S. ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research* Vol. 35, No. 14, 1697–1716, 2016.

[30] Jeon, J.; Jung, Y.; Kim, H.; Lee, S. Texture map generation for 3D reconstructed scenes. *The Visual Computer* Vol. 32, Nos. 6–8, 955–965, 2016.

[31] Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2432–2443, 2017.

[32] Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 573–580, 2012.

[33] Zhou, Q. Y.; Miller, S.; Koltun, V. Elastic fragments for dense scene reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, 473–480, 2013.

[34] Xiao, J. X.; Owens, A.; Torralba, A. SUN3D: A database of big spaces reconstructed using SfM and object labels. In: Proceedings of the IEEE International Conference on Computer Vision, 1625–1632, 2013.

[35] Handa, A.; Whelan, T.; McDonald, J.; Davison, A. J. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1524–1531, 2014.

[36] Hua, B. S.; Pham, Q. H.; Nguyen, D. T.; Tran, M. K.; Yu, L. F.; Yeung, S. K. SceneNN: A scene meshes dataset with aNNotations. In: Proceedings of the 4th International Conference on 3D Vision, 92–101, 2016.

[37] Wasenmüller, O.; Meyer, M.; Stricker, D. CoRBS: Comprehensive RGB-D benchmark for SLAM using Kinect v2. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1–7, 2016.

[38] Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niebner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3D: Learning from RGB-D data in indoor environments. In: Proceedings of the International Conference on 3D Vision, 667–676, 2017.

[39] McCormac, J.; Handa, A.; Leutenegger, S.; Davison, A. J. SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation? In: Proceedings of the IEEE International Conference on Computer Vision, 2697–2706, 2017.

[40] Palazzolo, E.; Behley, J.; Lottes, P.; Giguère, P.; Stachniss, C. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 7855–7862, 2019.

[41] Li, W. B.; Saeedi, S.; McCormac, J.; Clark, R.; Leutenegger, S. InteriorNet: Mega-scale Multi-sensor Photo-realistic indoor scenes dataset. *arXiv preprint* arXiv:1809.00716, 2018.

[42] Straub, J.; Whelan, T.; Ma, L. N.; Chen, Y. F.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint* arXiv:1906.05797, 2019.

[43] Shi, X. S.; Li, D. J.; Zhao, P. P.; Tian, Q. B.; Tian, Y. X.; Long, Q. W.; Zhu, C.; Song, J.; Qiao, F.; Song, L.; et al. Are we ready for service robots? The OpenLORIS-scene datasets for lifelong SLAM. In: Proceedings of the IEEE International Conference on Robotics and Automation, 3139–3145, 2020.

[44] Kadambi, A.; Taamazyan, V.; Shi, B. X.; Raskar, R. Polarized 3D: High-quality depth sensing with polarization cues. In: Proceedings of the IEEE International Conference on Computer Vision, 3370–3378, 2015.

[45] Berger, M.; Tagliasacchi, A.; Seversky, L.; Alliez, P.; Levine, J.; Sharf, A.; Silva, C. State of the art in surface reconstruction from point clouds. In: Proceedings of the Eurographics 2014 - State of the Art Reports, 161–185, 2014.

[46] Chen, K.; Lai, Y. K.; Hu, S. M. 3D indoor scene modeling from RGB-D data: A survey. *Computational Visual Media* Vol. 1, No. 4, 267–278, 2015.

[47] Stotko, P. State of the art in real-time registration of RGB-D images. In: Proceedings of the Central European Seminar on Computer Graphics for Students, 2016.

[48] Xu, K.; Kim, V. G.; Huang, Q. X.; Mitra, N.; Kalogerakis, E. Data-driven shape analysis and processing. In: Proceedings of the SIGGRAPH ASIA 2016 Courses, Article No. 4, 2016.

[49] Zollhöfer, M.; Stotko, P.; Görlitz, A.; Theobalt, C.; Nießner, M.; Klein, R.; Kolb, A. State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum* Vol. 37, No. 2, 625–652, 2018.

[50] Han, X. F.; Laga, H.; Bennamoun, M. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 5, 1578–1604, 2021.

[51] Roldão, L.; Charette, R. D.; Verroust-Blondet, A. 3D semantic scene completion: A survey. *arXiv preprint* arXiv:2103.07466, 2021.

[52] Liu, Y. Z.; Fu, Y. J.; Chen, F. D.; Goossens, B.; Zhao, H. Simultaneous localization and mapping related datasets: A comprehensive survey. *arXiv preprint* arXiv:2102.04036, 2021.

[53] Nguyen, C. V.; Izadi, S.; Lovell, D. Modeling kinect sensor noise for improved 3D reconstruction and tracking. In: Proceedings of the 2nd International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, 524–530, 2012.

[54] Sarbolandi, H.; Lefloch, D.; Kolb, A. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Computer Vision and Image Understanding* Vol. 139, 1–20, 2015.

[55] Wasenmüller, O.; Stricker, D. Comparison of kinect V1 and V2 depth images in terms of accuracy and precision. In: *Computer Vision – ACCV 2016 Workshops. Lecture Notes in Computer Science, Vol. 10117*. Chen, C. S.; Lu, J.; Ma, K, K. Eds. Springer Cham, 34–45, 2017.

[56] Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In: Proceedings of the 6th International Conference on Computer Vision, 839–846, 1998.

[57] Li, J. W.; Gao, W.; Wu, Y. H. Elaborate scene reconstruction with a consumer depth camera. *International Journal of Automation and Computing* Vol. 15, No. 4, 443–453, 2018.

[58] Sterzentsenko, V.; Saroglou, L.; Chatzitofis, A.; Thermos, S.; Zioulis, N.; Doumanoglou, A.; Zarpalas, D.; Daras, P. Self-supervised deep depth denoising.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 1242–1251, 2019.

[59] Ferstl, D.; Rüther, M.; Bischof, H. Variational depth superresolution using example-based edge representations. In: Proceedings of the IEEE International Conference on Computer Vision, 513–521, 2015.

[60] Kopf, J.; Cohen, M. F.; Lischinski, D.; Uyttendaele, M. Joint bilateral upsampling. *ACM Transactions on Graphics* Vol. 26, No. 3, 96–es, 2007.

[61] Kiechle, M.; Hawe, S.; Kleinsteuber, M. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, 1545–1552, 2013.

[62] Park, J.; Kim, H.; Tai, Y. W.; Brown, M. S.; Kweon, I. S. High-quality depth map upsampling and completion for RGB-D cameras. *IEEE Transactions on Image Processing* Vol. 23, No. 12, 5559–5572, 2014.

[63] Hui, T. W.; Loy, C. C.; Tang, X. O. Depth map super-resolution by deep multi-scale guidance. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9907*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 353–369, 2016.

[64] Riegler, G.; Ferstl, D.; Rüther, M.; Bischof, H. A deep primal-dual network for guided depth super-resolution. In: Procedings of the British Machine Vision Conference, 2016.

[65] Riegler, G.; Rüther, M.; Bischof, H. ATGV-net: Accurate depth super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9907*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 268–284, 2016.

[66] Zhang, R.; Tsai, P. S.; Cryer, J. E.; Shah, M. Shape-from-shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 21, No. 8, 690–706, 1999.

[67] Han, Y.; Lee, J.-Y.; Kweon, I. S. High quality shape from a single RGB-D image under uncalibrated natural illumination. In: Proceedings of the IEEE International Conference on Computer Vision, 1617–1624, 2013.

[68] Yu, L. F.; Yeung, S. K.; Tai, Y. W.; Lin, S. Shading-based shape refinement of RGB-D images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1415–1422, 2013.

[69] Wu, C. L.; Zollhöfer, M.; Nießner, M.; Stamminger, M.; Izadi, S.; Theobalt, C. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 200, 2014.

[70] Or-El, R.; Rosman, G.; Wetzler, A.; Kimmel, R.; Bruckstein, A. M. RGBD-fusion: Real-time high precision depth recovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5407–5416, 2015.

[71] Cui, Z. P.; Gu, J. W.; Shi, B. X.; Tan, P.; Kautz, J. Polarimetric multi-view stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 369–378, 2017.

[72] Ba, Y.; Gilbert, A. R.; Wang, F.; Yang, J.; Chen, R.; Wang, Y.; Yan, L.; Shi, B.; Kadambi, A. Deep shape from polarization. *arXiv preprint* arXiv:1903.10210, 2019.

[73] Deschaintre, V.; Lin, Y. M.; Ghosh, A. Deep polarization imaging for 3D shape and SVBRDF acquisition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15562–15571, 2021.

[74] Information on https://github.com/yindaz/DeepCompletionRelease.

[75] Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7576*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 746–760, 2012.

[76] Chen, Q. F.; Koltun, V. Fast MRF optimization with application to depth reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3914–3921, 2014.

[77] Ma, F. C.; Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: Proceedings of the IEEE International Conference on Robotics and Automation, 4796–4803, 2018.

[78] Chen, Z.; Badrinarayanan, V.; Drozdov, G.; Rabinovich, A. Estimating depth from RGB and sparse sensing. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11208*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 176–192, 2018.

[79] Cheng, X. J.; Wang, P.; Yang, R. G. Depth estimation via affinity learned with convolutional spatial propagation network. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11220*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 108–125, 2018.

[80] Cheng, X. J.; Wang, P.; Yang, R. G. Learning depth with convolutional spatial propagation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 42, No. 10, 2361–2379, 2020.

[81] Lee, B. U.; Jeon, H. G.; Im, S.; Kweon, I. S. Depth completion with deep geometry and context guidance. In: Proceedings of the International Conference on Robotics and Automation, 3281–3287, 2019.

[82] Cheng, X. J.; Wang, P.; Guan, C. Y.; Yang, R. G. CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 10615–10622, 2020.

[83] Imran, S.; Long, Y. F.; Liu, X. M.; Morris, D. Depth coefficients for depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12438–12447, 2019.

[84] Zhu, L. Y.; Mousavian, A.; Xiang, Y.; Mazhar, H.; Eenbergen, J. V.; Debnath, S.; Fox, D. RGB-D local implicit function for depth completion of transparent objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4647–4656, 2021.

[85] Li, J. W.; Gao, W.; Wu, Y. H. High-quality 3D reconstruction with depth super-resolution and completion. *IEEE Access* Vol. 7, 19370–19381, 2019.

[86] Slavcheva, M.; Kehl, W.; Navab, N.; Ilic, S. SDF-2-SDF: Highly accurate 3D object reconstruction. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9905.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 680–696, 2016.

[87] Zeng, A.; Song, S. R.; Nießner, M.; Fisher, M.; Xiao, J. X.; Funkhouser, T. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 199–208, 2017.

[88] Lee, J. K.; Yea, J.; Park, M. G.; Yoon, K. J. Joint layout estimation and global multi-view registration for indoor reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, 162–171, 2017.

[89] Besl, P. J.; McKay, N. D. Method for registration of 3-D shapes. In: Proceedings of the SPIE 1611, Sensor Fusion IV: Control Paradigms and Data Structures, 586–606, 1992.

[90] Low, K.-L. Linear least-squares optimization for point-to-plane ICP surface registration. Technical Report TR04-004. Department of Computer Science, University of North Carolina at Chapel Hill, 2004.

[91] Kerl, C.; Sturm, J.; Cremers, D. Robust odometry estimation for RGB-D cameras. In: Proceedings of the IEEE International Conference on Robotics and Automation, 3748–3754, 2013.

[92] Whelan, T.; Johannsson, H.; Kaess, M.; Leonard, J. J.; McDonald, J. Robust real-time visual odometry for dense RGB-D mapping. In: Proceedings of the IEEE International Conference on Robotics and Automation, 5724–5731, 2013.

[93] Johnson, A. E.; Kang, S. B. Registration and integration of textured 3D data. *Image and Vision Computing* Vol. 17, No. 2, 135–147, 1999.

[94] Haehnel, D.; Thrun, S.; Burgard, W. An extension of the icp algorithm for modeling nonrigid objects with mobile robots. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, 915–920, 2003.

[95] Segal, A.; Haehnel, D.; Thrun, S. Generalized-ICP. *Robotics: Science and Systems* Vol. 2, No. 4, 435, 2009.

[96] Serafin, J.; Grisetti, G. NICP: Dense normal based point cloud registration. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 742–749, 2015.

[97] Wang, Y.; Solomon, J. Deep closest point: Learning representations for point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3522–3531, 2019.

[98] Wang, Y.; Solomon, J. M. PRNet: Self-supervised learning for partial-to-partial registration. *arXiv preprint* arXiv:1910.12240, 2019.

[99] Aoki, Y.; Goforth, H.; Srivatsan, R. A.; Lucey, S. PointNetLK: Robust & efficient point cloud registration using PointNet. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7156–7165, 2019.

[100] Ginzburg, D.; Raviv, D. Deep Weighted Consensus: Dense correspondence confidence maps for 3D shape registration. *arXiv preprint* arXiv:2105.02714, 2021.

[101] Mur-Artal, R.; Tardós, J. D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics* Vol. 33, No. 5, 1255–1262, 2017.

[102] Sarlin, P. E.; Unagar, A.; Larsson, M.; Germain, H.; Toft, C.; Larsson, V.; Pollefeys, M.; Lepetit, V.; Hammarstrand, L.; Kahl, F.; et al. Back to the feature: Learning robust camera localization from pixels to pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3246–3256, 2021.

[103] Choi, C.; Trevor, A. J. B.; Christensen, H. I. RGB-D edge detection and edge-based registration. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 1568–1575, 2013.

[104] Lu, Y.; Song, D. Z. Robust RGB-D odometry using point and line features. In: Proceedings of the IEEE International Conference on Computer Vision, 3934–3942, 2015.

[105] Zhou, Q. Y.; Koltun, V. Depth camera tracking with contour cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 632–638, 2015.

[106] Taguchi, Y.; Jian, Y. D.; Ramalingam, S.; Feng, C. Point-plane SLAM for hand-held 3D sensors. In: Proceedings of the IEEE International Conference on Robotics and Automation, 5182–5189, 2013.

[107] Salas-Moreno, R. F.; Glocken, B.; Kelly, P. H. J.; Davison, A. J. Dense planar SLAM. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 157–164, 2014.

[108] Ma, L. N.; Kerl, C.; Stückler, J.; Cremers, D. CPA-SLAM: Consistent plane-model alignment for direct RGB-D SLAM. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1285–1291, 2016.

[109] Shi, Y. F.; Xu, K.; Nießner, M.; Rusinkiewicz, S.; Funkhouser, T. PlaneMatch: Patch coplanarity prediction for robust RGB-D reconstruction. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11212*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 767–784, 2018.

[110] Yunus, R.; Li, Y. Y.; Tombari, F. ManhattanSLAM: Robust planar tracking and mapping leveraging mixture of Manhattan frames. In: Proceedings of the IEEE International Conference on Robotics and Automation, 6687–6693, 2021.

[111] Kehl, W.; Tombari, F.; Ilic, S.; Navab, N. Real-time 3D model tracking in color and depth on a single CPU core. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 465–473, 2017.

[112] Schönberger, J. L.; Pollefeys, M.; Geiger, A.; Sattler, T. Semantic visual localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6896–6906, 2018.

[113] Yin, Z. C.; Shi, J. P. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1983–1992, 2018.

[114] Tang, S.; Tang, C.; Huang, R.; Zhu, S.; Tan, P. Learning camera localization via dense scene matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1831–1841, 2021.

[115] Puri, P.; Jia, D. Y.; Kaess, M. GravityFusion: Real-time dense mapping without pose graph using deformation and orientation. In: Procedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 6506–6513, 2017.

[116] Dong, W.; Wang, Q. Y.; Wang, X.; Zha, H. B. PSDF fusion: Probabilistic signed distance function for on-the-fly 3D data fusion and scene reconstruction. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11213*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 714–730, 2018.

[117] Steinbrucker, F.; Kerl, C.; Cremers, D.; Sturm, J. Large-scale multi-resolution surface reconstruction from RGB-D sequences. In: Proceedings of the IEEE International Conference on Computer Vision, 3264–3271, 2013.

[118] Sumner, R. W.; Schmid, J.; Pauly, M. Embedded deformation for shape manipulation. *ACM Transactions on Graphics* Vol. 26, No. 3, 80–es, 2007.

[119] Zheng, Z. R.; Yu, T.; Wei, Y. X.; Dai, Q. H.; Liu, Y. B. DeepHuman: 3D human reconstruction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7738–7748, 2019.

[120] Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; Lovegrove, S. DeepSDF: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 165–174, 2019.

[121] Sitzmann, V.; Zollhöfer, M.; Wetzstein, G. Scene representation networks: Continuous 3D-structure-aware neural scene representations. *arXiv preprint* arXiv:1906.01618, 2019.

[122] Li, Z. Q.; Niklaus, S.; Snavely, N.; Wang, O. Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6494–6504, 2021.

[123] Pfister, H.; Zwicker, M.; van Baar, J.; Gross, M. Surfels: Surface elements as rendering primitives. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, 335–342, 2000.

[124] Andersen, V.; Aans, H.; Brentzen, J. A. Surfel based geometry resonstruction. In: *Theory and Practice of Computer Graphics*. The Eurographics Association, 39–44, 2010.

[125] Keller, M.; Lefloch, D.; Lambers, M.; Izadi, S.; Weyrich, T.; Kolb, A. Real-time 3D reconstruction

in dynamic scenes using point-based fusion. In: Proceedings of the International Conference on 3D Vision, 1–8, 2013.

[126] Mihajlovic, M.; Weder, S.; Pollefeys, M.; Oswald, M. R. DeepSurfels: learning online appearance fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14519–14530, 2021.

[127] Mandikal, P.; Radhakrishnan, V. B. Dense 3D point cloud reconstruction using a deep pyramid network. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1052–1060, 2019.

[128] Wolff, K.; Kim, C.; Zimmer, H.; Schroers, C.; Botsch, M.; Sorkine-Hornung, O.; Sorkine-Hornung, A. Point cloud noise and outlier removal for image-based 3D reconstruction. In: Proceedings of the 4th International Conference on 3D Vision, 118–127, 2016.

[129] Casajus, P. H.; Ritschel, T.; Ropinski, T. Total denoising: Unsupervised learning of 3D point cloud cleaning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 52–60, 2019.

[130] Delaunoy, A.; Prados, E. Gradient flows for optimizing triangular mesh-based surfaces: Applications to 3D reconstruction problems dealing with visibility. *International Journal of Computer Vision* Vol. 95, No. 2, 100–123, 2011.

[131] Wang, P. S.; Liu, Y.; Tong, X. Mesh denoising via cascaded normal regression. *ACM Transactions on Graphics* Vol. 35, No. 6, Article No. 232, 2016.

[132] Schertler, N.; Tarini, M.; Jakob, W.; Kazhdan, M.; Gumhold, S.; Panozzo, D. Field-aligned online surface reconstruction. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 77, 2017.

[133] Tsai, C. Y.; Sankaranarayanan, A. C.; Gkioulekas, I. Beyond volumetric albedo—A surface optimization framework for non-line-of-sight imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1545–1555, 2019.

[134] Firman, M.; Aodha, O. M.; Julier, S.; Brostow, G. J. Structured prediction of unobserved voxels from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5431–5440, 2016.

[135] Huang, J.; Dai, A.; Guibas, L. J.; Niessner, M. 3Dlite: Towards commodity 3D scanning for content creation. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 203, 2017.

[136] Zollhöfer, M., Dai, A., Innmann, M., Wu, C. L.; Stamminger, M., Theobalt, C.; Nießner, M. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics* Vol. 34, No. 4, Article No. 96, 2015.

[137] Xu, D.; Duan, Q.; Zheng, J. M.; Zhang, J. Y.; Cai, J. F.; Cham, T. J. Shading-based surface detail recovery under general unknown illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 2, 423–436, 2018.

[138] Chen, Z. Q.; Kim, V. G.; Fisher, M.; Aigerman, N.; Zhang, H.; Chaudhuri, S. DECOR-GAN: 3D shape detailization by conditional refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15735–15744, 2021.

[139] Liu, Z. N.; Cao, Y. P.; Kuang, Z. F.; Kobbelt, L.; Hu, S. M. High-quality textured 3D shape reconstruction with cascaded fully convolutional networks. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 1, 83–97, 2021.

[140] Huang, J. W.; Thies, J.; Dai, A.; Kundu, A.; Jiang, C. Y.; Guibas, L. J.; Nießner, M.; Funkhouser, T. Adversarial texture optimization from RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1556–1565, 2020.

[141] Davis, J.; Marschner, S. R.; Garr, M.; Levoy, M. Filling holes in complex surfaces using volumetric diffusion. In: Proceedings of the International Symposium on 3D Data Processing Visualization and Transmission, 428–441, 2002.

[142] Rock, J.; Gupta, T.; Thorsen, J.; Gwak, J.; Shin, D.; Hoiem, D. Completing 3D object shape from one depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2484–2493, 2015.

[143] Harary, G.; Tal, A.; Grinspun, E. Context-based coherent surface completion. *ACM Transactions on Graphics* Vol. 33, No. 1, Article No. 5, 2014.

[144] Wu, Z. R.; Song, S. R.; Khosla, A.; Yu, F.; Zhang, L. G.; Tang, X. O.; Xiao, J. 3D ShapeNets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1912–1920, 2015.

[145] Sharma, A.; Grau, O.; Fritz, M. VConv-DAE: Deep volumetric shape learning without object labels. In: *Computer Vision – ECCV 2016 Workshops. Lecture Notes in Computer Science, Vol. 9915*. Hua, G.; Jégou, H. Eds. Springer Cham, 236–250, 2016.

[146] Wu, J.; Zhang, C.; Xue, T.; Freeman, W. T.; Tenenbaum, J. B. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Proceedings of the 29th Conference on Neural Information Processing System, 82–90, 2016.

[147] Riegler, G.; Ulusoy, A. O.; Bischof, H.; Geiger, A. OctNetFusion: Learning depth fusion from data.

In: Proceedings of the International Conference on 3D Vision, 57–66, 2017.

[148] Dai, A.; Qi, C. R.; Nießner, M. Shape completion using 3D-encoder-predictor CNNs and shape synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6545–6554, 2017.

[149] Nicastro, A.; Clark, R.; Leutenegger, S. X-section: Cross-section prediction for enhanced RGB-D fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 1517–1526, 2019.

[150] Hou, J.; Dai, A.; Nießner, M. RevealNet: Seeing behind objects in RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2095–2104, 2020.

[151] Zhang, J. Z.; Chen, X. Y.; Cai, Z.; Pan, L.; Zhao, H. Y.; Yi, S.; Yeo, C. K., Dai, B.; Loy, C. C. Unsupervised 3D shape completion through GAN inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1768–1777, 2021.

[152] Silberman, N.; Shapira, L.; Gal, R.; Kohli, P. A contour completion model for augmenting surface reconstructions. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8691*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 488–503, 2014.

[153] Sung, M.; Kim, V. G.; Angst, R.; Guibas, L. Data-driven structural priors for shape completion. *ACM Transactions on Graphics* Vol. 34, No. 6, Article No. 175, 2015.

[154] Song, S. R.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; Funkhouser, T. Semantic scene completion from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 190–198, 2017.

[155] Dzitsiuk, M.; Sturm, J.; Maier, R.; Ma, L. N.; Cremers, D. De-noising, stabilizing and completing 3D reconstructions on-the-go using plane priors. In: Proceedings of the IEEE International Conference on Robotics and Automation, 3976–3983, 2017.

[156] Dai, A.; Ritchie, D.; Bokeloh, M.; Reed, S.; Sturm, J.; Nießner, M. ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4578–4587, 2018.

[157] Li, J.; Liu, Y.; Yuan, X.; Zhao, C. X.; Siegwart, R.; Reid, I.; Cadena, C. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics and Automation Letters* Vol. 5, No. 1, 219–226, 2020.

[158] Endres, F.; Hess, J.; Engelhard, N.; Sturm, J.; Cremers, D.; Burgard, W. An evaluation of the RGB-D SLAM system. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1691–1696, 2012.

[159] Ferstl, D.; Reinbacher, C.; Ranftl, R.; Ruether, M.; Bischof, H. Image guided depth upsampling using anisotropic total generalized variation. In: Proceedings of the IEEE International Conference on Computer Vision, 993–1000, 2013.

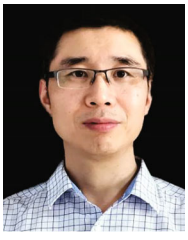
[160] Zhang, Y. D.; Song, S. R.; Yumer, E.; Savva, M.; Lee, J. Y.; Jin, H. L.; Funkhouser, T. Physically-based rendering for indoor scene understanding using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5057–5065, 2017.

Jianwei Li received her B.Sc. degree in measurement and control technology and instruments, and her M.Sc. degree in detection technology and automatic equipment from Beijing Jiaotong University, China, in 2008 and 2011 respectively, and her Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2019. She is currently a lecturer in Beijing Sports University. Her research interests include intelligent sensing technology for motion capture, computer vision for human action analysis, 3D reconstruction from images, and SLAM technology.

Wei Gao received his B.Sc. degree in computational mathematics and his M.Sc. degree in pattern recognition and intelligent systems from Shanxi University and his Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2002, 2005, and 2008, respectively. In July 2008, he joined the Robot Vision Group of the National Laboratory of Pattern Recognition where he is currently an associate professor. His research interests include 3D reconstruction from images and SLAM technology.

Yihong Wu received her Ph.D. degree from the Institute of Systems Science of the Chinese Academy of Sciences in 2001. She is currently a professor at the Institute of Automation of the Chinese Academy of Sciences. Her research interests include vision geometry, image matching, camera calibration, camera pose determination, SLAM, and their applications.

Yangdong Liu received his B.S. degree from the Northern University of China and his M.S. degree from Beijing University of Posts and Telecommunications, in 2012 and 2015 respectively, and his Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2019. Currently, he is an engineer at Huawei Technologies Co., Ltd. His research focuses on computer vision, 3D reconstruction, and AR/VR.



Yanfei Shen received his B.S. and M.S. degrees in computer science from the Key Laboratory of Multimedia and Network Communication, Wuhan University, China, in 1999 and 2002 respectively, and his Ph.D. degree in computer applications from the Chinese Academy of Sciences in 2014. He has served as an associate professor in the Institute of Computing Technology, Chinese Academy of Sciences and Beijing University of Posts and Telecommunications. He is currently a professor in Beijing Sports University. His research interests include intelligent sensing technology for motion capture, computer vision for human action recognition and analysis, and motion performance analysis for team sports.