International Journal of Geo-Information

MDPI

Review

A Review of Techniques for 3D Reconstruction of Indoor Environments

Zhizhong Kang [1,2,*], Juntao Yang [1,2], Zhou Yang [1,2] and Sai Cheng [1,2]

1 School of Land Science and Technology, China University of Geosciences, No. 29 Xueyuan Road, Haidian District, Beijing 100083, China; jtyang@cugb.edu.cn (J.Y.); zyang0301@cugb.edu.cn (Z.Y.); chengsai@cugb.edu.cn (S.C.)

2 Shanxi Key Laboratory of Resources, Environment and Disaster Monitoring, No. 380 Yingbin West Street, Yuci District, Jinzhong 030600, China

* Correspondence: zzkang@cugb.edu.cn

Received: 26 January 2020; Accepted: 15 May 2020; Published: 19 May 2020

Abstract: Indoor environment model reconstruction has emerged as a significant and challenging task in terms of the provision of a semantically rich and geometrically accurate indoor model. Recently, there has been an increasing amount of research related to indoor environment reconstruction. Therefore, this paper reviews the state-of-the-art techniques for the three-dimensional (3D) reconstruction of indoor environments. First, some of the available benchmark datasets for 3D reconstruction of indoor environments are described and discussed. Then, data collection of 3D indoor spaces is briefly summarized. Furthermore, an overview of the geometric, semantic, and topological reconstruction of the indoor environment is presented, where the existing methodologies, advantages, and disadvantages of these three reconstruction types are analyzed and summarized. Finally, future research directions, including technique challenges and trends, are discussed for the purpose of promoting future research interest. It can be concluded that most of the existing indoor environment reconstruction methods are based on the strong Manhattan assumption, which may not be true in a real indoor environment, hence limiting the effectiveness and robustness of existing indoor environment reconstruction methods. Moreover, based on the hierarchical pyramid structures and the learnable parameters of deep-learning architectures, multi-task collaborative schemes to share parameters and to jointly optimize each other using redundant and complementary information from different perspectives show their potential for the 3D reconstruction of indoor environments. Furthermore, indoor–outdoor space seamless integration to achieve a full representation of both interior and exterior buildings is also heavily in demand.

Keywords: indoor environment; geometric modeling; semantic modeling; topological modeling; scene reconstruction

1. Introduction

According to an investigation by the Environmental Protection Agency, more than 75% of the population throughout the world live in towns and cities and spend almost 90% of their time inside buildings [1]. Human beings usually perform many indoor activities related to work, shopping, leisure, dining, sport, and so on. To facilitate these human activities, an indispensable factor is the availability of indoor spatial information representation for satisfying the requirement of different applications. Moreover, with the rise of urban populations and the prevalence of large-scale buildings (e.g., airports, train stations, shopping malls, and hospitals) in current society, there is also a growing demand for up-to-date spatial layouts of indoor environments [2] and information regarding the objects contained within them [3].

In recent years, three-dimensional (3D) reconstruction of indoor environments has emerged as a significant and challenging task [4] in terms of the provision of a semantically rich and geometrically accurate indoor model. As a matter of fact, it is also fundamental for many applications, such as navigation guidance [5], emergency management [6], building maintenance and renovation planning [7], and a range of indoor location-based services (e.g., way-finding, contextualized content delivery) [8]. For example, people without any interior experience can browse online furniture departments, which offer 3D digital models of their products, and decide what to buy with the help of experts or digital expert systems [9]. Furthermore, semantically rich and geometrically accurate indoor models provide critical information (such as door locations used for exits, opening directions of doors, locations of indoor spaces and their topological relationships, and semantic attributes of indoor spaces) for indoor navigation services [5].

The generation of an indoor space model first needs proper implementation of sensors (e.g., camera (monocular [10], stereo [11], video [12], or panoramic [13], laser scanning [14], depth camera [15], etc.) to collect indoor scene data, which help to accurately model the whole indoor scene and which make contributions to the efficiency of the reconstructed model in the subsequent procedures. For example, an inexpensive depth sensor, such as Microsoft Kinect, can be used to quickly digitize and reconstruct an indoor 3D model using streams of depth and color images. Moreover, terrestrial or mobile laser-scanning systems are significantly advanced because they can easily and rapidly capture the detailed geometry information of indoor environments. Following this, a proper algorithm is required to reconstruct an indoor space model from the incoming data. The modeling of an indoor environment can be represented, ranging from mesh models [16] and computer-aided design (CAD) geometrical models [17] to object-based parametric models [18,19]. These models are generally reconstructed to satisfy the requirements of different applications. For instance, digital geometric modeling is offered for the purpose of architectural design and simulation [20]. In addition to this model, a building's spatial information is also defined and organized by several spatial data standards, such as Industry Foundation Classes (IFC) [21], City Geography Markup Language (CityGML) [22], and Indoor Geography Markup Language (IndoorGML), to support location-based services and other indoor applications.

However, unlike their outdoor counterparts, the 3D reconstruction of indoor environments still poses specific challenges due to the nature of the complicated layout of the indoor structure, the complex interactions between objects, clutter, and occlusions [23]. For example, a lack of view coverage makes it hard to obtain data about walls, floors, and other structures of interest during data collection, leading to unsatisfactory reconstruction results [24]. Another typical example is the difficulty in recovering interior structures and the topological relationships (e.g., connectivity, containment, or adjacency) between them. Additionally, weakly textured regions (such as featureless walls or floors) commonly exist in indoor environments, which causes photo-consistency measurement errors [25]. Obviously, sensor noise and outliers further complicate the reconstruction processes. Furthermore, major appearance differences, illumination, and view changes across different scenes also make it remarkably challenging to automatically and robustly produce an indoor model.

To address these challenges, numerous methods have been developed in recent years related to 3D reconstruction of indoor environments; however, this is still an active research topic in both computer graphics and computer vision fields. By searching for keywords in Web of Science, Figure 1 summarizes the publication statistics related to the 3D reconstruction of indoor environments, covering the period from 2000 to 2019. These keywords included combinations of "indoor scene," "indoor environment," "reconstruction," "modeling," "depth estimation," "simultaneous localization and mapping (SLAM)," "layout estimation," "semantic segmentation," and "topology." From 2010 onward, a significant increase in the number of such studies can be observed, which suggests the growing importance of this research topic. Therefore, this paper presents a systematic review of the techniques of 3D reconstruction of indoor environments. Because of the importance in the development of indoor environment reconstruction, in Section 2 we first provide a brief description of some available

benchmark datasets in order to evaluate the performance of different algorithms. Then, the data collection process of 3D indoor space features is explained in Section 3, before a detailed analysis and summaries of the existing studies related to the 3D reconstruction of indoor environments are offered in Section 4, which can be divided into different categories according to their inherent principles and application requirements. This section focuses on the introduction of the relevant theories, as well as their advantages and disadvantages, instead of a quantitative analysis. Finally, a discussion of future research considerations is presented with our own conclusions about the state of the art of 3D reconstruction of indoor environments in Sections 5 and 6, respectively.

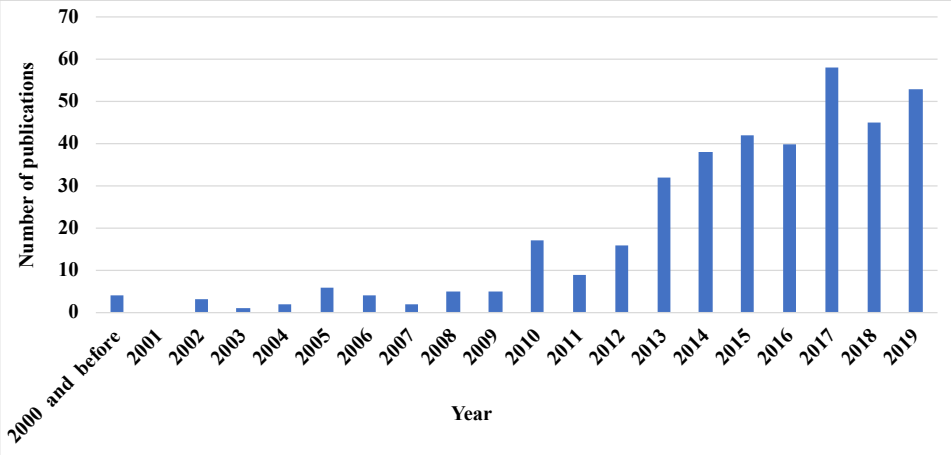


Figure 1. Publication statistics related to the three-dimensional (3D) reconstruction of indoor environments.

2. Benchmark Datasets

Benchmark datasets play a significant role in the objective verification of the performance and robustness of developed algorithms. To date, an increasing number of datasets dedicated to various applications have become available, which are of great importance to measure the current state of the art. In this section, the available benchmark datasets are classified according to their primary purposes, and corresponding brief descriptions are also provided. It should be noted that most available benchmark datasets are constructed for multi-tasks, rather than a single specific task only. Moreover, to the best of our knowledge, there is no benchmark available for qualitatively and quantitatively evaluating topological modeling, although numerous methods have been developed to represent topological relationships.

2.1. Geometric Modeling Benchmark Datasets

2.1.1. Imperial College London and National University of Ireland Maynooth (ICL-NUIM)

The Imperial College London and National University of Ireland Maynooth (ICL-NUIM) dataset [26] is collected from handheld RGB-D ("D" refers to a "depth" or "distance" channel) camera sequences for evaluating visual odometry, 3D reconstruction, and SLAM algorithms, which not only offers ground-truth camera pose information for every frame, but also enables the user to fully quantify the accuracy of the final map or surface reconstruction produced.

2.1.2. Technical University of Munich (TUM)

The Technical University of Munich (TUM) dataset [27] contains RGB and depth images captured using Microsoft Kinect. The 39 image sequences are recorded over an office environment and an industrial hall, and the corresponding ground-truth trajectory is provided for evaluating tasks such as visual odometry and SLAM. This dataset covers a large variety of scenes and cameras for evaluating specific situations (e.g., slow motion, longer trajectories with or without loop closure detection).

2.1.3. European Robotics Challenge (EuRoc)

The European Robotics Challenge (EuRoC) dataset [28] consists of 11 stereo sequences covering three different environments: two indoor rooms and one industrial scene. According to the flight speed, lighting conditions, and texture conditions of the drone, different datasets are presented. Each dataset provides a complete image frame and accurate ground-truth data, as well as important parameters for capturing the camera's internal information and that of other sensors.

2.1.4. Multisensory Indoor Mapping and Position (MiMAP)

The International Society for Photogrammetry and Remote Sensing (ISPRS) benchmark on multisensory indoor mapping and position (MiMAP) is the first dataset that links the multiple tasks of light detection and ranging (LiDAR)-based SLAM, building information model (BIM) feature extraction, and smartphone-based indoor positioning all together [29]. This benchmark includes three datasets: an indoor LiDAR-based SLAM dataset, a BIM feature extraction dataset, and an indoor positioning dataset. Each scene in the dataset contains point clouds from the multi-beam laser scanner, images from fisheye lens cameras, and records from the attached smartphone sensors based in indoor environments of various complexities. The MiMAP project provides a common framework for the evaluation and comparison of LiDAR-based SLAM, automated BIM feature extraction, and multisensory indoor positioning.

2.1.5. International Society for Photogrammetry and Remote Sensing (ISPRS) Benchmark on Indoor Modeling

The ISPRS benchmark on indoor modeling datasets [30] consists of several point clouds (including Technische Universität Braunschweig 1 (TUB1), TUB2, Fire Brigade, Uvigo, University of Melbourne (UoM), and Grainger Museum) captured from different sensors in indoor environments of different complexities to enable comparison of the performance of indoor modeling methods.

2.2. Semantic Modeling Benchmark Datasets

2.2.1. SUN RGB-D

The Scene Understanding (SUN) RGB-D dataset [31] contains 10,335 RGB-D images from four different sensors with dense annotations, and it also includes 146,617 two-dimensional (2D) polygons and 64,595 3D bounding boxes with accurate object orientations, as well as a 3D room layout and scene category for each image. For this dataset, half of the images are allocated to training, while the other half to testing, in order to evaluate scene classification, semantic segmentation, 3D object detection, object orientation, room layout estimation, and total scene understanding.

2.2.2. ScanNet

ScanNet [32] is a large RGB-D video dataset with 2.5 million data frames across 1513 scenes. These 1513 scans represent 707 distinct spaces, including small ones such as closets, bathrooms, and utility rooms, and large spaces such as classrooms, apartments, and libraries, annotated with 3D camera poses, surface reconstructions, and semantic segmentation. This dataset is divided into 1205 scans for training and another 312 scans for testing.

2.2.3. New York University (NYU) Depth

The New York University (NYU) depth dataset consists of video sequences from various indoor scenes. NYU depth v1 [33] and v2 [34] were introduced in 2011 and 2012, respectively, both captured using Microsoft Kinect. For NYU depth v1, there are 64 different indoor scenes, 7 scene-level classes, and 2347 dense labeled frames, divided into 60% for training and 40% for testing. NYU depth v2 contains 1449 RGB-D images describing 464 diverse indoor scenes across 26 scene-level classes,

and detailed annotations are provided for each image. This dataset is divided into 795 images for training and 664 images for testing.

2.2.4. Stanford Two-Dimensional-Three-Dimensional (2D-3D)-Semantic Dataset

The Stanford 2D-3D-Semantic dataset [35] offers a large-scale indoor space dataset, capturing 2D, 2.5D, and 3D data with instance-level semantic and geometric annotations. The dataset is collected in six large-scale indoor areas, including three different buildings of educational use and three of office use, which covers more than 6000 m^2 and contains 70,000 RGB images, along with the corresponding depths, surface normal, semantic annotations, global XYZ images, and camera information.

2.2.5. Matterport3D

Matterport3D [36] offers a large and diverse RGB-D dataset for indoor environments. It contains 10,800 panoramic views from 194,400 RGB-D images describing indoor scenes of 90 buildings. Unlike other datasets, it includes both depth and color 360 panoramas for each viewpoint, annotated by surface reconstructions, camera poses, and 2D and 3D semantic segmentations. The precise global alignment and comprehensive, diverse panoramic views over entire buildings enable keypoint matching, view overlap prediction, normal prediction from color, semantic segmentation, and region classification.

3. Data Collection of 3D Indoor Spaces

Nowadays, there are many ways to collect or obtain 3D data of indoor spaces. In this section, we summarize the following three ways: single-view depth estimation, multi-view depth estimation and SLAM. It should be noted that an increasing number of data-collection methods or sensors (such as radar [37]) are emerging for different indoor applications.

3.1. Single-View Depth Estimation

The main goal of depth estimation is to directly estimate the depth of image pixels for reflecting the real 3D scene observed in an image. Single-view depth estimation allows the depth information to be recovered only using a monocular image [38]. The searching keywords included combinations of "single image/view depth estimation," "monocular depth estimation," and "indoor scene/environment." Figure 2 lists the publication statistics related to monocular depth estimation. It is well-known that spatial context is usually introduced to encode the relationships between neighboring entities for guaranteeing spatial consistency across the images. To introduce the spatial context, Liu et al. [39] developed a discrete–continuous graphical model, where the data term was defined based on the retrieved candidate depth map, while the smooth term was constructed using the occlusion relationships between neighboring super-pixels. Without any additional priors, Zhuo [40] organized the super-pixel, region, and layout layers into a hierarchical framework, where the local, mid-level, and high-level information was encoded under the conditional random field model, with the primary purpose of estimating depth from a monocular image.

Unlike the previous work that used hand-crafted features with prior information, deep-learning techniques show superior performance in feature representation for depth estimation. Eder et al. [41] established a convolutional neural network into which the principal curvature was integrated for jointly estimating dense depth, surface normal, and the boundaries from a single omnidirectional image. Roy et al. [42] combined a convolutional neural network and regression forests, where individual regression results were integrated into a final depth estimation, based on a monocular image. Without considering any prior information, Liu et al. [43] jointly adopted both a continuous conditional random field and a deep convolutional neural network for carrying out depth estimations from a single image.

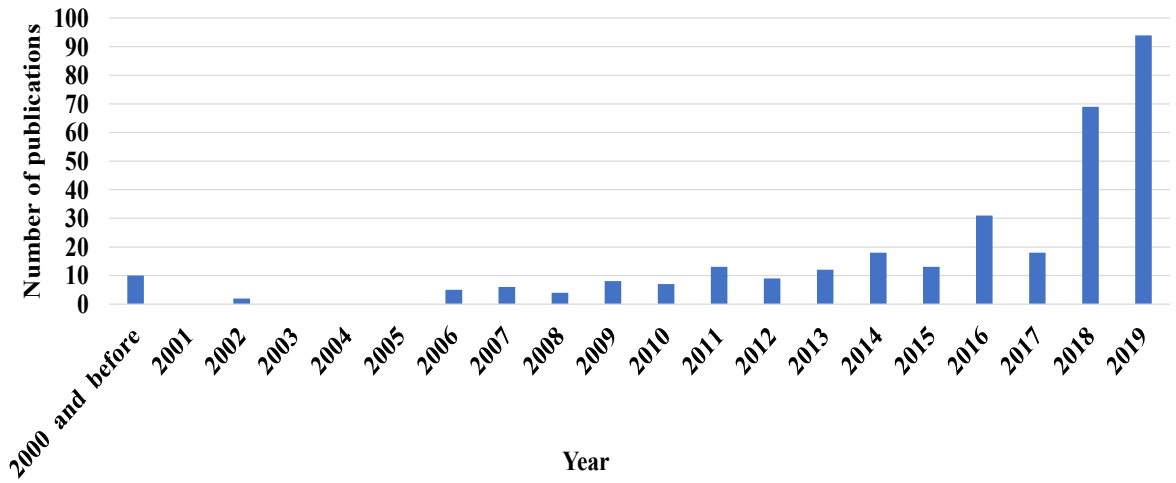


Figure 2. Publication statistics related to monocular depth estimation.

3.2. Multi-View Depth Estimation

Multi-view depth estimation collects multi-view images from calibrated or uncalibrated cameras to reconstruct a dense 3D representation of the scene based on camera parameters and poses, which has always been a hot topic in photogrammetry and computer vision communities. By searching for keywords in Web of Science, Figure 3 summarizes the publication statistics related to multi-view depth estimation ranging from 2000 to 2019. The searching keywords included combinations of "multi-view depth estimation," "indoor scene/environment," and "multi-view stereo."

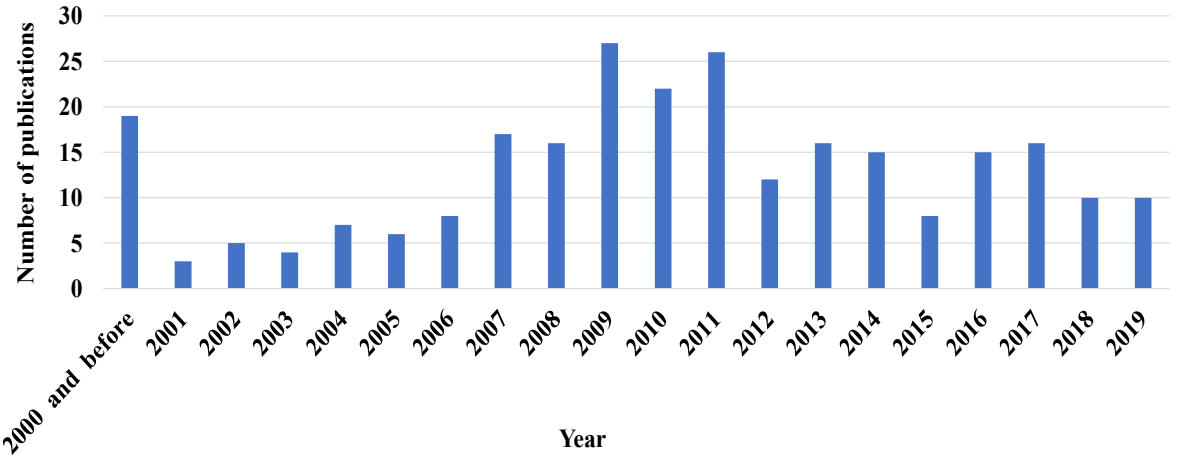


Figure 3. Publication statistics related to multi-view depth estimation.

Figure 4 shows an example of multi-view reconstruction with multiple calibration cameras. Given the calibrated images and the reconstructed geometry, a super-resolution texture map based on 3D geometry models from multiple images can be estimated [44]. This method produces better-textured models with a higher level of detail than that based on a single image.

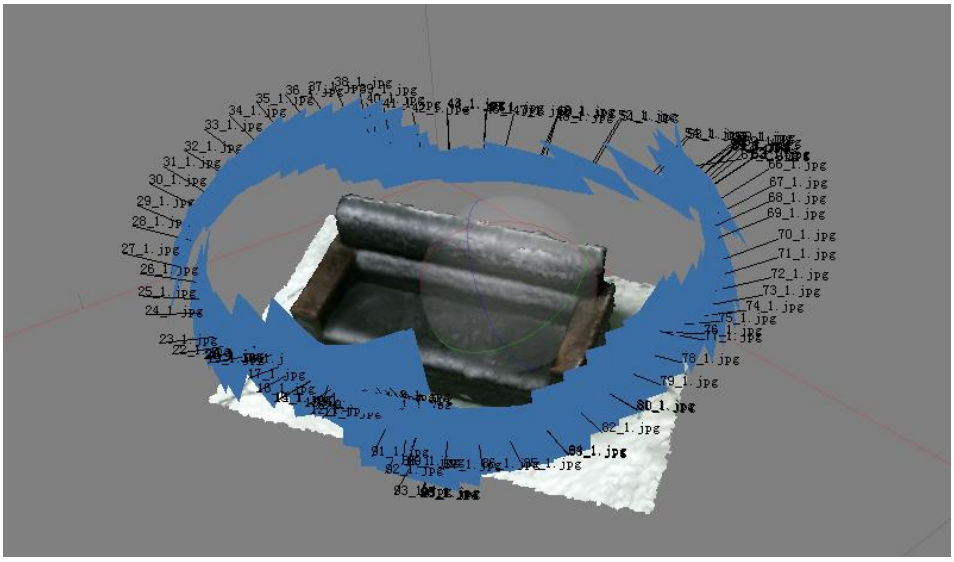


Figure 4. An example of multi-view reconstruction.

However, the image alignment easily affects the quality of the texture map, and the computation volume remarkably increases with the number of input images. Besides the classical algorithm of bundle adjustment, traditional multi-view stereo methods focus on computing plane sweep volume and optimizing photometric error functions to measure the similarity between patches for depth estimation. Collins et al. [45] proposed a simple plane sweeping method for sequentially regularizing the cost volume. Furukawa et al. [46] presented a multi-view stereo algorithm by introducing local photometric consistency and global visibility constraints to produce a dense depth map. Galliani et al. [47] iteratively propagated and refined per-view 3D depth and normal field to maximize the defined photo consistency measure for generating high-quality multi-view matching. Moreover, other complementary information has been added into the reconstruction procedure in order to enhance the performance. For example, Langguth et al. [48] introduced shape-from-shading energy into the reconstruction optimization procedure. Based on the Manhattan assumption, Langguth et al. [48] also used Markov random field to perform plane hypothesis-based optimization for improving the reconstruction performance, especially at the weakly textured regions. Häne et al. [49] jointly conducted semantic segmentation and dense reconstruction, where the former is able to provide information about the likelihood of the surface direction and the latter can offer the likelihood of the semantic class. The structure-from-motion (SfM) technique [50], as a type of multi-view depth estimation, is a successful reconstruction method that utilizes RGB information. Most SfM systems for unordered image collections are incremental, starting with a few images, repeatedly matching features between two images, adding matched images, triangulating feature matches, and performing optimizations to refine camera poses.

Just like other applications, deep-learning techniques have also shown their superior performance in multi-view depth estimation. In more recent years, convolutional neural networks have been introduced to measure the similarity among patches [51]. Ji et al. [52] developed an end-to-end multi-view stereo learning framework, which directly learns both the photo consistency and geometric structure relationships of surfaces. Huang et al. [53] used a deep convolutional neural network to aggregate information from a set of pre-produced plane sweep volumes for multi-view stereo reconstruction. Yao et al. [54] developed an end-to-end multi-view stereo architecture for depth estimation, where the reconstruction problem is decoupled into per-view depth estimation by establishing the cost volume. To address the memory consumption from which the previous work [53] suffered in terms of reconstructing high-resolution scenes, Yao et al. [55] also designed a recurrent neural network-based multi-view stereo framework by regularizing the cost volume in a sequential way.

3.3. Simultaneous Localization and Mapping (SLAM)

A more general framework as a real-time version of the SfM technique is the SLAM method [56], where a mapping system starts motion from an unknown location in an unknown environment to estimate its own positions according to the surrounding environment during the movement, while using its own positions to build an incremental map. SLAM is the process of capturing and recovering the geometrical structures of the whole scene by using either active (e.g., laser sensor [57] and depth cameras [58]) or passive (e.g., monocular [59] and binocular stereo [60]) sensing techniques. Figure 5 summarizes the publication statistics related to SLAM, covering the period from 2000 to 2019, after searching for the keywords in Web of Science. The searching keywords included combinations of "indoor scene, indoor environment, reconstruction, simultaneous localization and mapping (SLAM)." Moreover, according to the type of sensor, several public datasets with ground truth data (e.g., ICL-NUIM, TUM, EuRoC) have been released for evaluating the performance of different methods. Leonard and Whyte [61] originally developed a SLAM system called extended Kalman filter (EKF)-SLAM, where the probabilistic method was used to alleviate the effect of the inaccurate sensors, and this system has become the standard implementation.

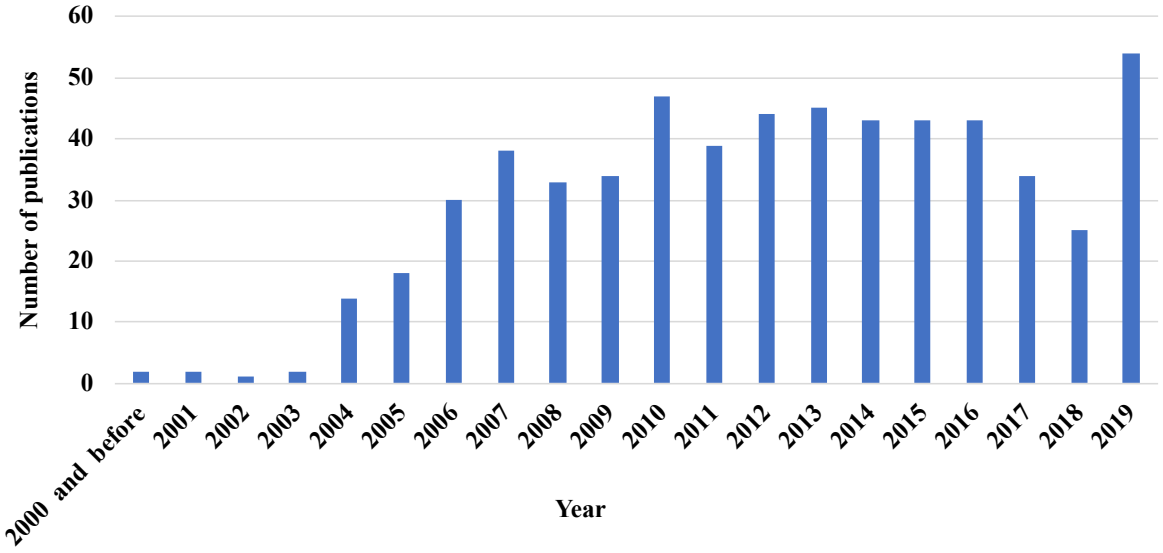


Figure 5. Publication statistics related to simultaneous localization and mapping (SLAM).

After a period of development of more than 30 years, visual SLAM has achieved remarkable results, especially with the appearance of Oriented FAST and Rotated BRIEF (ORB)-SLAM2, which have pushed the traditional visual SLAM system to its peak. However, when single-feature-based SLAM systems are used in low-textured scenes, the performance of their positioning and mapping is often degraded. Modern visual SLAM has gradually developed into a multi-feature, multi-sensor, and deep learning-based method. Unlike the ORB-SLAM solution, the Point-Line-Semi-direct Monocular Visual Odometry (PL-SVO) [62] was developed based on point–line features, where incremental pose estimation is performed and the non-linear minimum reprojection error of point–line features is jointly used to achieve camera motion estimation. Furthermore, a complete binocular Point-Line- SLAM system (PL-SLAM) was also developed, which was built on PL-SVO [62]. Based on the ORB-SLAM system, Pumarola [63] incorporated line features for designing a monocular point and line SLAM that builds a tracking model. Gomez et al. [62] proposed adaptively weighing the errors of different features according to their covariance matrix. In addition to the traditional reprojection error of the endpoints for establishing a Gauss–Newton estimation model by minimizing the angle observation, Wang et al. [64] proposed adjusting the weight ratio of the points and lines based on the estimation of the camera state residuals. Semantic SLAM raises image features to the object level and provides an understanding of the surrounding environment. Combined with the currently emerging deep-learning

methods, the results of semantic segmentation and target detection can provide SLAM with higher-level information. More recently, researchers have used the expectation maximization (EM) method [65], as well as different dynamic object detection methods, to remove the dynamic object. Consequently, data association is performed through the probability model, where the objects detected in the image are correctly mapped to the 3D objects that are already present in the map data, thus generating more robust results.

4. Overview of the Research Methodology for 3D Reconstruction of Indoor Environments

This section covers the progress in the field of indoor environment modeling, which can be divided into three aspects according to the modeling purpose: geometry modeling, semantic modeling, and topological modeling. Figure 6 summarizes techniques available for the 3D reconstruction of indoor environments. The main purpose of geometry modeling is to fully recover 3D geometry of the indoor environment, including polygonal structural elements modeling (e.g., walls, floors, ceils, doors, windows), room layout estimation, and indoor–outdoor seamless modeling. These generated models can be evaluated based on the benchmarks (e.g., ICL-NUM, TUM, EuRoc, MiMAP, ISPRS benchmark on indoor modeling). Semantic modeling focuses on semantic labeling (e.g., object types) using the machine-learning or deep-learning framework. The common datasets (e.g., SUN RGB-D, ScanNet, NYU depth, Stanford 2D-3D-Semantic dataset, Matterport3D) are generally used as benchmarks to verify the interpretation results. Topological modeling is mainly used to recover the topological relationships (e.g., connectivity, containment, or adjacency) between indoor space units, which is represented by subdivision-based methods, grid-based methods, and hybrid methods.

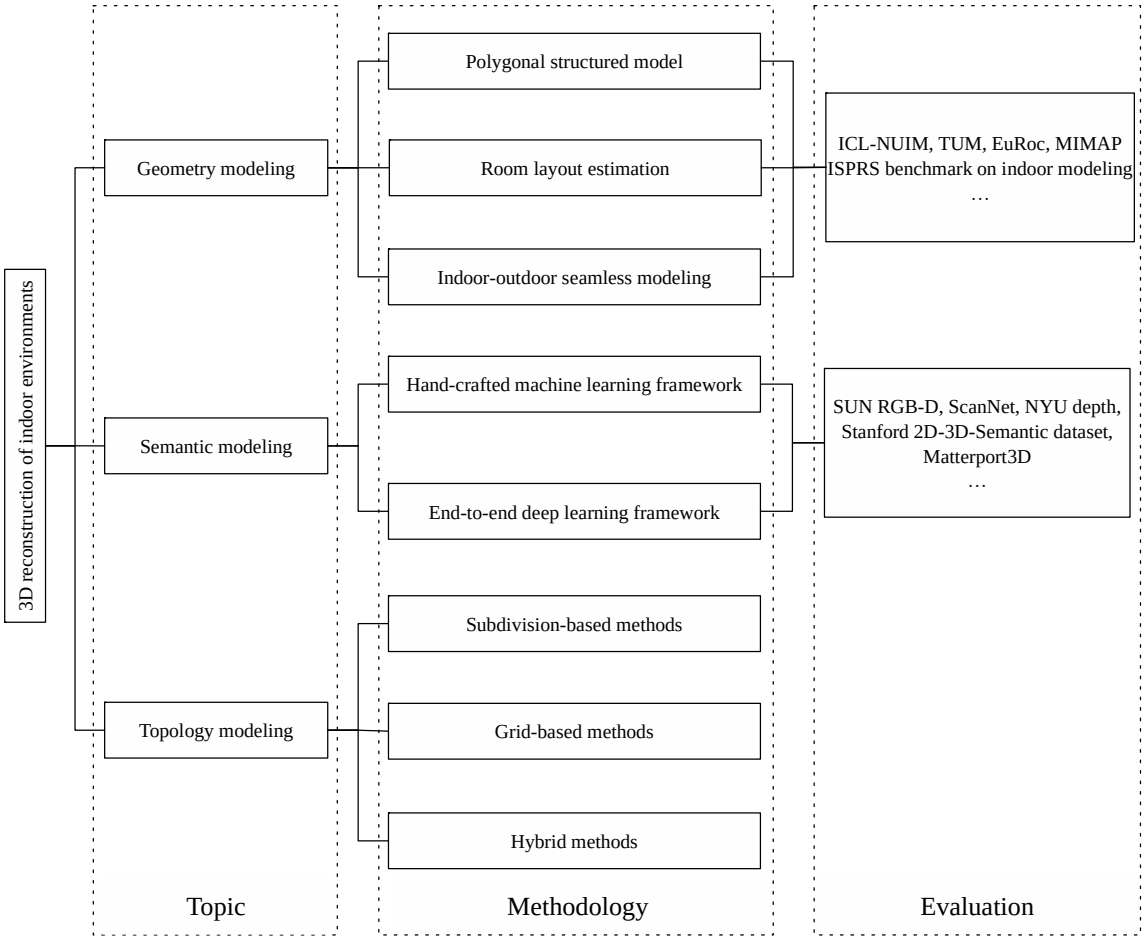


Figure 6. Technique overview for the 3D reconstruction of indoor environments.

4.1. Geometry Modeling

4.1.1. Polygonal-Structured Model

The generation of 3D building models has many far-reaching applications in the architectural engineering and construction community. However, manually creating a polygonal 3D model of a set of floor plans is non-trivial and requires skill and time. Currently, researchers are trying to automate the reconstruction of 3D building models. In this section, it is shown that a polygonal-structured model mainly focuses on determining structural elements (e.g., walls, ceilings, and floors) or wall-surface features (e.g., windows and doors). In addition, public datasets, such as the ISPRS benchmark on the indoor reconstruction dataset, are used as benchmark datasets and a common evaluation framework is used for a comparison of the performance of different methods.

To design and modify a complex 3D architectural building, 2D architectural drawings often provide a more effective means to this end. With architectural and semantic information, 2D architectural drawings as effective data sources have been used for the reconstruction of 3D building models. Nevertheless, most drawings take the form of floor plans, which portray an orthographic projection of each building level using a set of standardized symbols regarding architectural elements. Based on this fact, vectorization and symbol recognition play important roles in reconstruction, and numerous methods for which 2D architectural drawings serve as the input for generating the associated 3D models have been developed. For instance, So et al. [66] developed a semi-automated reconstruction method of virtual 3D building models from 2D architectural drawings to improve the efficiency of the previous manual reconstruction process. Lu et al. [67] combined the architectural information in multiple drawings, semantics, and prior knowledge to reconstruct a 3D building model from 2D architectural drawings. Lee et al. [68] presented a framework for viewing images in order to implement perception-based techniques, as well as a hinging-angle scheme for interactive sketching. Horna et al. [69] proposed a formal representation of the consistency constraint for 3D reconstruction and its associated topological models. Li et al. [70] performed an efficient reconstruction from architectural drawings, where matching and classification algorithms were used to reduce the interaction during the 3D reconstruction. Although these methods based on 2D architectural drawings have been presented, complete automation for generating 3D building models from 2D architectural drawings has failed to be achieved due to the ambiguities or inconsistencies of architectural representations [71].

Besides digitalization and extrusion from 2D architectural plans into 3D polygonal models, the development of laser-scanning systems (e.g., terrestrial or mobile) and photogrammetric techniques using images offers practical solutions to render real 3D indoor environments. The existing methods related to parametric-structured models directly derived from raw 3D data can be generally divided into structural element extraction methods and space decomposition-and-reconstruction methods.

Structural element extraction methods usually follow the strong Manhattan assumption when modeling complex indoor environments, and divide the indoor scene into ceilings, walls, and floors using normal analysis methods [72], least square-based methods [73], region-growing methods [73], random sample consensus (RANSAC)-based methods [74], and threshold-independent Bayesian sampling consensus (BaySAC) [75]. Figure 7 shows an example of a structural element extraction method, which can be considered a polygonal-structured modeling paradigm, including the building structure detection stage and the associated parameterized (or vectorized) stage. Under the strong Manhattan assumption, the main architecture of buildings consists of a set of planar primitives, which offers key insights behind most structural element extraction methods. However, the Manhattan assumption might not be true in the real environment, hence limiting the usage of this method. Jung et al. [76] conducted plane detection using the RANSAC algorithm and tracked the boundaries for outline extraction. Wang et al. [77] detected and classified building components using the region-growing algorithm for building an information model. Ning et al. [72] identified wall structures based on a normal variant and created a 3D model using the extracted wall structures. To produce a more realistic parametric model, Previtali et al. [74] extracted planar primitives using a modified

RANSAC technique, further refined the extracted planar primitives using a graph cut algorithm, and finally detected both doors and windows using the ray tracing method for generating the CityGML or IFC model. Without using any prior information, Shi et al. [78] carried out 3D segmentation by using a combination of different information, and then reconstructed the room layout for generating a semantic 3D model. Instead of extracting the 3D main structures of buildings, Hong et al. [79] modeled the main wall structures in the 2D domain to extract the 2D floor boundary, which can then be further combined with vertical modeling for producing a 3D wireframe model from point cloud data. To accurately detect the main architectural structures, Michailidis and Pajarola [80] focused on openings (e.g., windows and doors) extraction of cluttered and occluded indoor scenes from point clouds using Bayesian graph cut methods. More recently, research focusing on the classification and labeling issue of objects shows superior performance for machine-learning and deep-learning techniques, which is a prerequisite for subsequent applications. Unlike the previous planar detection methods to extract the main structures of buildings, Wang et al. [14] presented semantic classification methods to establish a semantic line-based framework for indoor building modeling, and to use the conditional generative adversarial nets method for optimizing the occlusion situation.

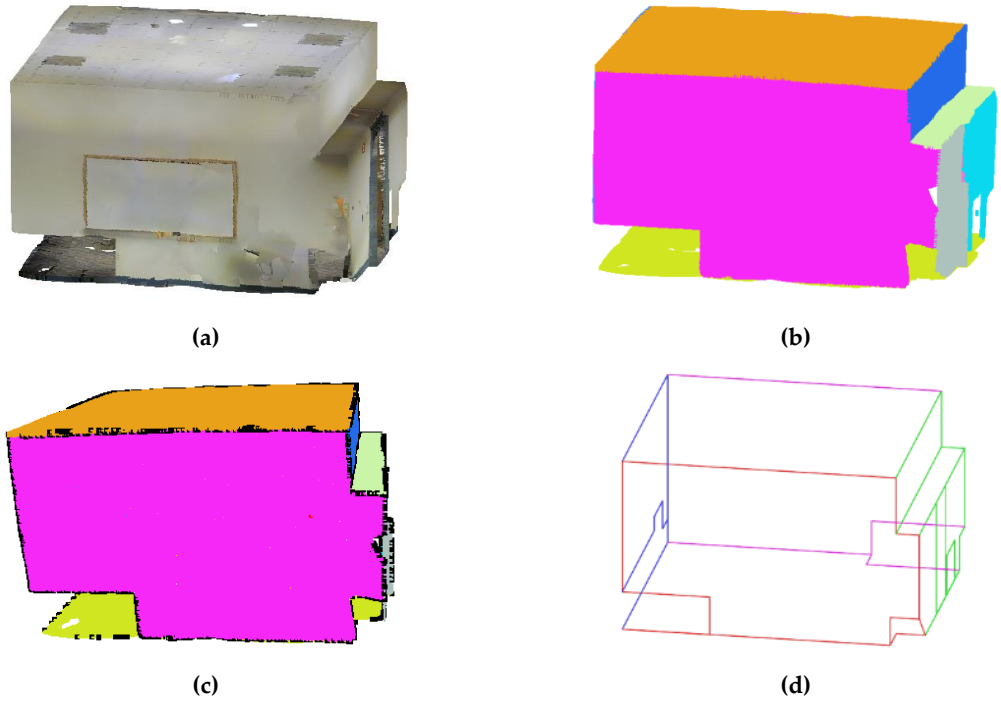
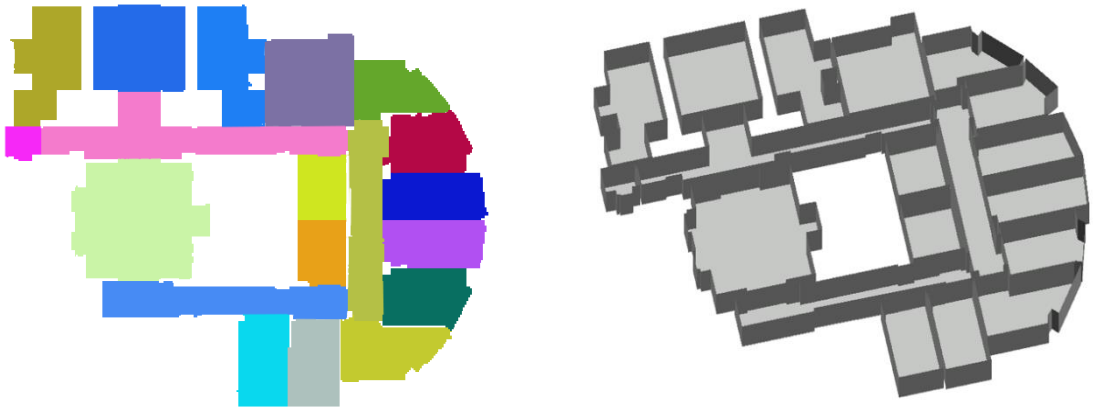


Figure 7. An example of a structural element extraction method. (**a**) Raw point clouds; (**b**) plane segmentation; (**c**) boundary extraction; (**d**) final wireframe model. In (**b**), different planes are randomly rendered in different colors. In (**c**), the extracted boundary points are rendered in black.

The space decomposition-and-reconstruction methods use prior knowledge of indoor architectural structures and formulate the scene reconstruction issue into an indoor space decomposition-and-reconstruction problem in order to achieve a more robust reconstruction result. Similar to structural element extraction methods, most of the space decomposition-and-reconstruction methods are also conducted based on the strong Manhattan assumption. With regard to the space decomposition-and-reconstruction methods, the room units are defined from different perspectives. Figure 8 outlines examples of space decomposition-and-reconstruction methods [20], where the units of the building (i.e., rooms) are partitioned and modeled, separately. Although these space decomposition-and-reconstruction methods improve the robustness and performance or reconstruction by partitioning the complex indoor spaces into individual simple rooms, the quality of the 3D models is also affected by the segmentation results. More recently, the scanner positions or the motion trajectories

usually offer important information for individual room segmentation. By exploiting the knowledge of the scanner position, Mura et al. [81] carried out the candidate wall-based space partitioning algorithm for automatically segmenting individual rooms, and modeled each room as a 3D polygon model. Tang et al. [82] partitioned the indoor environment into several units by the start–stop strategy, and then extracted and modeled the main architectural structures (e.g., walls, floors, ceilings, and windows) from RGB-D images for establishing semantically rich 3D indoor models stored in CityGML 3.0 standard. Although the scanner positions or the motion trajectories enhance partition performance, they are not always available. Without using timestamp information, Wang et al. [83] presented a graph cut-based labeling framework for reconstructing structural walls from the extracted line primitives, and then reconstructed rooms for generating 3D pylon models. To guarantee the integrity of the space partitioning and space geometric regularity, Li et al. [84] developed a comprehensive segmentation-based method for multi-story indoor environment reconstruction, where multi-story buildings can be segmented into several floors using a peak–nadir–peak strategy and where each floor can be partitioned into rooms and corridors. Similar to the previous line primitive-based methods [83], Pang et al. [85] integrated space boundary and Boolean difference for indoor space extraction to produce an indoor model with geometric, semantic, and relationship information. Instead of assuming the Manhattan world, Yang et al. [20] constructed complex indoor environments with both straight and curved wall structures by minimizing an energy function based on Markov random field formulation.



(a) (b)

Figure 8. Examples of space decomposition-and-reconstruction methods. (**a**) Space decomposition; (**b**) reconstruction result.

In summary, the existing 2D architectural drawing-based solutions share a common pipeline for automated conversion of 2D environments into 3D models. However, most architectural drawings take the form of floor plans with various levels of detail and use varying graphic symbols, which results in ambiguities or inconsistencies in architectural representations. Structural element extraction methods usually exploit model-fitting methods to fit geometric primitives (such as planes, cylinders, spheres, and cones) and then use these to compute intersections and corners [86], which is easily affected by noises or occlusions. Meanwhile space decomposition-and-reconstruction methods formulate scene reconstruction into an indoor space decomposition problem, and can be used in cluttered indoor environments. However, the over-segmentation issues from which the room segmentation algorithms suffer need to be optimized at the post-processing stage. Also, both structural element extraction methods and space decomposition-and-reconstruction methods mainly focus on the architectural structures of buildings. In addition to the architectural structures of buildings, interior objects are also modeled by using model retrieval methods [10,18,24,87,88] to generate more complete indoor scene models. Moreover, the quantitative evaluation in geometry between the derived model and the ground truth data is useful for comparing the geometric quality [89].

4.1.2. Room Layout Estimation

Reconstructing the room layout of an indoor environment is based on locating the boundaries of the walls, floors, and ceilings [90], which offers high-level contextual information about the scene (even in cluttered environments) [91]. However, since the presence of clutter results in the occlusions, most of the existing methods are generally based on inherent geometric assumptions (e.g., Manhattan or even box-shaped layout) to delineate the key corner positions or boundaries that specify the room layout. According to the different perspectives of layout estimation, the existing methods can be categorized into area-based methods [92] (as shown in Figure 9a), edge-based methods [93] (as shown in Figure 9b), keypoint (corner)-based methods [94] (as shown in Figure 9c), and hybrid methods [95] (as shown in Figure 9d). Obviously, these methods are based on prior knowledge of the real indoor environment and interpret the scene layouts from different perspectives (i.e., region, edge, corner). However, most of these methods fall under a strong geometry assumption (e.g., Manhattan geometry). For complicated layouts (e.g., L-shaped layouts, curved walls), these methods might be not applicable. Meanwhile, public datasets, e.g., NYU, SUN RGB-D, Stanford 2D-3D-Semantic dataset, are used as benchmarks to evaluate these methods.

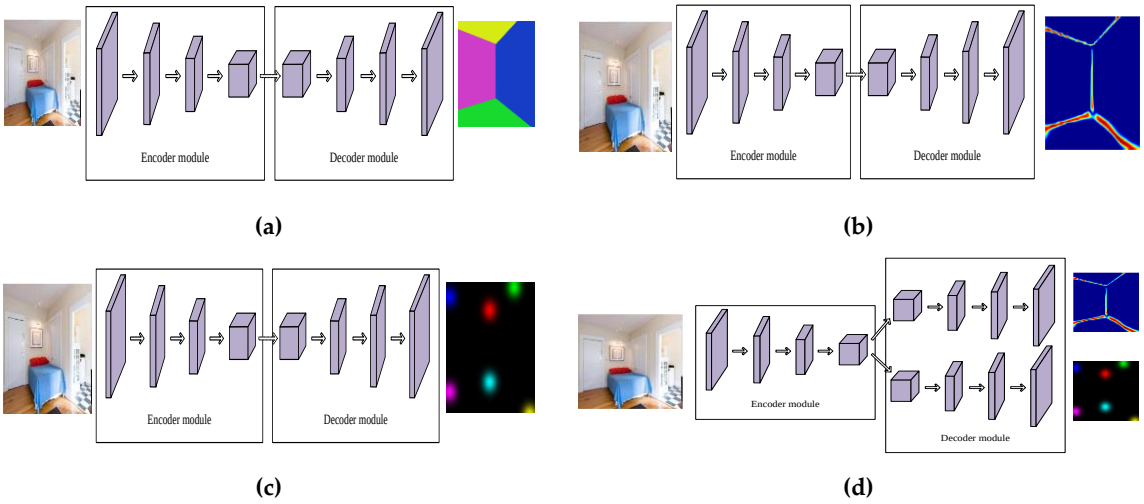
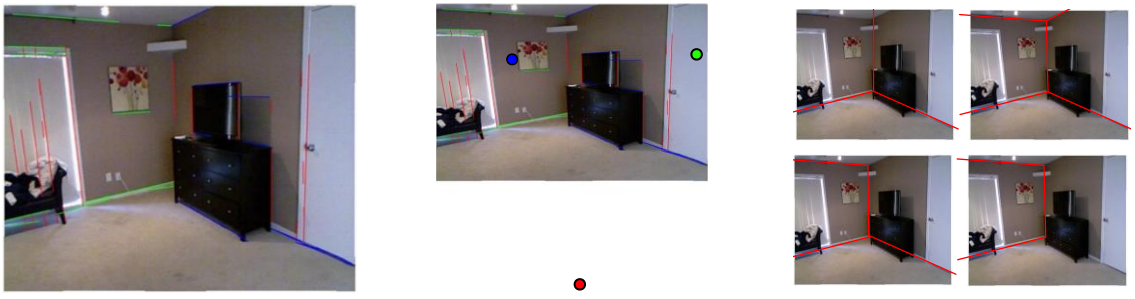


Figure 9. Typical examples of different methods. (**a**) Area-based methods; (**b**) edge-based methods; (**c**) keypoint (corner)-based methods; (**d**) hybrid methods.

Area-based methods generally rely on the geometric features of local regions, such as the geometric context and orientation map, to determine the spatial orientation using the vanishing point-based method [96] (as shown in Figure 10); then, the optimal layout scheme is selected from a set of layout candidates. The vanishing point-based method is based on the perspective geometry theory, and is also limited by the strong Manhattan assumption. Since the vanishing point-based layout hypotheses generation methods fail in highly cluttered indoor environments, Chao et al. [97] exploited the relationship between people and the room box, as well as introducing both geometric and semantic cues to improve the performance of vanishing point estimation. Park et al. [98] established a conditional random field framework, where semantic features derived from semantic segmentation architecture and orientation maps are integrated, to estimate an indoor layout from single images. Due to their superior performance, deep-learning techniques are also used to provide high-level features for layout estimation or to achieve an end-to-end layout estimation framework. Dasgupta et al. [2] classified a single RGB image into five categories (i.e., left wall, front wall, right wall, ceiling, and ground) using a fully convolutional network, and a refinement framework was then conducted to recover the spatial layout. Lin et al. [92] designed an end-to-end fully convolutional network architecture by defining an adaptive edge penalty and smoothness terms to achieve planar semantic layout estimation from a single image.



(a) (b) (c)

Figure 10. An illustration of the vanishing point-based method. (**a**) Detected line segments; (**b**) computed vanishing points; (**c**) room layout hypothesizes. In (**a**), a set of line segments are detected and used to find the vanishing points. The line segments in (**a**) are colored using the associated colors of the vanishing points they correspond to. Then, several room layout hypothesizes are generated, which can be ranked by a defined measure function to select the optimal one.

Edge-based methods consider layout estimation as the problem of finding all of the wall–floor, wall–wall, and wall–ceiling boundaries. Mallya et al. [99] defined and found informative edge probability maps using single images through a fully convolutional network, and then used a maximum margin-structured classifier for generating layout estimation. Jahromi et al. [100] presented a hypothesizing–verifying method, where a set of hypotheses were generated through both the random line segment intersection and the virtual rays of vanishing points, and then used the geometric reasoning-based verifying method to search for the optimal hypothesis for reconstructing the spatial layout of a corridor scene. Zhang et al. [93] combined a deconvolutional network with an adaptive sampling strategy to generate a high-quality edge map for room layout estimation.

Keypoint (or corner)-based methods focus on predicting the positions of the 2D keypoints that define the room layout. For instance, Hirzer et al. [101] carried out a corner-targeted hypothesize-and-test strategy, where different segmentation hypotheses were developed based on the different numbers of visible walls to localize the corners and to predict the spatial layout based on a monocular image. Instead of most hypotheses-verifying room layout estimations, Lee et al. [94] developed an end-to-end room layout estimation framework for use on a monocular image, where the keypoints that specify the room layout are localized and ordered. Fernandez et al. [91] also developed a corner-based layout estimation framework from 360 images in an end-to-end manner.

Hybrid methods try to combine different layout elements and complement the limitations of each other. Based on the Manhattan assumption, Chang et al. [102] formulated the indoor spatial layout estimation problem into an energy optimization process from a monocular image, where the cost function is defined using boundary and surface consistency. Zou et al. [95] proposed a method based directly on a panoramic image to predict cuboid and more general layout through a combination of multiple layout elements (e.g., corners and boundaries). Kruzhilov et al. [103] combined keypoint maps and edge maps to predict the room layout using a double refinement network.

Integrating room layout estimation with other tasks (e.g., semantic segmentation or object detection) is simultaneously performed, where room layout estimation can provide the spatial layout knowledge for the reasoning in semantic segmentation or object detection task, while semantic segmentation or object detection can also offer semantic information for room layout estimation. Figure 11 provides an example of multi-task integration, where the edge map and segmentation map are jointly estimated. This shows a general framework about multi-task integration based on a hierarchical structure, which is composed of a single encoder and multiple heads for predicting pre-defined scores based on different requirements. In this case, multiple predictions can be applied to different applications for different purposes, while the learnable parameters are shared for high computational efficiency, especially in settings where the ground truth data for all tasks are expensive to collect. Hedau et al. [96] developed an indoor spatial layout estimation method based on single images,

where spatial layout and surface labeling are jointly conducted and complemented. They proposed a structure-learning algorithm to predict the parametric 3D box model of the global room space while carrying out surface labeling of pixels. Taking the 3D interaction between objects and the spatial layout into consideration, Gupta et al. [104] presented volumetric reasoning between indoor objects and their layout to improve the existing structured prediction framework for layout reconstruction. Pero et al. [105] combined layout prediction with object detection and localization in single images using 3D reasoning and a Bayesian inference for better understanding of indoor scenes. Schwing et al. [106] proposed a branch and bound method for simultaneously performing room layout estimation and object detection in cluttered indoor scenes. Zhang et al. [107] integrated both depth and appearance features from RGB-D images to jointly predict the room layout and the clutter in indoor environments. To jointly predict the 3D room structure and the interior objects, Bao et al. [108] integrated both geometry information extracted from structure-from-motion points with semantic information using multi-view images. Zhang et al. [109] designed an encoder–decoder network architecture, where the edge map and semantic information were integrated, to reconstruct the room layout from a monocular image.

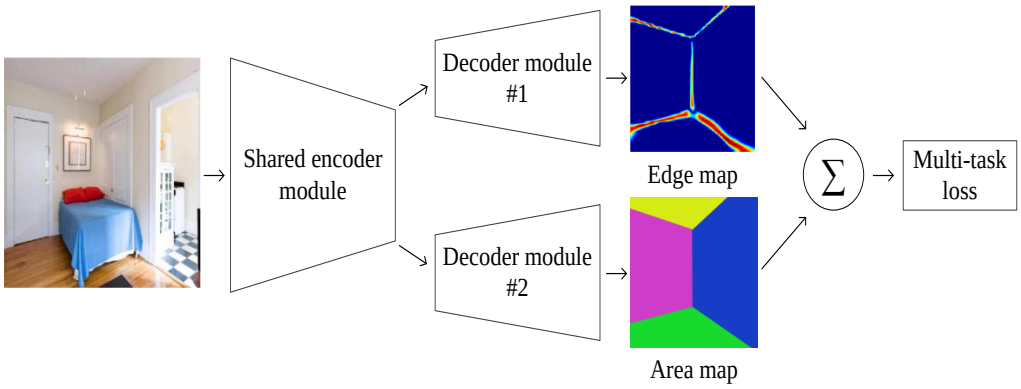


Figure 11. An example of multi-task integration.

In summary, layout estimation is achieved by identifying architectural structures (e.g., ceiling, floor, wall). However, this process is very challenging due to considerable amounts of clutter, varying lighting, large intra-class variance, and occlusions from the furniture. These types of methods attempt to detect the keypoint, regions or edges to locate the architectural structures for achieving layout estimation. More recently, deep-learning techniques effectively provide semantic information to assist in decision making, and its hierarchical structure plays an important role in multi-task collaborative optimization to produce a more reliable layout estimation.

4.1.3. Indoor–Outdoor Seamless Modeling

With the rising demand for indoor location-based service applications, the integration of indoor and outdoor models over the same scene is also a novel topic. To the best of our knowledge, there is a lack of focus in the literature on aligning indoor and outdoor models. Cohen et al. [110] took window information into consideration to match indoor and outdoor scenes. Similarly, Koch et al. [111] also detected the line segments of windows to automatically align indoor and outdoor models. The geometric shape priors of windows provide important information for connecting indoor and outdoor models, which offers a key insight for indoor–outdoor seamless modeling. Nevertheless, the existing methods are limited by these geometric shape priors.

4.2. Semantic Modeling

Unlike geometric modeling, which does not care about what it contains, semantic modeling mainly focuses on semantic labeling (e.g., object types), which plays an important role in semantically rich and geometrically accurate indoor models. Moreover, public datasets, such as SUN RGB-D, ScanNet,

NYU, Stanford 2D-3D-Semantic dataset, Matterport3D, etc., provide pixel-wise or point-wise dense semantic annotation for evaluating the performance of the semantic segmentation results derived from different methods. Most previous research primarily relied on hand-crafted features as the input of the frequently-used classifier for automatic classification and of the probabilistic graphical models, such as conditional random fields (CRFs) [33,112]. Also, Markov random fields (MRFs) traditionally have been used to encode the contextual information in semantic segmentation. Figure 12 summarizes the general classification framework based on hand-crafted features. Silberman and Fergus [33] developed a CRF-based model, combining 3D location, derived beforehand from depth channels, with features captured from both the depth and color channels for indoor scene segmentation. Ren et al. [113] adopted the kernel-based framework for transforming the pixel-level similarity within each super-pixel into a patch descriptor, which was then integrated with contextual information under the framework of MRFs for labeling RGB-D images. Silberman et al. [34] used depth cues for enabling a more detailed and accurate geometric structure to interpret the major surfaces of indoor scenes, i.e., floors, walls, supporting surfaces, and object regions, from RGB-D images and to recover the physical support relations. Gupta et al. [114] effectively made use of depth information for optimizing image segmentation, and defined the features of super-pixels for automatic classification using the random forest classifier and the support vector machine (SVM) classifier. Khan et al. [112] combined the appearance, location, boundaries, and layout of pixels under the framework of CRFs for reasoning about a set of semantically meaningful classes from RGB-D images. Müller and Behnke [115] established a CRF-based framework without prior information regarding scene layout, into which color, depth, and 3D scene features were incorporated, for semantic annotation of RGB-D images. Deng et al. [116] made full use of the 3D geometric structure derived from Kinect, and integrated the global object co-occurrence constraint, the relative height relationship constraint, and the local support relationship constraint using a CRF-based framework for segmenting and annotating RGB-D images of indoor scenes. Unfortunately, these conventional methods usually consist of bottom-up segmentation, feature extraction, and classification, and their final results depend on the results of each stage [3].

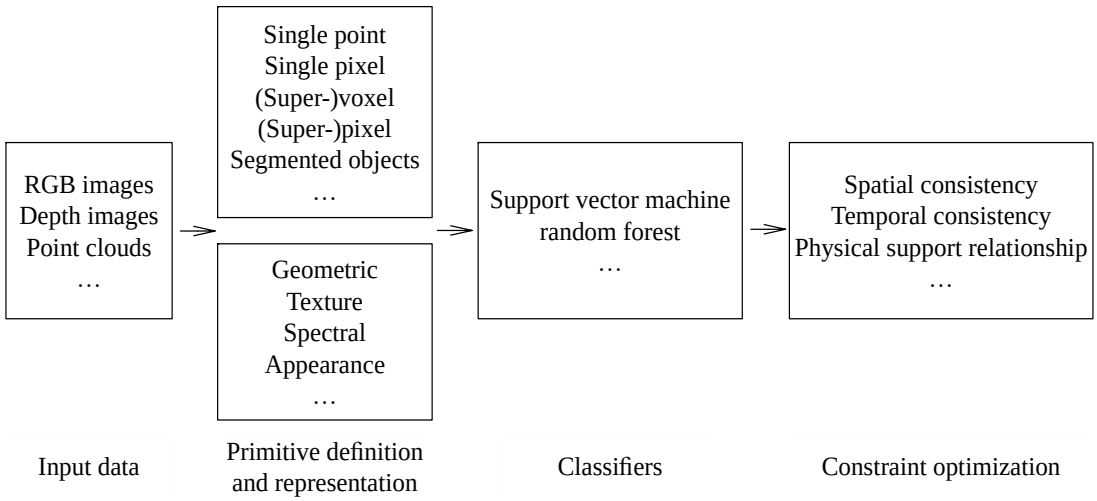


Figure 12. General classification framework based on hand-crafted features.

With the availability of a convolutional neural network (CNN) in many applications, various CNN-based architectures have been developed to extract high-level semantic features for semantic segmentation in recent years. Without the need for engineered features, Farabet et al. [117] captured texture, shape, and contextual information from regions with multiple sizes centered on each pixel to extract dense feature vectors using a CNN architecture for scene labeling. Following the idea of a region CNN (R-CNN) method [118], Gupta et al. [119] encoded the height above ground, the angle with gravity for each pixel, and the horizontal disparity. Then, they extracted a set of regions of interest

from an input image, computed the features for each extracted region using a CNN architecture, and classified each extracted region using an SVM classifier for object detection. Finally, they established a super-pixel-based classification framework [114] on the output of the object detectors for semantic scene segmentation. Although they can produce a powerful feature representation based on CNN architectures, these CNN-based region proposal-based methods exhibit a large amount of repeated computation and make the system itself commit potential errors in the front-end segmentation algorithm [120].

To deal with the limitations from which these CNN-based region proposal methods suffer, fully convolutional network (FCN)-based methods [121] demonstrate efficient feature extraction and end-to-end training, and thus have become increasingly popular for semantic segmentation. Figure 13 outlines an end-to-end classification framework based on deep learning, where some typical modules [122–124] are summarized. Additionally, this type of network can take arbitrarily sized inputs and produce outputs of corresponding size. Hazirbas et al. [125] established an encoder–decoder-based architecture, where the complementary depth cues were fused into the extracted RGB feature maps during the encoder procedures for indoor semantic segmentation. Husain et al. [3] combined semantic features from the color images, the geometric features, and the proposed distance-from-wall feature for object class segmentation of an indoor scene. Jiang et al. [126] combined an RGB-D-based fully convolutional neural network with a depth-sensitive fully-connected conditional random field to refine semantic segmentation results. Cheng et al. [127] incorporated locally visual and geometric information for recovering sharp object boundaries, and used a gated fusion layer for adjusting the combination of RGB and depth cues for improving the performance of object detection. Lin et al. [128] used the depth cues to split the scene into multiple layers with similar visual characteristics, and then proposed a context-aware receptive field for making full use of the common visual characteristics of the observed scenes. Finally, they implemented a multi-branch-based network model for segmenting RGB-D images. To sufficiently exploit contextual information, Li et al. [129] carried out a two-stream FCN to determine the RGB and depth features, and gradually fused these features from a high level to a low level for indoor scene semantic segmentation. Jiang et al. [130] developed an encoder–decoder architecture to extract RGB information and depth information separately and to fuse the information over several layers for indoor semantic segmentation. By incorporating the depth information, the spatial geometric information, which is more invariant to illumination changes and appearances, can be derived for the improvement of semantic segmentation. Guo and Chen [131] adopted a deep CNN for estimating a depth map from a single RGB image and integrated the estimated depth image with the original RGB image for enhancing the performance of indoor semantic segmentation.

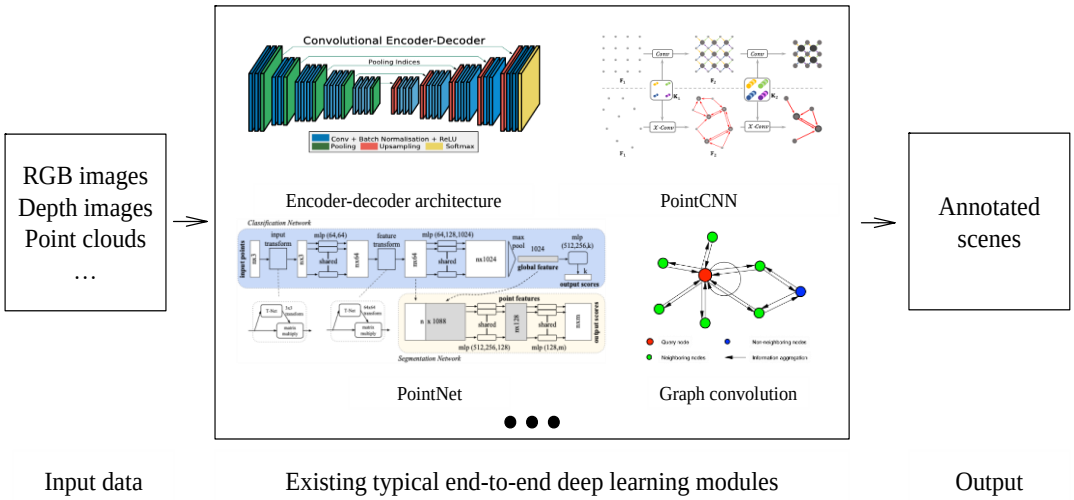


Figure 13. An end-to-end classification framework based on deep learning. CNN, convolutional neural network.

In summary, compared with hand-crafted feature-based methods, an end-to-end framework based on deep-learning architecture exhibits superior performance, in which different scales of information are extracted and combined under a hierarchical structure. Meanwhile, the strong computational capability, especially when using a graphics processing unit (GPU), enables the use of large-scale data for learning. Moreover, due to its superior capability in information aggregation, convolutional operation used not only the regular/structured data (e.g., image or RGB-D image), but also the non-structured data (e.g., point clouds), which provide more solutions to better interpret the indoor scene data and allows more information from different perspectives for making the decision.

4.3. Topological Modeling

Unlike outdoor spaces, where a specific location is defined and represented by coordinate reference systems, the position in an indoor space is usually marked using a cell identifier or a symbolic code. Additionally, the distance between any two positions in an outdoor space can be determined and computed by the linear length in between. However, the situation in an indoor space is different and is affected by architectural structures (such as walls and doors). As a consequence, any spatial information model suitable for outdoor space cannot simply be applied to an indoor space due to their differences (such as the influences of architectural structures and lack of Global Navigation Satellite System (GNSS) signals). Therefore, to manage and maintain complicated indoor space information efficiently, it is necessary to extend and develop 3D building spatial information models for supporting location-based services through topological information [132]. That is, the semantically-rich representation of an indoor environment requires not only geometry and semantics, but also topology [74], which offers useful information about geometrical objects such as door locations or exit points in a building to enable the capability of dynamic routing and door-to-door movement [133]. As shown in Figure 14, a dual graph is constructed to represent connectivity relations (black solid lines) and adjacency relations (red dashed lines) [80]. Thus, spatial relationship recovery plays an important role in indoor applications, ranging from path planning and indoor navigation to localization-based services.

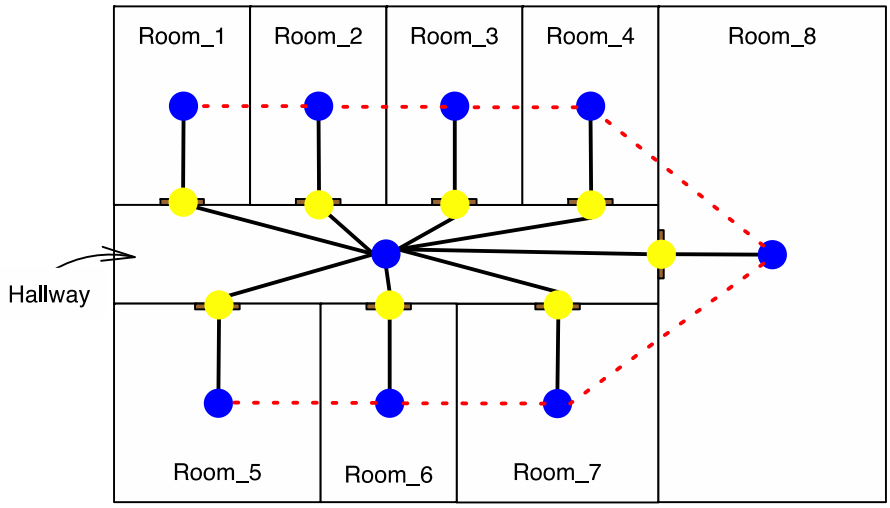


Figure 14. Dual graph representing connectivity relations (black solid lines) and adjacency relations (red dashed lines).

Topology suggests the way in which rooms, corridors, stairs, and doors are interrelated or arranged, usually providing the following information: "Which room is adjacent to which room?" and "Which door connects which room?" [134] The existing methods available to model indoor topological relationships generally can be categorized into three classes: subdivision method models, grid models, and hybrid models. Based on these indoor topological models, complicated spatial analysis (e.g., finding the shortest and most natural path for specific situations) can be implemented [135].

The subdivision method partitions the interior space into a set of non-overlapped cells to analyze and extract their topological relationships, which is usually represented by node-and-edge graph structures. Thrun et al. [136] used artificial neural networks and Bayesian integration to extract high-level information from a grid map, where the critical points on a Voronoi diagram were identified, for generating a topological map. Joo et al. [137] performed corner feature-based virtual door detection and adopted a genetic algorithm to generate a topological map from an occupancy grid map. Portugal and Rocha [138] also extracted a simple topological graph-like representation from a grid map to assist the navigation task. Yang and Worboys [139] transformed a combinational map that included both geometry and semantic information into graph-based indoor structures to describe the topology and connectivity. Tran et al. [134] divided point clouds into navigable and non-navigable space, and then used grammar rules for topological relationship reconstruction. Rather than the subdivision of indoor spaces, Sithole et al. [140] constructed a generalized graph of indoor spaces by iteratively simplifying their floor maps. Based on the graph structures, the subdivision method has high efficiency in terms of path and query analyses, but ignores spatial location information. Grid models are established based on the voxel-based method to consider path size for offering users several indoor paths. Demyen et al. [141] represented the indoor environment based on an irregular triangulation network and implemented pathfinding analysis. Li et al. [142] established a grid graph-based model to represent an indoor space by considering its structural and spatial properties, and showed the proposed model's potential through indoor space analysis. Arjona et al. [143] simplified occupancy grid indoor mapping for low-cost robots. Xu et al. [144] extracted the required geometric and semantic information from the BIM model and established a 2D grid navigational network by considering obstacles. Compared with the subdivision method, grid models retain positional information but increase the data volume and computational complexity, as well as reducing spatial analysis efficiency. To combine the advantages of the subdivision method and the regular grid models, and to alleviate their limitations, the development of a hybrid model can improve the balance of a single model. Li and Lee [145] proposed a lattice-based indoor topology representation model, which is able to provide an explicit representation of not only containment and overlap, but also adjacency on the lattice concept. Lin et al. [146] combined a topology graph and a grid model to enhance the performance of spatial analysis and to provide positional information.

In order to make the interoperability of indoor spatial data available, several spatial data standards (e.g., IFC, CityGML, and IndoorGML) [147] have been developed in recent years to define, organize, and store detailed indoor physical environments. Among these data standards, each satisfies the unique needs of a specific application, and numerous studies regarding topology generation have been proposed. Khan et al. [148] performed an automatic geometric, semantic, and topological transformation from an existing semantic 3D building model (e.g., IFC or CityGML LoD4) to IndoorGML. Mirvahabi and Abbaspour [149] derived the IndoorGML data file from the OpenStreetMap file. Referring to IndoorGML, Zhu et al. [150] defined Indoor Multi-Dimensional Location GML mainly for indoor location and navigation. Teo and Yu [151] converted the topological relationships from BIM into an indoor network model stored in IndoorGML structures. Srivastava et al. [152] extracted relevant geometric entities and their topological relationships from CAD drawings and added semantic information to extend the existing IndoorGML. Tessema et al. [153] extracted both semantic and topological information from an occupancy grid map and translated this extracted information into a semantic node-relation graph stored in IndoorGML format. Motivated by a museum case study, Kontarinis et al. [154] established a semantic indoor trajectory model by combining the existing semantic outdoor trajectory models and the semantically rich hierarchical indoor space symbolic model. Flikweert et al. [155] automatically extracted a navigation graph from trajectory-based mobile point clouds, where connected spaces, walkable spaces, rooms, and corridors were identified and organized, and stored them in an IndoorGML format. Mortari et al. [156] proposed a navigation model using geometric information from CityGML and introduced semantic elements (e.g., openings) into the establishing indoor navigation model for meeting human needs. Starting from the raw point clouds,

Nikoohemat et al. [157] developed a 3D indoor reconstruction method, which also involved obstacles, to perform navigation and path planning.

Although the existing data standards define, organize, manage, and store detailed indoor spatial information to support different indoor applications, more efforts are needed to discover and implement their potential aspects and to further complement their current versions, which might not be explicitly explained in the standard document [158]. For instance, IndoorGML version 1 was developed to satisfy the requirements of indoor navigation applications (e.g., indoor location-based services and routing services), but other requirements (e.g., facility management) are expected to be handled in its future versions [145]. Furthermore, there have been some combinations to complement each other for satisfying the specific requirements of different indoor applications. As complementary to one another, Kim et al. [159] generated IndoorGML data from CityGML LoD4 and offered external references from IndoorGML to an object in CityGML LoD4. Liu et al. [160] integrated IndoorGML with IndoorLocationGML, where IndoorGML offers the subdivision of indoor spaces, while the location semantics in IndoorLocationGML can be introduced. To address the limitations of the standardized level of detail (LoD), Tang et al. [161] extended an indoor LoD specification using both IFC and IndoorGML to achieve a full LoD specification for 3D building models. Zeng and Kang [162] extracted navigational elements from RGB-D data and encoded them under the framework of IndoorGML. Alattas et al. [163] considered the access rights of indoor spaces and developed a conceptual model that combines the Land Administration Domain Model (LADM) and IndoorGML to define the rights, restrictions, and responsibilities of users. Moreover, some recommendations have been discussed with the intent of improving the future standard. Diakité et al. [164] pointed out that several concepts (e.g., cell sub-spacing) in the current version of IndoorGML should be improved, and proposed several criteria to automatically perform indoor sub-division processes.

5. Trends and Challenges

In recent years, the 3D reconstruction of indoor environments has become increasingly popular for the development of many indoor applications, such as navigation guidance, emergency management, building maintenance, and renovation planning, as well as a range of indoor location-based services (e.g., way-finding and contextualized content delivery). Therefore, this literature review provides a summary related to the state-of-the-art techniques for 3D reconstruction of indoor environments. With the ongoing development of research, the challenges and trends of the current techniques are discussed in this section.

5.1. Non-Manhattan Assumption

The major limitation in the existing methods is generally related to the strong Manhattan assumption [92,94,96,97,102–104,109], where the main structures of buildings are assumed to be rectangular and to intersect orthogonally. However, indoor spaces with complex geometric structures (e.g., cylindrical walls, spherical ceilings, L-shaped layout, or other non-planar structures [20,95,165–167]) occur in current indoor environments. Thus, recovering a complete 3D indoor model with arbitrary geometric shapes is an interesting topic for future study [95,168].

5.2. Multi-Task Collaborative Optimization

The objective of a single task usually fails to generate enough information for robustly achieving the ideal requirements. Multi-task collaborative optimization offers redundant and complementary information from different perspectives to assist task completion. For example, for collaboratively estimating room layout and for detecting 3D objects [106,169], room layout estimation provides information of spatial constraints for object detection, while object detection also offers occlusion information for room layout estimation, thus complementing one another. Another example is to jointly optimize the room layout and the semantic segmentation task [109]. Moreover, multi-task collaborative optimization can also achieve shared weights in designed network architectures for

reducing training and computational costs [170,171]. Thus, the integration of multi-task collaborative optimization and deep-learning techniques will also be a future focus for indoor environment modeling and applications.

5.3. Indoor Scene Understanding by Combining Both Spatial and Temporal Consistencies

Previous work mainly focuses on modeling the spatial information or the spatial context based on a single RGB(-D) image, where spatial consistency has been widely used as an inherent constraint in indoor environment modeling and applications. However, little attention has been paid to the consistency of the sequences of images [172]. Therefore, temporal consistency of the sequences in many indoor modeling and real-world applications is an issue that needs to be resolved [127,173–179] since, for complex dynamic scenes, moving objects exhibit similar shapes and appearances, both spatially and temporally [180]. Hence, the combination of both spatial and temporal consistencies will also be important in order to improve the robustness of indoor scene understanding (e.g., semantic segmentation and object detection).

5.4. Automatic Reconstruction of Indoor Models with Different Levels of Detail

Until now, the structural modeling of buildings has only focused on the main structural elements (e.g., walls, floors, ceilings, and openings) [14,76,79,83,181,182]. To facilitate a range of location-based services, a more detailed 3D indoor model with interior furniture (specifically obstacles) should be defined and established for fine-grained applications [183], where its space structures are described and the interior furniture is also modeled [10,184]. Although the interior objects are involved in LoD4, it is obvious that the current LoD definition might be insufficient to meet the needs for the rising variety of applications, which may require interior structures with different accuracy levels. In future work, more geometric features should be explored in order to produce more robust indoor spatial information models with accurate geometry and rich semantics.

5.5. Indoor–Outdoor Space Seamless Integration

To enable people to move seamlessly between buildings and surrounding areas, a full representation of both interior and exterior buildings is necessary to offer support in the indoor–outdoor context [161,185]. However, this is challenging due to the differences in their physical structures, spatial relationships between entities, and so on [186]. For example, there are few visual correspondences that are part of both interior and exterior models for their alignment [111]. Moreover, it is also inaccurate to perform a complete reconstruction for capturing the entire scene from the outside to inside because of drifts or a lack of matchable features in most cases [111]. Thus, robustly stitching indoor and outdoor models together with no/few visual overlaps has become an active topic.

6. Conclusions

This paper conducted a comprehensive review in order to analyze and summarize the 3D reconstruction of indoor environments. Because of the importance of the development of 3D reconstruction techniques for indoor environments, we first provided a brief description of the available benchmark datasets that can be used to evaluate the performance of different algorithms, as well as the data-collection methods regarding 3D indoor spaces. Since indoor scene reconstruction is far from being satisfactory in terms of accuracy and precision, an increasing number of datasets are being released as benchmarks to enable comparisons among different reconstruction methods (including geometric, semantic, and topological) and to promote international state-of-the-art research. Meanwhile, with the development of sensor technology, the use of a variety of sensors, e.g., cameras (monocular, stereo, video, or panoramic), laser scanners, and depth cameras, for dealing with data collection tasks and the SLAM-based framework, especially using monocular vision due to its low cost, is becoming mainstream for collecting data.

Therefore, this review offered a detailed analysis and summary of the existing studies related to the 3D reconstruction of indoor environments, which can be divided into different categories according to their inherent principles and application requirements. Rather than offering a quantitative analysis among different methods, we mainly focused on introducing the relevant theories (e.g., geometric, semantic, and topological reconstruction of indoor environments) and their advantages and disadvantages. With regard to geometric reconstruction, most of the existing methods are based on the strong Manhattan assumption and mainly focus on the reconstruction of main architectural structures. The reconstruction of interior furniture and the seamless modeling of indoor and outdoor spaces are still in their infancy. Moreover, although the cluttered and occluded characteristics of indoor environments still pose great challenges to semantic modeling, deep-learning techniques and strong data operational capability, especially based on the use of a GPU, can improve the generalization of learnable models to obtain better interpretation results. In addition, the hierarchical pyramid structures of deep-learning architectures enable multi-task collaborative schemes to jointly optimize each other using redundant and complementary information from different perspectives, which demonstrates their potential for use in the 3D reconstruction of indoor environments. In addition, the occurrence of spatial data standards (e.g., IFC, CityGML, and IndoorGML) is beneficial for defining, organizing, and storing detailed indoor physical environments, and satisfies the unique needs of a specific application. Although the existing data standards define, organize, manage, and store detailed indoor spatial information to support different indoor applications, more efforts are needed to discover and implement their potential aspects and to further complement their current versions, which might not be explicitly explained in the standard document.

Finally, the challenges and trends faced by the current techniques, such as indoor scene reconstruction without the strong Manhattan assumption, multi-task collaborative optimization to complement each other, scene understanding by combining both spatial and temporal consistency, automatic reconstruction of indoor models with different levels of detail, and indoor–outdoor space seamless integration, were summarized and discussed.

Author Contributions: Zhizhong Kang proposed and refined the major framework for the review. Juntao Yangand, Zhizhong Kang organized the literature and wrote the manuscript. Zhou Yang and Sai Cheng assisted in the collection of the literature. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Natural Science Foundation of China [No. 41471360]. And the APC was funded by the National Natural Science Foundation of China.

Acknowledgments: Sincere thanks are given for the comments and contributions of the anonymous reviewers and the members of the editorial team.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. US Environmental Protection Agency. *Buildings and Their Impact on the Environment: A Statistical Summary*; US Environmental Protection Agency Green Building Workgroup: Washington, DC, USA, 2009.
2. Dasgupta, S.; Fang, K.; Chen, K.; Savarese, S. Delay: Robust spatial layout estimation for cluttered indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 616–624.
3. Husain, F.; Schulz, H.; Dellen, B.; Torras, C.; Behnke, S. Combining semantic and geometric features for object class segmentation of indoor scenes. *IEEE Robot. Autom. Lett.* **2016**, *2*, 49–55. [CrossRef]
4. Sequeira, V.; Gonçalves, J.G.; Ribeiro, M.I. 3D reconstruction of indoor environments. In Proceedings of the 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 19 September 1996; pp. 405–408.
5. Isikdag, U.; Zlatanova, S.; Underwood, J. A BIM-Oriented Model for supporting indoor navigation requirements. *Comput. Environ. Urban Syst.* **2013**, *41*, 112–123. [CrossRef]

6. Ahmed, A.A.; Al-Shaboti, M.; Al-Zubairi, A. An indoor emergency guidance algorithm based on wireless sensor networks. In Proceedings of the 2015 International Conference on Cloud Computing (ICCC), Riyadh, Saudi Arabia, 26–29 April 2015; pp. 1–5.
7. Chen, C.; Tang, L. BIM-based integrated management workflow design for schedule and cost planning of building fabric maintenance. *Autom. Constr.* **2019**, *107*, 102944. [CrossRef]
8. Tian, X.; Shen, R.; Liu, D.; Wen, Y.; Wang, X. Performance analysis of RSS fingerprinting based indoor localization. *IEEE Trans. Mob. Comput.* **2016**, *16*, 2847–2861. [CrossRef]
9. Chen, K.; Lai, Y.-K.; Hu, S.-M. 3D indoor scene modeling from RGB-D data: A survey. *Comput. Vis. Media* **2015**, *1*, 267–278. [CrossRef]
10. Zhang, Y.; Liu, Z.; Miao, Z.; Wu, W.; Liu, K.; Sun, Z. Single image-based data-driven indoor scene modeling. *Comput. Graph.* **2015**, *53*, 210–223. [CrossRef]
11. Engel, J.; Stückler, J.; Cremers, D. Large-scale direct SLAM with stereo cameras. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 1935–1942.
12. Leiva, J.; Martinez, P.; Perez, E.; Urdiales, C.; Sandoval, F. 3D reconstruction of static indoor environment by fusion of sonar and video data. In Proceedings of the International Symposium on Intelligent Robotic Systems, Toulouse, France, 18–20 July 2001.
13. Yang, H.; Zhang, H. Modeling room structure from indoor panorama. In Proceedings of the 13th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry, Shenzhen, China, 30 November–2 December 2014; pp. 47–55.
14. Wang, C.; Hou, S.; Wen, C.; Gong, Z.; Li, Q.; Sun, X.; Li, J. Semantic line framework-based indoor building modeling using backpacked laser scanning point cloud. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 150–166. [CrossRef]
15. Bokaris, P.-A.; Muselet, D.; Trémeau, A. 3D reconstruction of indoor scenes using a single RGB-D image. In Proceedings of the 12th International Conference on Computer Vision Theory and Applications (VISAPP 2017), Porto, Portugal, 27 February–1 March 2017.
16. Valentin, J.P.; Sengupta, S.; Warrell, J.; Shahrokni, A.; Torr, P.H. Mesh based semantic modelling for indoor and outdoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2067–2074.
17. Jung, J.; Hong, S.; Yoon, S.; Kim, J.; Heo, J. Automated 3D wireframe modeling of indoor structures from point clouds using constrained least-squares adjustment for as-built BIM. *J. Comput. Civ. Eng.* **2015**, *30*, 04015074. [CrossRef]
18. Shao, T.; Xu, W.; Zhou, K.; Wang, J.; Li, D.; Guo, B. An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *ACM Trans. Graph. TOG* **2012**, *31*, 136. [CrossRef]
19. Ochmann, S.; Vock, R.; Wessel, R.; Klein, R. Automatic reconstruction of parametric building models from indoor point clouds. *Comput. Graph.* **2016**, *54*, 94–103. [CrossRef]
20. Yang, F.; Zhou, G.; Su, F.; Zuo, X.; Tang, L.; Liang, Y.; Zhu, H.; Li, L. Automatic Indoor Reconstruction from Point Clouds in Multi-room Environments with Curved Walls. *Sensors* **2019**, *19*, 3798. [CrossRef] [PubMed]
21. Froese, T.; Grobler, F.; Ritzenthaler, J.; Yu, K.; Akinci, B.; Akbas, R.; Koo, B.; Barron, A.; Kunz, J.C. Industry Foundation Classes for Project Management-A Trial Implementation. *ITcon* **1999**, *4*, 17–36.
22. Gröger, G.; Kolbe, T.H.; Nagel, C.; Häfele, K.-H. *OGC City Geography Markup Language (CityGML) Encoding Standard*; Open Geospatial Consortium, 2012. Available online: http://www.opengis.net/spec/citygml/2.0 (accessed on 17 May 2020).
23. Naseer, M.; Khan, S.; Porikli, F. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE Access* **2018**, *7*, 1859–1887. [CrossRef]
24. Li, Y.; Dai, A.; Guibas, L.; Nießner, M. Database-assisted object retrieval for real-time 3d reconstruction. *Computer Graph. Forum* **2015**, *34*, 435–446. [CrossRef]
25. Schwing, A.G.; Hazan, T.; Pollefeys, M.; Urtasun, R. Efficient structured prediction for 3d indoor scene understanding. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2815–2822.
26. Handa, A.; Whelan, T.; McDonald, J.; Davison, A.J. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In Proceedings of the 2014 IEEE international conference on Robotics and automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 1524–1531.

27. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; pp. 573–580.
28. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Res.* **2016**, *35*, 1157–1163. [CrossRef]
29. Wang, C.; Dai, Y.; El-Sheimy, N.; Wen, C.; Retscher, G.; Kang, Z.; Lingua, A. Progress on Isprs Benchmark on Multisensory Indoor Mapping and Positioning. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-2/W13*, 1709–1713. [CrossRef]
30. Khoshelham, K.; Vilariño, L.D.; Peter, M.; Kang, Z.; Acharya, D. The Isprs Benchmark on Indoor Modelling. *Int. Arch. Photogramm. Remote Sen. Spat. Inf. Sci.* **2017**, *42*, 367–372. [CrossRef]
31. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576.
32. 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, Barcelona, Spain, 6–13 November 2011; pp. 5828–5839.
33. Silberman, N.; Fergus, R. Indoor scene segmentation using a structured light sensor. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 601–608.
34. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012.
35. Armeni, I.; Sax, S.; Zamir, A.R.; Savarese, S. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv* **2017**, arXiv:1702.01105.
36. Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv* **2017**, arXiv:1709.06158.
37. Marck, J.W.; Mohamoud, A.; vd Houwen, E.; van Heijster, R. Indoor radar SLAM A radar application for vision and GPS denied environments. In Proceedings of the 2013 European Radar Conference, Nuremberg, Germany, 9–11 October 2013; pp. 471–474.
38. van Dijk, T.; de Croon, G.C. How Do Neural Networks See Depth in Single Images? In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 2183–2191.
39. Liu, M.; Salzmann, M.; He, X. Discrete-continuous depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 716–723.
40. Zhuo, W.; Salzmann, M.; He, X.; Liu, M. Indoor scene structure analysis for single image depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 614–622.
41. Eder, M.; Moulon, P.; Guan, L. Pano Popups: Indoor 3D Reconstruction with a Plane-Aware Network. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Québec, QC, Canada, 16–19 September 2019; pp. 76–84.
42. Roy, A.; Todorovic, S. Monocular depth estimation using neural regression forest. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5506–5514.
43. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5162–5170.
44. Goldlücke, B.; Aubry, M.; Kolev, K.; Cremers, D. A super-resolution framework for high-accuracy multiview reconstruction. *Int. J. Comput. Vis.* **2014**, *106*, 172–191. [CrossRef]
45. Collins, R.T. A space-sweep approach to true multi-image matching. In Proceedings of the CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 18–20 June 1996; pp. 358–363.
46. Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1362–1376. [CrossRef] [PubMed]

47. Galliani, S.; Lasinger, K.; Schindler, K. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 873–881.
48. Langguth, F.; Sunkavalli, K.; Hadap, S.; Goesele, M. Shading-aware multi-view stereo. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
49. Häne, C.; Zach, C.; Cohen, A.; Pollefeys, M. Dense semantic 3d reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1730–1743. [CrossRef] [PubMed]
50. Ullman, S. The interpretation of structure from motion. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **1979**, *203*, 405–426.
51. Hartmann, W.; Galliani, S.; Havlena, M.; Van Gool, L.; Schindler, K. Learned multi-patch similarity. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1586–1594.
52. Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; Fang, L. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2307–2315.
53. Huang, P.-H.; Matzen, K.; Kopf, J.; Ahuja, N.; Huang, J.-B. Deepmvs: Learning multi-view stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2821–2830.
54. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 767–783.
55. Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5525–5534.
56. Bailey, T.; Durrant-Whyte, H. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robot. Autom. Mag.* **2006**, *13*, 108–117. [CrossRef]
57. Aouina, A.; Devy, M.; Hernandez, A.M. 3d modeling with a moving tilting laser sensor for indoor environments. *IFAC Proc. Vol.* **2014**, *47*, 7604–7609. [CrossRef]
58. Salas-Moreno, R.F.; Newcombe, R.A.; Strasdat, H.; Kelly, P.H.; 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, Portland, OR, USA, 23–28 June 2013; pp. 1352–1359.
59. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef]
60. Schulz, V.H.; Bombardelli, F.G.; Todt, E. A SoC with FPGA Landmark Acquisition System for Binocular Visual SLAM. In Proceedings of the 2015 12th Latin American Robotics Symposium and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR), Uberlândia, Brazil, 28 October 28–1 November 2015; pp. 336–341.
61. Leonard, J.J.; Durrant-Whyte, H.F. Mobile robot localization by tracking geometric beacons. *IEEE Trans. Robot. Autom.* **1991**, *7*, 376–382. [CrossRef]
62. Gomez-Ojeda, R.; Briales, J.; Gonzalez-Jimenez, J. PL-SVO: Semi-direct Monocular Visual Odometry by combining points and line segments. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 4211–4216.
63. Pumarola, A.; Vakhitov, A.; Agudo, A.; Sanfeliu, A.; Moreno-Noguer, F. PL-SLAM: Real-time monocular visual SLAM with points and lines. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 4503–4508.
64. Wang, R.; Di, K.; Wan, W.; Wang, Y. Improved Point-Line Feature Based Visual SLAM Method for Indoor Scenes. *Sensors* **2018**, *18*, 3559. [CrossRef]
65. Bowman, S.L.; Atanasov, N.; Daniilidis, K.; Pappas, G.J. Probabilistic data association for semantic slam. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1722–1729.
66. So, C.; Baciu, G.; Sun, H. Reconstruction of 3D virtual buildings from 2D architectural floor plans. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, Taipei, Taiwan, 2–5 November 1998; pp. 17–23.
67. Lu, T.; Tai, C.-L.; Bao, L.; Su, F.; Cai, S. 3D reconstruction of detailed buildings from architectural drawings. *Comput. Aided Des. Appl.* **2005**, *2*, 527–536. [CrossRef]

68. Lee, S.; Feng, D.; Grimm, C.; Gooch, B. A Sketch-Based User Interface for Reconstructing Architectural Drawings. *Comput. Graph. Forum* **2008**, *27*, 81–90. [CrossRef]
69. Horna, S.; Meneveaux, D.; Damiand, G.; Bertrand, Y. Consistency constraints and 3D building reconstruction. *Comput. Aided Des.* **2009**, *41*, 13–27. [CrossRef]
70. Li, T.; Shu, B.; Qiu, X.; Wang, Z. Efficient reconstruction from architectural drawings. *Int. J. Comput. Appl. Technol.* **2010**, *38*, 177–184. [CrossRef]
71. Yin, X.; Wonka, P.; Razdan, A. Generating 3d building models from architectural drawings: A survey. *IEEE Comput. Graph. Appl.* **2008**, *29*, 20–30. [CrossRef] [PubMed]
72. Ning, X.; Ma, J.; Lv, Z.; Xu, Q.; Wang, Y. Structure Reconstruction of Indoor Scene from Terrestrial Laser Scanner. In *International Conference on E-Learning and Games*; Springer: Berlin/Heidelberg, Germany, 2018.
73. Edelsbrunner, H. Alpha shapes—A survey. *Tessellations Sci.* **2010**, *27*, 1–25.
74. Previtali, M.; Díaz-Vilariño, L.; Scaioni, M. Indoor building reconstruction from occluded point clouds using graph-cut and ray-tracing. *Appl. Sci.* **2018**, *8*, 1529. [CrossRef]
75. Kang, Z.; Zhong, R.; Wu, A.; Shi, Z.; Luo, Z. An efficient planar feature fitting method using point cloud simplification and threshold-independent BaySAC. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1842–1846. [CrossRef]
76. Jung, J.; Hong, S.; Jeong, S.; Kim, S.; Cho, H.; Hong, S.; Heo, J. Productive modeling for development of as-built BIM of existing indoor structures. *Autom. Constr.* **2014**, *42*, 68–77. [CrossRef]
77. Wang, C.; Cho, Y.K.; Kim, C. Automatic BIM component extraction from point clouds of existing buildings for sustainability applications. *Autom. Constr.* **2015**, *56*, 1–13. [CrossRef]
78. Shi, W.; Ahmed, W.; Li, N.; Fan, W.; Xiang, H.; Wang, M. Semantic Geometric Modelling of Unstructured Indoor Point Cloud. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 9. [CrossRef]
79. Hong, S.; Jung, J.; Kim, S.; Cho, H.; Lee, J.; Heo, J. Semi-automated approach to indoor mapping for 3D as-built building information modeling. *Comput. Environ. Urban Syst.* **2015**, *51*, 34–46. [CrossRef]
80. Michailidis, G.-T.; Pajarola, R. Bayesian graph-cut optimization for wall surfaces reconstruction in indoor environments. *Vis. Comput.* **2017**, *33*, 1347–1355. [CrossRef]
81. Mura, C.; Mattausch, O.; Villanueva, A.J.; Gobbetti, E.; Pajarola, R. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Comput. Graph.* **2014**, *44*, 20–32. [CrossRef]
82. Tang, S.; Zhang, Y.; Li, Y.; Yuan, Z.; Wang, Y.; Zhang, X.; Li, X.; Zhang, Y.; Guo, R.; Wang, W. Fast and Automatic Reconstruction of Semantically Rich 3D Indoor Maps from Low-quality RGB-D Sequences. *Sensors* **2019**, *19*, 533. [CrossRef]
83. Wang, R.; Xie, L.; Chen, D. Modeling indoor spaces using decomposition and reconstruction of structural elements. *Photogramm. Eng. Remote Sens.* **2017**, *83*, 827–841. [CrossRef]
84. Li, L.; Su, F.; Yang, F.; Zhu, H.; Li, D.; Zuo, X.; Li, F.; Liu, Y.; Ying, S. Reconstruction of Three-Dimensional (3D) Indoor Interiors with Multiple Stories via Comprehensive Segmentation. *Remote Sens.* **2018**, *10*, 1281. [CrossRef]
85. Pang, Y.; Zhang, C.; Zhou, L.; Lin, B.; Lv, G. Extracting Indoor Space Information in Complex Building Environments. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 321. [CrossRef]
86. Chen, J.; Chen, B. Architectural modeling from sparsely scanned range data. *Int. J. Comput. Vis.* **2008**, *78*, 223–236. [CrossRef]
87. Chen, K.; Lai, Y.; Wu, Y.-X.; Martin, R.R.; Hu, S.-M. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. *ACM Trans. Graph.* **2014**, *33*, 208. [CrossRef]
88. Liu, Z.; Zhang, Y.; Wu, W.; Liu, K.; Sun, Z. Model-driven indoor scenes modeling from a single image. In Proceedings of the 41st Graphics Interface Conference, Halifax, NS, Canada, 3–5 June 2015.
89. Tran, H.; Khoshelham, K.; Kealy, A. Geometric comparison and quality evaluation of 3D models of indoor environments. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 29–39. [CrossRef]
90. Chen, J.; Shao, J.; Zhang, D.; Wu, X. A Fast End-to-End Method with Style Transfer for Room Layout Estimation. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 964–969.
91. Fernandez-Labrador, C.; Facil, J.M.; Perez-Yus, A.; Demonceaux, C.; Civera, J.; Guerrero, J.J. Corners for Layout: End-to-End Layout Recovery from 360 Images. *arXiv* **2019**, arXiv:1903.08094. [CrossRef]

92. Lin, H.J.; Huang, S.W.; Lai, S.H.; Chiang, C.K. Indoor scene layout estimation from a single image. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
93. Zhang, W.; Zhang, W.; Liu, K.; Gu, J. Learning to predict high-quality edge maps for room layout estimation. *IEEE Trans. Multimed.* **2016**, *19*, 935–943. [CrossRef]
94. Lee, C.-Y.; Badrinarayanan, V.; Malisiewicz, T.; Rabinovich, A. Roomnet: End-to-end room layout estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4865–4874.
95. Zou, C.; Colburn, A.; Shan, Q.; Doiem, D. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 18–23 June 2018.
96. Hedau, V.; Hoiem, D.; Forsyth, D. Recovering the spatial layout of cluttered rooms. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009.
97. Chao, Y.-W.; Choi, W.; Pantofaru, C.; Savarese, S. Layout estimation of highly cluttered indoor scenes using geometric and semantic cues. In *International Conference on Image Analysis and Processing*; Springer: Berlin/Heidelberg, Germany, 2013.
98. Park, S.-J.; Hong, K.-S. Recovering an indoor 3D layout with top-down semantic segmentation from a single image. *Pattern Recognit. Lett.* **2015**, *68*, 70–75. [CrossRef]
99. Mallya, A.; Lazebnik, S. Learning informative edge maps for indoor scene layout prediction. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 936–944.
100. Jahromi, A.B.; Sohn, G. Edge Based 3d Indoor Corridor Modeling Using a Single Image. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*.
101. Hirzer, M.; Roth, P.M.; Lepetit, V. Smart Hypothesis Generation for Efficient and Robust Room Layout Estimation. *arXiv* **2019**, arXiv:1910.12257.
102. Chang, H.-C.; Huang, S.-H.; Lai, S.-H. Using line consistency to estimate 3D indoor Manhattan scene layout from a single image. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec, QC, Canada, 27–30 September 2015; pp. 4723–4727.
103. Kruzhilov, I.; Romanov, M.; Konushin, A. Double Refinement Network for Room Layout Estimation. In *Asian Conference on Pattern Recognition*; Springer: Cham, Switzerland, 2019.
104. Gupta, A.; Hebert, M.; Kanade, T.; Blei, D.M. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2010; pp. 1288–1296. Available online: http://papers.nips.cc/paper/4120-estimating-spatial-layout-of-rooms-using-volumetric-reasoning-about-objects-and-surfaces.pdf (accessed on 17 May 2020).
105. Del Pero, L.; Bowdish, J.; Fried, D.; Kermgard, B.; Hartley, E.; Barnard, K. Bayesian geometric modeling of indoor scenes. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2719–2726.
106. Schwing, A.G.; Fidler, S.; Pollefeys, M.; Urtasun, R. Box in the box: Joint 3d layout and object reasoning from single images. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 353–360.
107. Zhang, J.; Kan, C.; Schwing, A.G.; Urtasun, R. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1273–1280.
108. Bao, S.Y.; Furlan, A.; Fei-Fei, L.; Savarese, S. Understanding the 3D layout of a cluttered room from multiple images. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, 24–26 March 2014; pp. 690–697.
109. Zhang, W.; Zhang, W.; Gu, J. Edge-semantic learning strategy for layout estimation in indoor environment. *IEEE Trans. Cybern.* **2020**, *50*, 2730–2739. [CrossRef] [PubMed]
110. Cohen, A.; Schönberger, J.L.; Speciale, P.; Sattler, T.; Frahm, J.-M.; Pollefeys, M. Indoor-outdoor 3d reconstruction alignment. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 285–300.

111. Koch, T.; Korner, M.; Fraundorfer, F. Automatic alignment of indoor and outdoor building models using 3D line segments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016.
112. Khan, S.H.; Bennamoun, M.; Sohel, F.; Togneri, R. Geometry driven semantic labeling of indoor scenes. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014.
113. Ren, X.; Bo, L.; Fox, D. Rgb-(d) scene labeling: Features and algorithms. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2759–2766.
114. Gupta, S.; Arbelaez, P.; Malik, J. Perceptual organization and recognition of indoor scenes from RGB-D images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 564–571.
115. Müller, A.C.; Behnke, S. Learning depth-sensitive conditional random fields for semantic segmentation of RGB-D images. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 6232–6237.
116. Deng, Z.; Todorovic, S.; Jan Latecki, L. Semantic segmentation of rgbd images with mutex constraints. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 1733–1741.
117. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef] [PubMed]
118. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
119. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014.
120. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
121. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
122. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
123. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-transformed points. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 820–830.
124. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
125. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
126. Jiang, J.; Zhang, Z.; Huang, Y.; Zheng, L. Incorporating depth into both cnn and crf for indoor semantic segmentation. In Proceedings of the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 24–26 November 2017; pp. 525–530.
127. Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; Huang, K. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3029–3037.
128. Lin, D.; Chen, G.; Cohen-Or, D.; Heng, P.-A.; Huang, H. Cascaded feature network for semantic segmentation of RGB-D images. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 1311–1319.
129. Li, Y.; Zhang, J.; Cheng, Y.; Huang, K.; Tan, T. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1262–1266.

130. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv* **2018**, arXiv:1806.01054.
131. Guo, Y.; Chen, T. Semantic segmentation of RGBD images based on deep depth regression. *Pattern Recognit. Lett.* **2018**, *109*, 55–64. [CrossRef]
132. Kim, Y.; Kang, H.; Lee, J. Development of indoor spatial data model using CityGML ADE. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *40*, 41–45. [CrossRef]
133. Jamali, A.; Rahman, A.A.; Boguslawski, P.; Kumar, P.; Gold, C.M. An automated 3D modeling of topological indoor navigation network. *GeoJournal* **2017**, *82*, 157–170. [CrossRef]
134. Tran, H.; Khoshelham, K.; Kealy, A.; Díaz-Vilariño, L. Extracting topological relations between indoor spaces from point clouds. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 401. [CrossRef]
135. Sarda, N. Development of navigational structure for buildings from their valid 3D CityGML models. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 3341–3346.
136. Thrun, S.; Bücken, A. Integrating grid-based and topological maps for mobile robot navigation. In Proceedings of the National Conference on Artificial Intelligence, Portland, OR, USA, 4–6 August 1996.
137. Joo, K.; Lee, T.-K.; Baek, S.; Oh, S.-Y. Generating topological map from occupancy grid-map using virtual door detection. In Proceedings of the IEEE Congress on Evolutionary Computation, Barcelona, Spain, 18–23 July 2010.
138. Portugal, D.; Rocha, R.P. Extracting Topological Information from Grid Maps for Robot Navigation. In Proceedings of the 4th International Conference on Agents and Artificial Intelligence (ICAART-2012), Algarve, Portugal, 6–8 February 2012; pp. 137–143. [CrossRef]
139. Yang, L.; Worboys, M. Generation of navigation graphs for indoor space. *Int. J. Geograph. Inf. Sci.* **2015**, *29*, 1737–1756. [CrossRef]
140. Sithole, G. Indoor Space Routing Graphs: Visibility, Encoding, Encryption and Attenuation. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *XLII-4*, 579–585. [CrossRef]
141. Demyen, D.; Buro, M. Efficient triangulation-based pathfinding. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 06), Boston, MA, USA, 16–20 July 2006.
142. Li, X.; Claramunt, C.; Ray, C. A grid graph-based model for the analysis of 2D indoor spaces. *Comput. Environ. Urban Syst.* **2010**, *34*, 532–540. [CrossRef]
143. Gonzalez-Arjona, D.; Sanchez, A.; López-Colino, F.; De Castro, A.; Garrido, J. Simplified occupancy grid indoor mapping optimized for low-cost robots. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 959–977. [CrossRef]
144. Xu, M.; Wei, S.; Zlatanova, S.; Zhang, R. BIM-based indoor path planning considering obstacles. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, 417–423. [CrossRef]
145. Li, D.; Lee, D.L. A lattice-based semantic location model for indoor navigation. In Proceedings of the Ninth International Conference on Mobile Data Management (mdm 2008), Beijing, China, 27–30 April 2008; pp. 17–24.
146. Lin, Z.; Xu, Z.; Hu, D.; Hu, Q.; Li, W. Hybrid spatial data model for indoor space: Combined topology and grid. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 343. [CrossRef]
147. Lee, J.; Li, K.; Zlatanova, S.; Kolbe, T.; Nagel, C.; Becker, T. OGC IndoorGML–with Corrigendum. 2016. Available online: http://www.opengis.net/doc/IS/indoorgml/1.0 (accessed on 17 May 2020).
148. Khan, A.; Donaubauer, A.; Kolbe, T.H. A multi-step transformation process for automatically generating indoor routing graphs from semantic 3D building models. In Proceedings of the 9th 3D GeoInfo Conference, Dubai, UAE, 11–13 Novermber 2014.
149. Mirvahabi, S.; Abbaspour, R.A. Automatic extraction of IndoorGML core model from OpenStreetMap. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *40*, 459. [CrossRef]
150. Zhu, Q.; Li, Y.; Xiong, Q.; Zlatanova, S.; Ding, Y.; Zhang, Y.; Zhou, Y. Indoor multi-dimensional location gml and its application for ubiquitous indoor location services. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 220. [CrossRef]
151. Teo, T.-A.; Yu, S.-C. The Extraction of Indoor Building Information from Bim to Ogc Indoorgml. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 167–170. [CrossRef]
152. Srivastavaa, S.; Maheshwarib, N.; Rajanc, K. Towards Generating Semantically-Rich Indoorgml Data from Architectural Plans. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 4. [CrossRef]

153. Tessema, L.S.; Jäger, R.; Stilla, U. Extraction of IndoorGML Model from an Occupancy Grid Map Constructed Using 2D LiDAR. In Proceedings of the German Society for Photogrammetry, Remote Sensing and Geoinformation, 39st Conference, Vienna, Austria, February 2020; Available online: https://www.researchgate.net/publication/338690496_Extraction_of_IndoorGML_Model_from_an_Occupancy_Grid_Map_Constructed_Using_2D_LiDAR (accessed on 17 May 2020).
154. Kontarinis, A.; Zeitouni, K.; Marinica, C.; Vodislav, D.; Kotzinos, D. Towards a Semantic Indoor Trajectory Model. In Proceedings of the 2nd International Workshop on "Big Mobility Data Analytics" (BMDA) with EDBT, Lisbon, Portugal, 26 March 2019.
155. Flikweert, P.; Peters, R.; Díaz-Vilariño, L.; Voûte, R.; Staats, B. Automatic Extraction of a Navigation Graph Intended for Indoorgml from an Indoor Point Cloud. *ISPRS Ann. Photogramm. Remote Sens. Spat.Inf. Sci.* **2019**, *4*, 271–278. [CrossRef]
156. Mortari, F.; Clementini, E.; Zlatanova, S.; Liu, L. An indoor navigation model and its network extraction. *Appl. Geomat.* **2019**, *11*, 413–427. [CrossRef]
157. Nikoohemat, S.; Diakité, A.; Zlatanova, S.; Vosselman, G. Indoor 3d Modeling and Flexible Space Subdivision from Point Clouds. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *4*, 285–292. [CrossRef]
158. Kang, H.-K.; Li, K.-J. A standard indoor spatial data model—OGC IndoorGML and implementation approaches. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 116. [CrossRef]
159. Kim, J.-S.; Yoo, S.-J.; Li, K.-J. Integrating IndoorGML and CityGML for indoor space. In *International Symposium on Web and Wireless Geographical Information Systems*; Springer: Berlin/Heidelberg, Germany, 2014.
160. Liu, L.; Zlatanova, S.; Zhu, Q.; Li, K. Towards the integration of IndoorGML and IndoorlocationGML for indoor applications. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 343. [CrossRef]
161. Tang, L.; Li, L.; Ying, S.; Lei, Y. A Full Level-of-Detail Specification for 3D Building Models Combining Indoor and Outdoor Scenes. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 419. [CrossRef]
162. Zeng, L.; Kang, Z. Automatic Recognition of Indoor Navigation Elements from Kinect Point Clouds. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 431–437. [CrossRef]
163. Alattas, A.; Zlatanova, S.; Van Oosterom, P.; Chatzinikolaou, E.; Lemmen, C.; Li, K.-J. Supporting indoor navigation using access rights to spaces based on combined use of IndoorGML and LADM models. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 384. [CrossRef]
164. Diakité, A.A.; Zlatanov, S.; Li, K.-J. About the Subdivision of Indoor Spaces in Indoorgml. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 41–48. [CrossRef]
165. Mura, C.; Villanueva, A.J.; Mattausch, O.; Gobbetti, E.; Pajarola, R. Reconstructing Complex Indoor Environments with Arbitrary Wall Orientations. *Eurograph. Posters* **2014**, *19*, 38–40.
166. Nakagawa, M.; Kataoka, K.; Yamamoto, T.; Shiozaki, M.; Ohhashi, T. Panoramic Rendering-Based Polygon Extraction from Indoor Mobile Lidar Data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *2*, 181–186. [CrossRef]
167. Mura, C.; Mattausch, O.; Pajarola, R. Piecewise-planar Reconstruction of Multi-room Interiors with Arbitrary Wall Arrangements. *Comput. Graph. Forum* **2016**, *35*, 179–188. [CrossRef]
168. Hsiao, C.-W.; Sun, C.; Sun, M.; Chen, H.-T. Flat2Layout: Flat Representation for Estimating Layout of General Room Types. *arXiv* **2019**, arXiv:1905.12571.
169. Yang, Y.; Jin, S.; Liu, R.; Bing Kang, S.; Yu, J. Automatic 3d indoor scene modeling from single panorama. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3926–3934.
170. Nekrasov, V.; Dharmasiri, T.; Spek, A.; Drummond, T.; Shen, C.; Reid, I. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7101–7107.
171. Sarlin, P.-E.; Cadena, C.; Siegwart, R.; Dymczyk, M. From coarse to fine: Robust hierarchical localization at large scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12716–12725.
172. Zhao, Z.; Chen, X. Towards Spatio-Temporally Consistent Semantic Mapping. In *Robot Soccer World Cup*; Springer: Berlin/Heidelberg, Germany, 2014.
173. Zhao, Z.; Chen, X. Building temporal consistent semantic maps for indoor scenes. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 88–95.

174. Gupta, S.; Arbeláez, P.; Girshick, R.; Malik, J. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *Int. J. Comput. Vis.* **2015**, *112*, 133–149. [CrossRef]
175. Lei, P.; Todorovic, S. Recurrent temporal deep field for semantic video labeling. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
176. Mustafa, A.; Kim, H.; Guillemaut, J.-Y.; Hilton, A. Temporally coherent 4d reconstruction of complex dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4660–4669.
177. He, Y.; Chiu, W.-C.; Keuper, M.; Fritz, M. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4837–4846.
178. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9297–9307.
179. Choy, C.; Gwak, J.; Savarese, S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3075–3084.
180. Mustafa, A.; Hilton, A. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 422–431.
181. Previtali, M.; Barazzetti, L.; Brumana, R.; Scaioni, M. Towards automatic indoor reconstruction of cluttered building rooms from point clouds. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *2*, 281–288. [CrossRef]
182. Tran, H.; Khoshelham, K.; Kealy, A.; Díaz-Vilariño, L. Shape Grammar Approach to 3D Modeling of Indoor Environments Using Point Clouds. *J. Comput. Civ. Eng.* **2019**, *33*, 04018055. [CrossRef]
183. Diakité, A.A.; Zlatanova, S. Spatial subdivision of complex indoor environments for 3D indoor navigation. *Int. J. Geograph. Inf. Sci.* **2018**, *32*, 213–235. [CrossRef]
184. Zhang, Y.; Xu, W.; Tong, Y.; Zhou, K. Online structure analysis for real-time indoor scene reconstruction. *ACM Trans. Graph. TOG* **2015**, *34*, 1–13. [CrossRef]
185. Teo, T.-A.; Cho, K.-H. BIM-oriented indoor network model for indoor and outdoor combined route planning. *Adv. Eng. Inform.* **2016**, *30*, 268–282. [CrossRef]
186. Vanclooster, A.; Van de Weghe, N.; De Maeyer, P. Integrating indoor and outdoor spaces for pedestrian navigation guidance: A review. *Trans. GIS* **2016**, *20*, 491–525. [CrossRef]