DOI: 10.1111/cgf.14021
EUROGRAPHICS 2020
R. Mantiuk and V. Sundstedt
(Guest Editors)

Volume 39 (2020), Number 2
STAR – State of The Art Report

State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments

Giovanni Pintore[1]*, Claudio Mura[2]*, Fabio Ganovelli[3], Lizeth Fuentes-Perez[2], Renato Pajarola[2], and Enrico Gobbetti[1]

[1] Visual Computing, CRS4, Italy
[2] Visualization and MultiMedia Lab, University of Zurich, Switzerland
[3] Visual Computing Group, ISTI CNR, Italy

Abstract

Creating high-level structured 3D models of real-world indoor scenes from captured data is a fundamental task which has important applications in many fields. Given the complexity and variability of interior environments and the need to cope with noisy and partial captured data, many open research problems remain, despite the substantial progress made in the past decade. In this survey, we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. After providing a characterization of input sources, we define the structure of output models and the priors exploited to bridge the gap between imperfect sources and desired output. We then identify and discuss the main components of a structured reconstruction pipeline, and review how they are combined in scalable solutions working at the building level. We finally point out relevant research issues and analyze research trends.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics;*** *Shape modeling;* ***Computer vision;*** *Computer vision problems; Shape inference; Reconstruction;* • ***Applied computing*** → *Computer-aided design;*

1 Introduction

The automated reconstruction of 3D models from acquired data, be it images or 3D point clouds, has been one of the central topics in computer graphics and computer vision for decades. This field is now thriving, as a result of complementing scientific, technological and market trends. In particular, in recent years, the widespread availability and proliferation of high-fidelity visual/3D sensors (e.g., smartphones, commodity and professional stereo cameras and depth sensors, panoramic cameras, low-cost and high-throughput scanners) has been matched with increasingly cost-effective options for large data processing (e.g., cloud and GPU-accelerated computation), as well as with novel means of visual exploration, from mobile phones to immersive personal displays.

In this context, one of the rapidly emerging sub-fields is concerned with the automatic reconstruction of indoor environments. That is, a 3D representation of an interior scene must be inferred from a collection of measurements that sample its shape and/or appearance, exploiting and/or combining sensing technologies ranging from passive methods, such as single- and multi-view image capturing, to active methods, such as infrared or time-of-flight cameras, optical laser-based range scanners, structured-light scanners, and LiDAR scanners [BTS*17].

Based on the raw data acquired by these devices, many *general* surface reconstruction methods focus on producing accurate and dense 3D models that faithfully replicate even the smallest geometry and appearance details. In this sense, their main goal is to provide the most accurate representation possible of all the surfaces that compose the input scene, disregarding its structure and semantics or possibly only exploiting them to maximize the fidelity of the output surface model. A number of more *specialized* indoor reconstruction solutions focus, instead, on abstracting simplified high-level structured models that optimize certain application-dependent characteristics [IYF15].

The focus on high-level structured models is motivated by several reasons. First of all, their availability is necessary in many fields. For example, applications such as the generation or revision of building information models (BIM) require, at least, the determination of the bare architectural structure [MMJV*14, TCZ15]. On the other hand, information on the interior clutter, in terms of 3D footprint of major indoor objects, is necessary in many other use cases, such as guidance, energy management, security, evacuation planning, location awareness or routing [IYF15]. Even when the goal is solely for visualization, structured simplified models need to be extracted as a fundamental component of a renderable model. This is because

* Joint first authors

narrow spaces, windows, non-cooperative materials, and abundant clutter make the transition from the acquisition of indoor scenes to their modeling and rendering a very difficult problem. Thus applying standard dense surface reconstruction approaches, which optimize for completeness, resolution and accuracy, leads to unsatisfactory results, as noted in earlier works [KSF*12, XF14].

Automatic 3D reconstruction and modeling of indoor scenes, has thus attracted a lot of research in recent years, making it an emerging well-defined topic. In particular, the focus has been on developing specialized techniques for very common and very structured multi-room environments, such as residential, office, or public buildings, which have a substantial impact on architecture, civil engineering, digital mapping, urban geography, real estate, and more [IYF15]. Commercial solutions in these areas range from generic approaches to create virtual tours of buildings based on panoramic images and videos (e.g., *3DVista* [3DV99]), to frameworks for supporting the construction process by mapping acquired visual or laser data to a reference planimetry or 3D CAD (e.g., *StructionSite* [Str16] or *Reconstruct* [Rec16]), to ecosystems offering reconstruction and exploration of structured models in the form of services in a cloud environment (e.g., *Matterport* [Mat17]).

In the indoor reconstruction context, the fundamental tasks are the discovery of structural elements, such as rooms, walls, doors, and indoor objects, and their combination in a consistent structured 3D shape and visual representation. The research community working on these problems appears, however, fragmented, and many different vertical solutions have been proposed for the various motivating applications.

In the recent past, extensive surveys have been presented for several aspects of indoor capture, mainly focusing, however, on very specific acquisition and modeling aspects (e.g., general 3D reconstruction of all scene surfaces), or on specialized (non-graphic) applications (Sec. 2). Instead, in this survey we provide an up-to-date integrative view of the field, bridging complementary views coming from computer graphics and computer vision. The target audience of our report includes researchers in geometric modeling, as well as practitioners in the relevant application fields. Researchers will find a structured overview of the field, which organizes the various problems and existing solutions, classifies the existing literature, and indicates challenging open problems. Domain experts will, in turn, find a presentation of the areas where automated methods are already mature enough to be ported into practice, as well as an analysis of the kind of indoor environments that still pose major challenges.

After summarizing the related survey literature (Sec. 2), we discuss shape and color sources generated by indoor mapping devices and describe several open datasets available for research purposes (Sec. 3). We then provide an abstract characterization of the typical structured indoor models, and of the main problems that need to be solved to create such models from imperfect input data, identifying the specialized priors exploited to address significantly challenging imperfections in visual and geometric input (Sec. 4). The various solutions proposed in the literature, and their combination into global reconstruction pipelines are then analyzed (Sec. 5–9). We finally point out relevant research issues and analyze research trends (Sec. 10).

2 Related surveys

Reconstruction of visual and geometric models from images or point clouds is a very broad topic in computer graphics and computer vision. This survey focuses on the specific problems and solutions relating to the reconstruction of *structured 3D indoor models*. We do not specifically aim at reconstructing detailed surfaces from dense high-quality data, nor assigning semantic to existing geometry, but rather we cover the extraction of an *approximate structured geometry* connected to a *visual representation* from sparse and incomplete measurements.

A general coverage of methods for 3D surface reconstruction and primitive identification is available in recent surveys [BTS*17, KYZB19], and we will build on them for the definition of general problems and solutions. In the same spirit, we do not specifically cover interactive or online approaches; those interested in online reconstruction can find more detail on the topic in the survey by Zollhöfer et al. [ZSG*18]. We also refer the reader to an established state-of-the-art report on urban reconstruction [MWA*13] for an overview of the companion problem of reconstructing (from the outside) 3D geometric models of urban areas, individual buildings, façades, and further architectural details.

The techniques surveyed in this state-of-the-art review also have an overlap with the domains of Scan-to-BIM or Inverse-CAD, where the goal is the automatic reconstruction of full (volumetric) information models from measurement data. However, the overlap is only partial, since we do not cover the assignment of full semantic information and/or the satisfaction of engineering construction rules, and Scan-to-BIM generally does not cover the generation of visual representations, which is necessary for rendering. Moreover, most Scan-to-BIM solutions are currently targeting (dense) point cloud data, while we cover solutions starting from a variety of input sources. It should be noted that, obviously, relations do exist, and many of the solutions surveyed here can serve as good building blocks to tackle the full Scan-to-BIM problem. We refer the reader to established surveys in the Scan-to-BIM area for a review of related techniques based on point-cloud data [THA*10, VSS14, PAN*15], general computer vision [FDL15], and RGB-D data [CLH15].

3 Background on input data capture and representation

Indoor reconstruction starts from measured data obtained by surveying the indoor environment. Many options exist for performing capture, ranging from very low-cost commodity solutions to professional devices and systems. In this section, we first provide a characterization of the various input sources and then provide a link to the main public domain datasets available for research purposes.

3.1 Input data sources

Indoor mapping is required for a wide variety of applications, and an enormous range of 3D acquisition devices have been proposed over the last decades. From LiDAR to portable mobile mappers, these sensors gather shape and/or color information in an effective, often domain-specific, way [XAAH13, LKN*17]. In addition, many general-purpose commodity solutions, e.g., based on smartphones and cameras, have also been exploited for that purpose [PAG14, SS12]. However, a survey of acquisition methods is out of the scope

of this survey. We rather provide a classification in terms of the characteristics of the acquired information that have an impact on the processing pipeline.

Purely visual input sources. Imagery is perhaps the most obvious input source from which to attempt reconstruction since common images acquired indoors have the advantage of being very easy and affordable to obtain, store and exchange. For this reason, image datasets of various kinds have often been exploited as input for reconstruction, inferring all the geometric information from purely visual (typically RGB) signals. The most basic input source is the *single still image*. This, however, is inherently ambiguous and partial, and only appears in specialized solutions targeting room corners or interior objects rather than fully structured models [CY99, ZCSH18]. In particular, the small field-of-view of standard cameras makes contextual information insufficient for reliable object detection or full-room reconstruction. To overcome this limitation, a growing trend in single-image solutions is to use a 360° *full-view panorama* for indoor capture [ZSTX14, YZ16, YJL*18]. In contrast, taking multiple pictures from several viewpoints makes it possible to ensure a larger coverage both of single rooms, by reducing unseen areas due to occlusion, and multi-room environments, by distributing views over the entire floor plan. Moreover, Structure-from-Motion (SfM) techniques can be applied to recover geometric information, in terms of triangulated features and camera poses, which provides essential starting cues for further joint analysis of the correspondence between the measured colors and the inferred geometry. For this reason, *registered image collections* and *registered* 360° *full-view panorama collections* are becoming the most widespread purely-visual input sets [FCSS09a, BFFFS14, PGJG19]. In outdoor settings or for the capture of individual objects, such collections can be used in a relatively straightforward way to generate dense point clouds through multi-view-stereo methods [SCD*06, MWA*13]. However, this is not the case when dealing with interior scenes, due to the presence of texture-poor surfaces (such as painted walls), occluding structures (both permanent and temporary) that hamper visibility reasoning and feature triangulation. Furthermore, thin structures (e.g., walls, doors, tables) demand extremely high resolution to be resolved [FCSS09b]. Therefore, methods dealing with images should always consider that 3D evidence is sparse and uneven. Nowadays, such image collections are acquired using mobile setups, which, even in the simplest settings, typically provide additional information for each capture through sensors reading from an inertial measurement unit (IMU), composed of gyroscopes, accelerometers, magnetometers, or other sensors. It is therefore not uncommon to see indoor reconstruction systems to exploit this sort of input, from the tracking of trajectories to detect free paths in the environments [PGGS16a] or to the alignment of images to the gravity vector [PGJG19]. Similarly, since capture devices increasingly feature wireless connectivity, other authors have proposed, for the same purpose, to also exploit WiFi triangulation to infer the camera pose associated to each captured image [SCC12].

Purely geometric input sources. At the opposite end of the spectrum there are 3D point sets, which provide purely geometric information on the surveyed environment. While historically these datasets were acquired with survey-grade terrestrial laser range scanners, in recent times more often faster and often cheaper mobile scanning solutions have been used [LKN*17]. The shift to mobile solutions makes it easier to acquire a scene from multiple points of view, possibly up to a continuous stream [IKH*11] and consequently reduced amount of unseen areas. In this context, the sampling rate is generally assumed geometrically dense (e.g, sub-centimetric), and, due to active scanning, generally covers all typical indoor surfaces [THA*10] with a good sampling rate. Since capture is dense, local geometric information such as normals and curvature can typically be extracted. Many solutions rely on these features for the detection of the surfaces of structural elements (e.g., using normals for planar patch segmentation [MMP16]). While the most general representation is the *3D point cloud*, simply consisting of a list of discrete points that sample the scene, several variations exist. The most common one is to consider a *registered 3D range scan collection*, which provides knowledge of the pose of each of the scanning probes in a globally registered frame, and represents each scan as a range image. Such additional information is exploited in a number of structured reconstruction systems [TZ12, TZ13, MMJV*14].

Multimodal colorimetric and geometric input sources. While the two preceding input sources only provide measured information either on appearance or on geometry, it is increasingly common to exploit input sources that provide combined color and data measurements. The combination of active scanners with passive cameras to jointly acquire shape and color has a long history [PAD*98]. Currently, this area is again very active due to the many affordable solutions that are emerging both in the professional (e.g., backpacks [LKN*17]) and consumer markets (e.g., consumer RGB-D cameras [CLH15]). Note, however, that while modern low-cost mobile depth-sensing devices, such as generic RGB-D cameras, have become a promising alternative for widespread short-range 3D acquisition, rooms larger than a few meters, for example a hotel hall, are outside their depth range and make the acquisition process more time consuming [GH13, JGSC13]. For this reason, several solutions have been designed for specific indoor capture purposes [LKN*17]. Independent from the acquisition device and process, but instead from the processing point of view, there are at least three principal kinds of sources. The first input source is the *colored 3D point cloud*, which is typically generated by devices where scanning and color capture have a similar resolution. These clouds can be obtained directly by multi-modal devices or by subsequent registration of a photographic acquisition over a separately acquired raw 3D point cloud [PGCD17]. From the point of view of processing, this type of input presents the same characteristics of the plain 3D point clouds, and the additional color information is exploited to help segmentation and/or for visual display (Sec. 9). A second input source is the *3D point cloud with registered (panoramic) images*, generally acquired by combinations of rigidly aligned scanners and cameras. Typically, the geometric information is at much lower resolution with respect to the images, which, however, are taken from just a few positions. In this case, the 3D points can be used as anchors to provide 3D evidence during image analysis, and the known poses of the cameras associated with the images help with visibility analysis and geometric reasoning [WLL*18]. The last common input source is the *registered RGB-D collection*, which is a collection of color and range maps aligned in a global reference frame. This representation is becoming dominant today due to the increasing proliferation and diffusion of affordable sensing systems that cap-

ture RGB images along with per-pixel depth information [CLH15]. As for image collections, these RGB-D collections are enriched with the poses associated to the capture, as well as often with additional positioning information coming from IMUs, odometry, or other sensors [SGD*18]. Since reasonably dense color and geometry information is available for each pose, data fusion methods can be exploited to recover structures [LWF18a, CLWF19].

3.2 Open research data

A notable number of freely available datasets containing indoor scenes have been released in recent years for the purposes of benchmarking and/or training learning-based solutions. However, most of them are more focused on scene understanding [Uni16] than reconstruction, and often only cover portions of rooms [New12, Cor12, Was14, Pri15, Tec15, Sta16b]. Many of them have been acquired with RGB-D scanners, due to the flexibility and low-cost of this solution (see an established survey [Fir16] for a detailed list of them).

In the following, as well as in Tab. 1, we summarize the major open datasets that have been used in general 3D indoor reconstruction research:

- **SUN360 Database** [Mas12, XEOT12, ZSTX14, YZ16, PPG*18, PGP*18]: Comprehensive collection of equirectangular spherical panoramas of a large variety of indoor scenes filled with objects. To build the core of the dataset, the authors downloaded a massive amount of high-resolution panorama images from the Internet, and grouped them into different place categories. This is currently a reference dataset for single-panorama analysis. A tool is also provided to generate perspective images from the panorama and thus extend its use to the analysis of conventional pin-hole images. However, no depth information is provided as ground-truth.
- **SUN3D Database** [Pri13, XOT13, CZK15, CDF*17, DNZ*17]: 415 RGB-D image sequences captured by Microsoft Kinect from 254 different indoor scenes, in 41 different buildings across North America, Europe, and Asia. Semantic class polygons and instance labels are given on frames for some sequences. Camera pose for each frame is also provided for registration.
- **UZH 3D Dataset** [Uni14, MMJV*14, MPM*14, MMP16]: 3D point cloud models of 40 individual rooms and 13 multi-room interiors. Each model consists of separate scans (in grid format) and includes per-scan alignment information. The scans represent office environments and apartments, mostly obtained by real-world scanning but also including 4 synthetic scenes. The environments include sloped ceilings and arbitrary oriented walls that are challenging for most techniques. The real-world scans were acquired using a Faro Focus3D laser scanner based on phase-shift technology, which has a much higher precision than consumer-level cameras like Microsoft Kinect.
- **SUNCG Dataset** [Pri16, SYZ*17, LWF18a, ASZS17, CDF*17]: 45,622 synthetic indoor scenes with manually created room and furniture layouts, including annotations. Images can be rendered from the geometry, but are not provided by default. Due to legal issues it is necessary to contact the authors for accessing the data.
- **BundleFusion Dataset** [Sta16a, DNZ*17, HDGN17, FCW*17]: Sequences of RGB-D depth-image data for 7 small indoor scenes, captured using a *Structure Sensor* depth sensor coupled with an iPad color camera. The scenes, largely consisting of single rooms or studio type apartments, are extensively scanned resulting in average trajectories covering 60m and containing 5K frames.
- **ETH3D Dataset** [SSG*17, YLL*19]: 16 indoor scenes, captured as collections of registered RGB images. Scenes are portions of a variety of indoor environments, both small and large. As the purpose of the database is to benchmark multi-view stereo algorithms, ground truth point clouds and depth maps are provided.
- **ScanNet Data** [DCS*17a, DCS*17b, CDF*17]: RGB-D video dataset of academic buildings and small apartments, containing 2.5 million frames in more than 1500 scans, annotated with 3D camera poses, surface reconstructions and instance-level semantic segmentation. To collect this data, authors developed a pipeline that includes automated surface reconstruction and crowd-sourced semantic annotation. It provides automatically computed (and human verified) camera poses and surface reconstructions, instance and semantic segmentation on reconstructed mesh. Aligned 3D CAD models are also provided for each scene.
- **Matterport3D Dataset** [Mat17, CDF*17]: Large-scale RGB-D dataset containing 10,800 panoramic views from 194,400 RGB-D images of 90 luxurious houses. Annotations are provided with surface reconstructions, camera poses, and 2D and 3D semantic segmentations. It includes both depth and color panoramas for each viewpoint, samples human-height viewpoints uniformly throughout the entire environment, provides camera poses that are globally consistent and aligned with a textured surface reconstruction, includes instance-level semantic segmentation into region and object categories, and provides data collected from living spaces in private homes. Due to instrument limitation, the visual coverage is sometimes limited and in general does not cover the hemispheres of the panorama, so the upper parts of the spherical image are missing or completed by inpainting [CDF*17], thus limiting the ability to reconstruct using only the visual data.
- **2D-3D-S Dataset** [Sta17, ASZS17]: 6 large-scale indoor scans of office spaces, captured by using the same *Matterport* system of the Matterport3D dataset. The dataset contains over 70,000 RGB images, along with the corresponding depths, surface normals, semantic annotations, global XYZ images (all in forms of both regular and 360 equirectangular images) as well as camera information. It also includes registered raw and semantically annotated 3D meshes and point clouds.
- **FloorNet Dataset** [LWF18b, LWF18a, CLWF19]: RGB-D video streams for 155 residential houses or apartments acquired with Google Tango phones, annotated with their complete floor plan information, such as architectural structures, icons, and room types.
- **CRS4/ViC Research Datasets** [CRS18, PPG*18, PGP*18, PGJG19]: Registered sets of high-resolution equirectangular panoramas covering 360x180 full view for a variety of real-world indoor scenes and the objects within. Provided scenes include multi-room environments, sloped ceilings, walls not aligned on a Cartesian grid, and many features which are usually challenging for existing techniques. The ground truth measures of the floor plans, obtained through laser measurement, and the height of the first camera (170 cm from the floor for almost all datasets) are provided, thus allowing the metric scaling of the models.
- **Replica Dataset** [SWM*19]: A dataset of 18 highly photo-realistic 3D indoor scene reconstructions at room and building

Name	Data	Source	Coverage	Capture	Notes
SUN 360 Database [Mas12]	Individual RGB	Real	Panoramic	Tripod	Whole rooms;
SUN 3D Database [Pri13]	Registered RGB-D	Real	Perspective	Hand-held video	Whole rooms; PL; 3D models
UZH 3D Dataset [Uni14]	Registered PC	Real/Synth	Scan	Tripod	Large-scale; multi-room; 3D models
SunCG Dataset [Pri16]	CAD models	Synth	All	Manual modeling	Large-scale; FL
BundleFusion Dataset [Sta16a]	Registered RGB-D	Real	Perspective	Hand-held video	Room-scale; FL; 3D models
ETH3D Dataset [ETH17]	Registered RGB	Real	Perspective	Tripod	Scene parts; ground truth (PC+DM)
Matterport 3D [Mat17]	Registered RGB-D	Real	Panoramic	Tripod	Large-scale; multi-room; FL
ScanNet [DCS*17a]	Registered RGB-D	Real	Perspective	Hand-held video	Large-scale; multi-room; FL;3D models
2D-3D-S [Sta17]	Registered RGB-D	Real	Panoramic	Tripod	Large-scale; multi-room; FL
FloorNet Data [LWF18b]	Registered RGB-D	Real	Perspective	Hand-held video	Large-scale; FL
CRS4/ViC Datasets [CRS18]	Registered RGB	Real	Panoramic	Tripod	Large-scale; multi-room; 3D models
Replica Dataset [SWM*19]	CAD models	Synth	All	Manual modeling	Highly realistic; FL
Structured3D Dataset [ZZL*19]	CAD models	Synth	All	Manual modeling	Large scale; FL

Table 1: ***Open indoor datasets.*** *Major datasets that are available for research purposes and have been used for structured 3D reconstruction. PC: point clouds (ground truth); DM: depth maps (ground truth); PL: dataset is partially labeled; FL: full labeling (objects and background).*

scale. Each scene consists of a dense mesh, high-resolution high-dynamic-range (HDR) textures, per-primitive semantic class and instance information, and planar mirror and glass reflectors.

- **Structured3D Dataset** [ZZL*19, SHSC19]: A synthetic dataset providing large-scale photo-realistic images with rich 3D structure annotations.

4 Targeted structured 3D model

The goal of structured 3D indoor reconstruction is to transform an input source containing a sampling of a real-world interior environment into a compact structured model containing both geometric and visual abstractions. Each distinct input source, as described in Sec. 3, tends to produce only partial coverage and imperfect sampling, making reconstruction difficult and ambiguous. For this reason, research has concentrated on defining priors in order to combat imperfections and focus reconstruction on very specific expected indoor structures, shapes, and visual representations. In the following, we first characterize the artifacts typical of indoor model measurement (Sec. 4.1), before defining the structure and priors commonly used in structured 3D indoor reconstruction research (Sec. 4.2) and the sub-problems connected to its generation (Sec. 4.3).

4.1 Artifacts

The general properties of the input source, be it geometric, visual, or both, are an important factor in understanding the behavior of reconstruction methods. Berger et al. [BTS*17] have characterized sampled sources according to the properties that have the most impact on reconstruction algorithms, identifying them into *sampling density*, *noise*, *outliers*, *misalignment*, and *missing data*. While the characterization was introduced for point clouds, it can be adopted for all the sources described in Sec. 3.1.

In the particular case of indoor environments, the artifacts associated with each one of these characteristics have some specific forms. In particular, in terms of *density*, not only 3D scans typically produce a nonuniform sampling on the surface depending on scanning geometry, but also 3D data derived from visual sources is very sparse and strongly depends on the amount of texture [FDL15]. *Noise* and *outliers* are very common in all acquisition modalities, in particular due to the widespread presence of transparent or reflective surfaces in interiors (e.g., windows and glass surfaces) [LKN*17], as well as the great amount of clutter in front of structures of interest. This is in contrast to other typical scanning scenarios [AH11]. Moreover, while *misalignments*, including loop closure problems due to drift, are not substantially different than in other incremental scanning approaches [ZXTZ15], the amount of *missing data* is extremely large for all kinds of input sources. The lack of data is due to the difficulty in covering all the structures because of furniture and narrow spaces, as well as the dominance of texture-poor surfaces that make 3D triangulation ambiguous in the case of visual sources [FCSS09b].

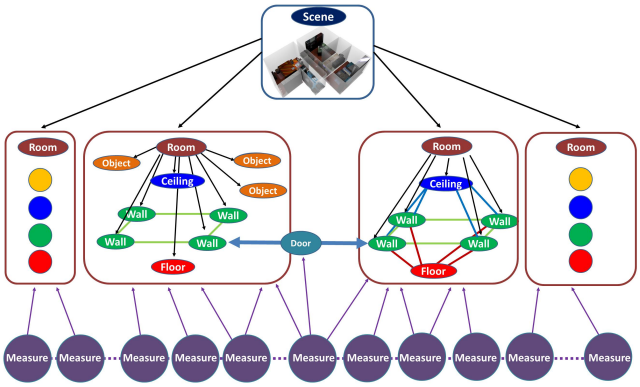


Figure 1: ***Abstract structured indoor model.*** *The typical output of a structured reconstruction pipeline is an architectural structure defined by a graph of rooms bounded by walls, floor, and ceiling, as well as connected by doors/passages and containing objects, such as furniture and wall-mounted items. The structured model thus combines a topological part (the connection graph), a geometric part (the shape of the various components) and a visual part (the appearance model of the different nodes). Moreover, an explicit mapping between input sources (e.g., scans or shots) and regions of the model is often generated as well.*

4.2 Reconstruction priors and abstract indoor model

A wide variety of representations could be used to describe the structure, geometry and appearance of an indoor scene automatically reconstructed from measurements. However, due to the dominance of artifacts in all kinds of datasets, it is extremely challenging to produce complete high-quality and high-detail 3D indoor scene models in the general case. In particular, without prior assumptions, the reconstruction problem is ill-posed, since an infinite number of solutions may exist that fit under-sampled or partially missing data.

For this reason, research in structured indoor reconstruction has focused its efforts on formally or implicitly restricting the target output model, in order to cover a large variety of interesting use-cases while making reconstruction tractable. Several authors [IYF15, AHG*19] have proposed a structured representation in the form of a scene graph, where nodes correspond to elements with their geometry and visual appearance, and edges to geometric relationships (e.g., adjacency). Moreover, there is a clear distinction between permanent structures and movable objects. Armeni et al. [AHG*19] have proposed to use such a graph in the context of scene understanding as a unified structure on which to base all the semantic information that can be recovered from an indoor scan. Furthermore, an explicit connection between input sources (e.g., scans or pictures) and the model is often generated and included in the structure as well. This mapping between input and output is often used both for navigation applications (see Sec. 9) and for serving as a basis for further semantic analysis [AHG*19].

In this context, the desired output of a structured reconstruction pipeline is an architectural data structure defined by a graph of rooms bounded by walls, floor and ceiling, as well as connected by portals, such as doors or passages (see Fig. 1). Rooms may contain indoor objects, typically including furniture or other movable items, as well as wall-mounted items. The structured model thus combines a topological part (the connection graph), a geometric part (the shape of the various components) and a visual part (the appearance model of the different nodes).

Such a structural prior is used to guide the reconstruction. However, it is often insufficient by itself to ensure recovery in the presence of partial or corrupted data. Thus in addition to exploiting typical priors from the surface reconstruction domain, such as surface and volumetric smoothness, knowledge of known geometric primitives or global regularity such as symmetry or repetition [BTS*17], very specific geometric priors for structural recovery have been introduced in the indoor reconstruction literature. The following priors are the most commonly used ones:

- **[FW]** *floor-wall* [DHN06]: The environment is composed by a single flat floor and straight vertical walls; with this prior, the ceiling is completely ignored in the reconstruction.
- **[CB]** *cuboid* [HHF09]: The environment is a single room of cuboid shape; a room's boundary surface is thus made from six rectangles placed at right angles.
- **[MW]** *Manhattan world* [CY99]: The environment has an horizontal floor and ceiling, and vertical walls which all meet at right angles; i.e., the boundary of a room is formed by planes lying in one of three mutually orthogonal orientations; note that the floors and ceilings may be at different elevations.
- **[AW]** *Atlanta world* (a.k.a. *Augmented Manhattan World*) [SD04]: The environment has horizontal floor, ceiling and vertical walls, possibly at different elevations; this is similar to MW, without the restriction of walls meeting at right angles.
- **[IWM]** *Indoor World Model* [LHK09]: MW with single floor and single ceiling; note that this prior introduces a symmetry between the shape of the ceiling and floor, simplifying geometric reasoning in case of occlusions.
- **[VW]** *Vertical Walls* [PGP*18]: Floor and ceiling can be sloping, but walls are vertical; as for the other priors with vertical walls, this permits to exploit top-down views to find floor plans; however, 3D reasoning must be employed to determine floor and ceiling shape.
- **[PW]** *Piecewise planarity* [FCSS09b]: The indoor environment is piecewise planar, and thus rooms are general polyhedra; this assumption imposes only slight restrictions on the possible shapes but necessitates full 3D reasoning.

4.3 Main problems

Starting from the above definitions, we can identify a core set of basic problems that need to be solved to construct the model from observed data. The list is the following:

- **room segmentation** – separate the observed data into different rooms (Sec. 5);
- **bounding surfaces reconstruction** – find the geometry bounding the room shapes, i.e., walls, ceilings, floor and other permanent structures (Sec. 6);
- **indoor object detection and reconstruction** – identify objects contained in rooms to remove clutter and/or reconstruct their footprint or shape (Sec. 7);
- **integrated model computation** – assemble all the individual rooms into a single consistent model, finding portals, building the graph structure (Sec. 8);
- **visual representation generation** – generate a representation suitable for graphics purpose by enriching the structured representation with visual attributes (Sec. 9).

In the following sections, we provide details on the methods that have been proposed for each of these sub-problems.

5 Room segmentation

While a number of early methods focused on reconstructing the bounding surface of the environment as a single entity, without considering the problem of recognizing individual sub-spaces within it, structuring the 3D model of an indoor environment according to its subdivision into different rooms has gradually become a fundamental step in all modern indoor modeling pipelines, regardless of the type of input they consider (e.g. visual vs. 3D data) or of their main intended goal (e.g. virtual exploration vs. as-built BIM) [IYF15].

Room segmentation is important for several reasons. First of all, segmenting the *input* before the application of the reconstruction pipeline makes it possible to apply per-room reconstruction methods using only carefully selected samples, improving performance and accuracy through the pre-filtering of massive amounts of outliers [PGJG19]. Second, structuring the *output* 3D model according to its subdivision into different rooms is of paramount importance for several different application scenarios, including navigation and path planning, emergency management, office space management and automatic generation of furnishing solutions [OVK19].

One major challenge involved in this task is the lack of a clear definition for a room [TZ14], which leads to the application of a variety of approaches that are combined together at various stages of the pipeline (see Table 2 for an overview).

Method	Input type	Partition type	Output	Features	Techniques
Turner et al. [TZ14]	Dense PC	Reco. subdiv.	Labeled cells	Scanner pos.	Over-segmentation + merging
Mura et al. [MMJV*14]	Dense PC	Reco. subdiv.	Clusters of cells	Scanner pos.; polygonal regions	Iterative binary k-medoids + merging
Ikehata et al. [IYF15]	Dense RGB-D	Reco. subdiv.	Clusters of pixels	3D points visibility	k-medoids clustering
Ochmann et al. [OVWK16]	Dense PC	Reco. subdiv.	Labeled cells	Scanner pos.	Over-segmentation + merging
Armeni et al. [ASZ*16]	Dense PC	Input part.	Labeled points	Wall detection using 2D filters	Over-segmentation + merging
Mura et al. [MMP16]	Dense PC	Input part.	Labeled cells	Scanner pos.	Markov clustering
Ambrus et al. [ACW17]	Dense PC	Reco. subdiv.	Labeled cells	Synthetic viewpoints	Over-segmentation + merging
Mura et al. [MP17]	Dense PC	Input part.	Clusters of patches	Synthetic viewpoints	Markov clustering
Murali et al. [MSOP17]	Dense PC	Reco. subdiv.	Clusters of cuboids	Synthetic viewpoints	Over-segmentation + merging
Bobkov et al. [BKHS17]	Dense PC	Reco. subdiv.	Labeled points	Voxel-based distance field	Hierarchical DBSCAN clustering
Pintore et al. [PGP*18]	Sparse RGB	Input part.	Clusters of images	MV feat. visibility; camera path	LSD clustering
Ochmann et al. [OVK19]	Dense PC	Input part.	Clusters of patches	Synthetic viewpoints (patches)	Markov clustering
Pintore et al. [PGJG19]	Sparse RGB	Input part.	Clusters of images	1D photoconsistency	Weighted graph
Chen et al. [CLWF19]	Dense RGB-D	Reco. subdiv.	Raster pixel mask	Disjoint regions	Mask R-CNN

Table 2: ***Room segmentation methods***. *Summary of the approaches described in Sec. 5, arranged by chronological order. Sparse/dense input type is related to spatial coverage (i.e., how many scans/poses). The partition type indicates whether the room segmentation is obtained by pre-partitioning the input data before reconstruction (input part.) or by subdividing the reconstructed model (reco. subdiv.).*

5.1 Input data partitioning

A pre-segmentation of the input into clusters, prior to, or independently from, any further 3D analysis to generate a structured model, is useful for both efficiency and accuracy reasons. It has been generally applied prior to the reconstruction pipeline, typically requiring user input to label input scans or input images in order to perform subsequent reconstruction steps in a more efficient local manner. A typical assumption here is, for instance, that the survey is planned to have a single scan per room, and to exploit this known partitioning for organizing all processing steps [OVW*14]. If multiple scans per room are present, Markov clustering can be applied to find rooms based on visible surface overlap [MMP16]. The same approach can be applied to an unordered point cloud by generating virtual scanning positions using the ExploreMaps technique [DBGBR*14], and applying the clustering to them.

More elaborate solutions, however, are necessary if 3D data is sparse or missing. In particular, grouping unordered image collections into room sets requires special care. A common approach is to just apply multi-view registration, and, then, group in the same room the images that share a set of 3D features, used both as indicators that the same surface is present in the two images, and that the lines of sight go through empty space [FCSS09b]. Similarly, Pintore et al. [PGP*18] have proposed to exploit a specialized approach to group input panoramic images exploiting triangulated multi-view features to estimate strong occlusions between camera poses and breaks among the camera trajectory. These approaches, however, are likely to fail for many indoor environments where 3D features are very sparse.

An alternative solution is to apply general instance-level image retrieval approaches, in which, given a starting image depicting a particular object, the aim is to retrieve all other similar images containing the same object/scene/architecture that may be captured under different views, illumination, or with occlusions, using a combination of global and local image similarity metrics [ZYT17]. These solutions are very appealing, but only solve part of the problem, since, especially in large-scale office settings, the presence of standardized furniture is likely to lead to many false positives. For this reason, solutions have been proposed specifically for indoor settings. Pintore et al. [PGJG19] have proposed an ad-hoc image-to-image similarity estimation to group panorama in same-room images. They measure how well the horizontal central slice of one image can be warped into the same portion of the other. Under the hypothesis that all panorama images are taken at approximately the same height, this measure tells how likely it is that the two images where taken in the same room, since the warping preserves the order of seen objects. They build a graph where nodes are images and edges are weighted with the similarity value of their extremes. Then, they partition of the images in groups, one group per room using a clustering method based on random walks [HK01]. Their method has shown to improve the accuracy of further reconstruction steps, both for room boundaries determination (Sec. 6) and interior object detection (Sec. 7).

5.2 Inferring the room subdivision for structured modeling

When 3D data is available, either as a result of dense 3D capture or as an outcome of previous reconstruction steps, room segmentation exploits geometric reasoning approaches. Many approaches move from the observation that different locations inside the same room view similar parts of the environment and cast room detection as a visibility-based clustering. Other researchers rely on the presence of a door to infer the separation between two distinct rooms. The length of the separating boundary between rooms has also been considered, based on the observation that the interface between rooms is typically small.

Much of the initial work on modeling interiors from 3D data has not tackled the separation into multiple rooms, but the task of reconstructing the bounding surfaces of the whole indoor space, considering it as a single object. In this context, the goal is to segment the overall volume into inside and outside regions. This amounts to detecting which regions in the space surrounding the input 3D model, often pre-partitioned into discrete regions, are inside or outside the permanent structures bounding the environment. This has a clear analogy with the more general (smooth) surface reconstruction, which is a fundamental and more well-studied problem in computer graphics [BTS*17].

Many approaches consider the number of points that fall inside

a region as an indicator of that region being inside (see Fig. 9a) - in the simplest case, by plain thresholding on the sheer number of points [BB10]. However, the presence of scanned points is more often regarded as a sign of a transition from inside to outside space, with more reliable visibility-based criteria being used to assess whether a region lies in the inner space. If the input 3D model does not include the position of the scanning device, ray-casting can be used to this purpose: at a given location, the fraction of rays shot in all directions and intersecting patches of scanned points can be interpreted as the probability that that location is inside the environment [OLA14], as shown in Fig. 9b. Many approaches, however, rely on the position (or the trajectory) of the scan device to identify a location as belonging to the inner space [TZ12, TZ13].

More recently, a clever use of the scan position has allowed to go beyond the sheer inner space detection and to integrate room segmentation in the reconstruction process. Mura et al. [MMJV*14] propose a pipeline in which room detection is incorporated in the reconstruction process. In their work, rooms are obtained as clusters of polygonal regions defined on the ground plane of the environment. An iterative binary clustering driven by diffusion distances extracts at each iteration one new room as a cluster of polygonal regions. Scan positions are used to define the termination condition: since each scan position must fall inside a room and, conversely, assuming that each room is scanned from at least one location inside its boundary, the clustering terminates when each input scan position has been assigned to a room cluster. Using this technique, over-segmentation can occur: this is fixed in a post-processing step, in which two adjacent room clusters are merged if no scanned evidence of separating structures is present along their border. Over-segmentation is used programmatically in the approach by Ochmann et al. [OVWK16], also based on detecting rooms as groups of 2D regions on the ground plane of the building. Their method initially assumes a one-to-one mapping between input scans and rooms and assigns each 2D region to a representative scan using a multi-label optimization procedure. The assignment results in clusters of regions, which may not correspond to the actual rooms since multiple scan positions can fall inside the same room. A Support Vector Machine (SVM) classifier is used in post-processing to determine whether the boundary between two adjacent clusters is plausible or not; in the latter case, the two clusters are merged.

The need for a merging step is avoided in a later work by directly clustering the input scan positions based on their visible surface overlap (see Fig. 2), for instance by using Markov Random clustering [MMP16]. In this approach, the correct number of rooms is available before room reconstruction, which allows the subsequent multi-label optimization to extract the final room models.

In the last few years, researchers have overcome the need for input scan positions by computing a set of synthetic viewpoints, which provide the set of labels for a multi-label optimization yielding the room models. Ambrus et al. [ACW17] compute such viewpoints by sampling locations on the medial axis of the occupancy map of the environment, which encodes the locations occupied by scanned points in a top-down view of the scene and thus denotes the regions that are inside the environment. Their intuition is that the points on the medial axis are maximally distant from the bounding walls and therefore correspond to locations from which most of the surrounding room is visible. New viewpoints are sampled from the medial axis in a greedy, iterative process, until most of the locations of the occupancy map are within a minimum distance from a viewpoint. As noted by the authors, this strategy can lead to oversegmentation; this is fixed in a post-processing step along the lines of previous work [MMJV*14, OVWK16]. Instead of optimizing the position of the viewpoints, Mura and Pajarola [MP17] generate an overly large set of view probes in the environment, selecting them as the centers of the leaf cells of an adaptive octree built around the scanned points. The rooms are then extracted using a visibility-based clustering, as in their previous work [MMP16]. An alternative approach is to segment the point cloud into small planar patches and use the centers of such patches as view probes [OVK19]; this has the advantage of not requiring the construction of a supporting data structure.

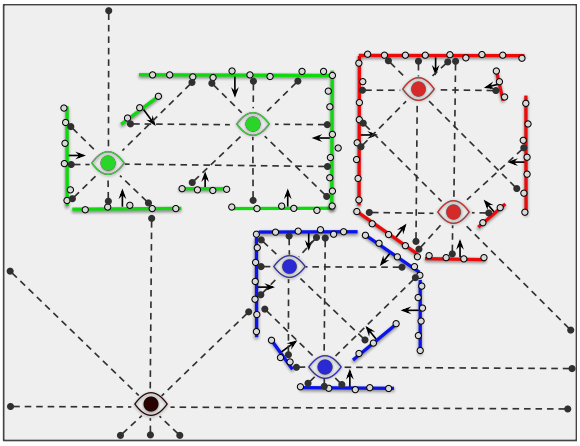


Figure 2: ***Room detection by visibility clustering.*** *The rooms of an environment can be extracted by clustering a set of viewpoints based on their* visible surface overlap*: this is the common set of surfaces (2D line segments in this top-down view) that are visible from all viewpoints. Such surfaces can be extracted from the input measurements (grey circles) and are considered* visible *from a viewpoint if they are not occluded and are facing the viewpoint. The resulting clusters (shown in color-coding) indicate the rooms of the environment; often, a special cluster (black viewpoint) is reserved to the space outside all rooms.*

Some approaches avoid relying on input scan positions for room segmentation, although this often comes at the cost of using strong priors or heuristics. Turner and Zakhor [TZ14] compute a Delaunay triangulation on the ground plane of the environment and select as seed locations for the rooms the triangles whose circumscribed circles are larger than those of neighboring triangles. This heuristic, however, results in over-segmentation and requires iterating the subsequent room reconstruction until convergence. Using the restrictive *Manhattan-World* assumption, Murali et al. [MSOP17] decompose the input scene into cuboids by detecting groups of four orthogonal wall planes and extract rooms by merging adjacent cuboids if the projection of scanned data on their adjacent face is not compatible with the presence of a wall or of a door. The *Manhattan-World* prior is also use in the *Building Parser* project [ASZ*16]. This work is based on detecting the main walls between rooms as *peak-gap-peak* patterns in a top-down density map of the environment. Walls induce an over-segmentation of the ground plane into disjoint sub-spaces; adjacent spaces are then collapsed if the *peak-gap-peak*

pattern is not found on their shared boundary, yielding the final room segmentation. Instead of correcting the number of rooms until convergence during the reconstruction, Bobkov et al. [BKHS17] apply Hierarchical DBSCAN clustering (HDBSCAN) to the cells of a coarse top-down map, driven by a combination of visibility-based distance, euclidean distance and a so-called Potential Field distance: this is computed for each cell of the top-down map from a voxel-based field that encodes the distance to the permanent structures of the environment. The room segmentation defined on the 2D cells is then propagated to the 3D point cloud provided as input, without explicitly reconstructing the bounding surfaces of the rooms. While many of the previous approaches solve the problem in the top-down 2D domain by determining a floor plan, several authors have proposed to detect rooms and room connections by explicitly exploiting also 3D information. A prominent example is the work of Ikehata et al. [IYF15], which propose a 3D modeling framework that reconstructs an indoor scene as a structured model exploiting panoramic *RGB-D* images. In a first phase, segmentation is performed in a 2D domain. Pixels are first divided into boundary pixels and interior pixels based on the amount of 3D points present in the vertical direction. A binary vector feature is then associated to each interior pixel, indicating which of the boundary pixels are visible. k-medoids, starting from an over-segmentation, is then used to cluster interior pixels, using as clustering metric the distance between features. Finally, the initial room segmentation is refined using 3D analysis, merging rooms where, in a further step, sufficiently large openings (e.g., doors) are not detected. The overall method has been recently improved by Chen et al. [CLWF19], which replaces the heuristic room segmentation step by a purely data driven approach in which the collection of input panorama scans are converted into a single 4-channel 256×256 point-density/normal map in a top-down view, which is then segmented into disjoint regions using a Recurrent Convolutional Neural Network (R-CNN) method called Mask R-CNN [HGDG17] (see Fig. 3).

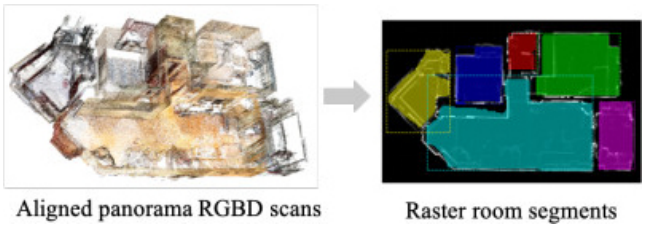


Figure 3: ***Room segmentation from aligned RGB-D scans****. Aligned panorama RGB-D images are turned into a floorplan graph starting from room segments (raster) from a top-down projection image, consisting of point density and mean surface normal. [CLWF19]*

5.3 Discussion

When using dense 3D data, input data partitioning before the application of the reconstruction pipeline has mostly found application as a means to create more scalable solutions. On the other hand, partitioning input images into rooms prior to the application of the pipeline has shown being beneficial to improve both performance and accuracy through the pre-filtering of massive amounts of outliers. Early indoor-specific solutions have been presented [PGJG19], but they are currently limited to very specific settings (panoramic images taken at standardized heights). Creating more general solutions is an interesting open research problem, especially since the grouping of images into per-room collections is a necessary step for visual indoor place recognition [PCJC10]. Performing this grouping early in the pipeline makes it possible to use semantic labeling for further specialized processing steps [FPF07].

Inferring the room partitioning by analysis of 3D data, has instead, attracted much research in the previous years, and current solutions are effective for both fairly dense input 3D data and when applied to post-process the output of a reconstruction pipeline, to separate rooms for further processing (see, in particular, Sec. 8). The vast majority of the methods perform this processing using similarity and visibility reasoning in a top-down 2D domain, extracting a 2D floor plan which is then extruded to 3D in further processing steps (Sec. 6). Since room labeling is often ambiguous in 3D, there have been some early attempts to refine this labeling using 3D structures (e.g., detecting doors) [IYF15]. Such 3D processing is, however, under-developed, and represents a major area of research, especially if one needs to perform labeling in complex multi-floor environments or just in the presence of non-vertical walls.

6 Bounding surfaces reconstruction

While room segmentation (Sec. 5) deals with the problem of decomposing an indoor space into disjoint spaces (e.g., hallways, rooms), the goal of bounding surface reconstruction is to further parse those spaces into the structural elements that bound their geometry (e.g. floor, ceiling, walls, etc.). This task is one of the major challenges in indoor reconstruction, since building interiors are typically cluttered with furniture and other objects. Not only are these elements not relevant to the structural shape of a building, and should therefore considered as outliers for this task, but they also generate viewpoint occlusions resulting in large amounts of missed sampling of the permanent structures. Larger amounts of missed 3D samplings are also present in visual input sources. Thus, generic surface reconstruction approaches are doomed to fail.

In the literature, a large variety of specific methods have been proposed to generate boundary surfaces of indoor spaces under clutter and occlusion (see Table 3). In general, in structured multi-room environments, these techniques are applied after room segmentation, on a room per room basis, and the partial results are then refined and assembled using the techniques presented in Sec. 8, which deal with the problem of optimizing the overall layout and building the connectivity graph. Some methods, however, see in particular Sec. 6.3, perform boundary extraction jointly with segmentation and optimization.

In terms of approaches, a primary distinction mainly depends on the amount and density of 3D information that is available for the analysis, either because it was part of the raw measures (e.g., for point cloud or RGB-D sources), or because it was derived with reasonable approximation by orthogonal techniques (e.g., using multi-view stereo to obtain a set of triangulated scene points).

When no 3D information is explicitly present, as in methods working with a single image per room, or with multiple non-overlapping images per room, the geometry must be inferred by extracting image features, such as corners, edges, and flat uniform regions, and interpreting them as geometry cues in the context of one of the specific

Method	Input type	Input requirements	Output	Priors	Features
Delage et al. [DHN06]	Single RGB	Single pinhole	Floor-wall planes	FW	Vertical-ground fold-lines
Hedau et al. [HHF09]	Single RGB	Single pinhole	Oriented box	CB	Geometric context (GC)
Lee et al. [LHK09]	Single RGB	Single pinhole	IWM planes	IWM	Orientation map (OM)
Furukawa et al. [FCSS09b]	Dense RGB	Multi pinhole	3D mesh; reg. images	MW	VF; FP evidence
Jenke et al. [JHS09]	Dense PC	Two scanners	3D mesh	MW	Cuboids merging
Flint et al. [FMMR10]	Single RGB	Single pinhole	Oriented planes	IWM	C-F homography
Budroni et al. [BB10]	Dense PC	Markers	3D mesh	IWM	Vertical walls via rotational sweep
Flint et al. [FMR11]	Dense RGB	Multi pinhole (video)	Oriented planes	IWM	GR+multi-view features
Turner et al. [TZ12]	Dense PC	Scan positions (per-point)	2D floorplan	VW	Curved walls
Turner et al. [TZ13]	Dense PC	Scan positions	3D mesh	MW	Voxel carving
Bao et al. [BFFFS14]	Dense RGB	Multi pinhole (video)	3D box	CB	GC+OM+multi-view features
PanoContext [ZSTX14]	Single RGB	Single panorama	Oriented box	IWM	GC+OM on panorama
Cabral et al. [CF14]	Sparse RGB	Multi panorama; dense PC	Textured 3D mesh	IWM	C-F homography; FP evidence
Oesau et al. [OLA14]	Dense PC	—	3D mesh	PW	2.5D cell complex
Turner et al. [TZ14]	Dense PC	Scan positions (per-point)	2D floorplan	VW	Triangulation of 2D wall samples
Mura et al. [MMJV*14]	Dense PC	Scan positions	3D mesh	AW	Occlusion-aware; diff. distances
Ikehata et al. [IYF15]	Dense RGB-D	Multi panorama	Structured 3D shape	MW	FS-S evidence
Yang et al. [YZ16]	Single RGB	Single panorama	Oriented 3D facets	MW	GC+OM; 3D facets
Ochmann et al. [OVWK16]	Dense PC	Scan positions (per-point)	3D mesh	AW	Parametric models; thick walls
Mura et al. [MMP16]	Dense PC	Scan positions; oriented points	3D mesh	PW	Fully 3D reconstruction
Pano2CAD [XSKT17]	Single RGB	Single panorama	3D shape	IWM	GC+OM on panorama
Ambrus et al. [ACW17]	Dense RGB-D	—	3D mesh	VW+PW	Artificial scan positions
Mura et al. [MP17]	Dense PC	Oriented points	3D mesh	PW	Artificial scan positions
Murali et al. [MSOP17]	Dense RGB-D	—	3D mesh	MW	Lightweight; cuboids merging
Liu et al. [LWKF17]	Dense RGB-D	Multi panorama; 2D florplan	Labeled 2.5 shape	MW	CNN+IP
Pintore et al. [PPG*18]	Sparse RGB	Single panorama	Textured 3D shape	AW	E2P;C-F homography
FloorNet [LWF18a]	Dense RGB-D	Video	2D floorplan	MW	Hybrid DNN architecture
Pintore et al. [PGP*18]	Sparse RGB	Multi panorama	Structured 3D shape	VW+PW	E2P facets
Yang et al. [YZS*19]	Sparse RGB	Dense point cloud	3D shape	IWM	Curved walls
DuLa-Net [YWP*19]	Single RGB	Single panorama	3D shape	IWM	E2P;C-F homography
HorizonNet [SHSC19]	Single RGB	Single panorama	3D shape	IWM	1D vectors encoding
Ochmann et al. [OVK19]	Dense PC	Oriented points	3D mesh	AW	2.5D cell complex; thick walls
Floor-SP [CLWF19]	Dense RGB-D	Multi panorama	2.5D floorplan	VW	Shortest polygonal loop

Table 3: ***Bounding surfaces reconstruction methods****. Summary of the approaches described in Sec. 6, arranged by chronological order. FW: Floor-Wall model; CB: cuboid; GC: pixel-wise geometric context; OM: pixel-wise orientation map; IWM: Indoor World Model; MW: canonical Manhattan World assumption; AW: Atlanta World assumption; GR: per image geometric reasoning based on IWM; VW: vertical walls; PW: piece-wise planarity; FS evidence: free-space evidence; VF: volumetric fusion; E2P: E2P transform; FS-S evidence: free-space and surface evidence; CNN: convolutional neural network; IP: integer programming. The methods that have multiple priors associated to them use different priors in different stages of the pipeline, as explained in the text.*

priors summarized in Sec. 4.2. The techniques, surveyed in Sec. 6.1, may use either a top-down approach, in which a known model (e.g., a cuboid) is fitted to the features, or a bottom-up approach in which the most plausible model is assembled from the feature set.

When only sparse 3D measurements are available, typically in the case of multi-view approaches that generate only very sparse 3D evidence in untextured indoors, data fusion techniques are exploited to incorporate known 3D data in 2D image analysis, for instance by associating heights or distances to the camera to pixels corresponding to triangulated features and propagating them to neighbors using geometric reasoning. The techniques, surveyed in Sec. 6.2, often impose less restrictive priors than single-image methods, and may also exploit multi-view geometry to perform a global registration, permitting the parallel joint reconstruction of multiple rooms.

Finally, when dense 3D measurements are available, the approach is often qualitatively different from the above, see Sec. 6.3. In most cases, the local shape information in terms of normal and curvature provided by high-density sampling is exploited to convert the dense and massive input point cloud into a manageable number of higher-level and more compact geometric primitives that describe the main surfaces of the environment. This patch-based representation, more expressive and leaner than the original point cloud, is then analyzed to perform boundary detection and reconstruction.

In the following, we analyze the major approaches for each of these different settings, focusing primarily on the extraction of walls, ceilings, and floors.

6.1 Reconstruction without geometric measures as input sources

A noticeable series of works concentrate on parsing the room layout from a single RGB image (see Tab. 3). However, since unconstrained 3D reconstruction from a single-view is essentially an ill-posed problem, the room structure may be uniquely inferred only if sufficient geometrical properties of a scene are known in advance. Based on the fact that man-made interiors often follow very strict rules, several successful approaches have been proposed by imposing one of the priors listed in Sec. 4.2.

Delage et al. [DHN06] presented one the first monocular ap-

proaches to automatically recover a 3D reconstruction from a single indoor image. They adopt a dynamic Bayesian network trained to recognize the *floor-wall* boundary in each column of the image, assuming the indoor scene consists only of a flat floor and straight vertical walls – i.e, the *Floor-Wall* (FW) model. However, in its original formulation, such a reconstruction is limited to partial views (e.g., a room corner).

Full-view *geometric context* (GC) estimation from appearance priors, i.e., the establishment of a correspondence between image pixels and geometric surface labels, was proposed as a method to analyze outdoor scenes by Hoiem et al. [HEH07], then successfully adopted for indoor scenes by many approaches (see Tab. 3).

Hedau et al. [HHF09] model the scene jointly in terms of a 3D box layout and surface labels of pixels, thus imposing the cuboid (CB) prior. The box layout coarsely models the space of the room as if it were empty. The surface labels provide precise localization of the visible object, wall, floor, and ceiling surfaces. The box layout and surface labels are difficult to estimate individually, but each provides cues that inform the other. They first find straight lines in the image and group them into three mutually orthogonal vanishing points, which specify the orientation of the box, providing constraints on its layout. By sampling rays from these vanishing points, many candidates are generated for the box layout, and the confidence for each is estimated using edge-based image features and learned models. The surface labels are then estimated given the most likely box candidate, providing a set of confidence maps from pixels to surfaces [HEH07]. The surface labels, in turn, allow more robust box layout estimation by providing confidence maps for visible surfaces and distinguishing between edges that fall on objects and those that fall on walls, floor, or ceiling. As a result, the box-labels combination that minimizes the error on global perspective cues is chosen.

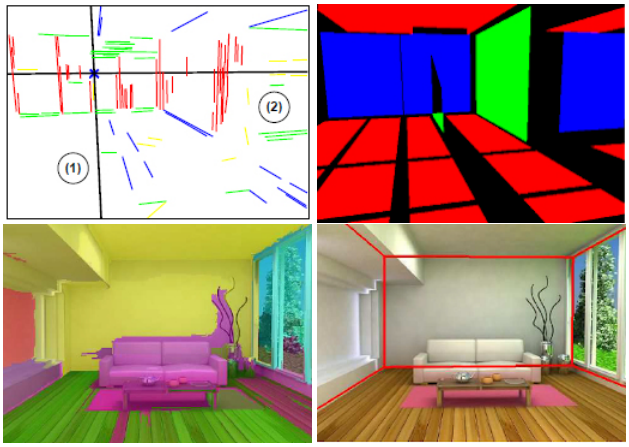


Figure 4: ***Orientation maps and geometric context****. Top left: line segments, vanishing points, and vanishing lines [LHK09, HHF09]. Top right: orientation map [LHK09]; lines segments and regions are colored according to the 3 main Manhattan World orientation. Bottom left: assigning a geometric context (GC) to an image means establishing a correspondence between image pixels and geometric surface labels [HHF09]. Bottom right: resulting 3D box layout from surface labels (GC) [HHF09].*

Lee et al. [LHK09] expands the approach of analysis of line segments by considering the much less constraining *Indoor World Model* (IWM), which combines the *Manhattan World* and *single-floor single-ceiling* priors. This model covers many real-world indoor environments and supports *geometric reasoning*. First of all, it is easy to represent a physically valid model of a scene in two dimensional top-down image space, which can be effortlessly translated into a three dimensional model. Another desirable property is the symmetry that it introduces between the shape of the ceiling and the floor. Evidence to infer building structure from a single image mostly comes from the position of boundaries between planes, but floor-wall boundaries are often occluded by objects. Even in those cases, ceiling-wall boundaries are rarely occluded, so observing ceiling-wall boundaries and assuming symmetry between them allows to infer the location of floor-wall boundaries. In particular, projections of buildings under the Indoor World can be fully represented by corners, so geometric constraints on corners guarantee the entire structure to be valid. For example, the simplest constraint on a corner is that it should consist of two junctions, one above the horizon and one below the horizon. This rule holds because the camera itself is between the floor and the ceiling (see Fig. 4). In the approach of Lee et al. [LHK09], finding the building structure is done simularly to Hedau et al. [HHF09], by first finding line segments and vanishing points, then generating many plausible building model hypotheses, and then ranking the hypotheses according to the match with the prior. In this case, however, hypotheses are generated by connecting line segments to create corners, and connecting corners to create building models, while testing is done against an orientation map (OM), which is a map of local belief of region orientations computed from line segments through heuristic rules. In particular, lines are swept towards vanishing points, and pixels covered by two orthogonal line sweeps are believed to be in a plane orthogonal to both sweeping directions.

Geometric reasoning on the IWM supports several efficient reconstruction methods. A notable example is the work of Flint et al. [FMMR10], who assume that scenes consist of vertical walls with horizontal floor and ceiling, so that the floor is related to the ceiling through a *homography*, and the structure classification problem is reduced to the estimation of the y-coordinate of the ceiling-wall boundary in each image column. The method, based on dynamic programming, has been extended to multi-view approaches [FMR11] (see also Sec. 6.2).

In general, geometric context (GC) [HHF09] and Orientation Map (OM) [LHK09] are in several ways the basis of almost all methods based on geometric reasoning on a single image.

One of the main limitation of single-image methods lies, in fact, on the restricted field of view (FOV) of conventional perspective images, which inevitably results in a limited geometric context [ZSTX14]. With the emergence of consumer-level 360° cameras, a wide indoor context can now be captured with one or at least few shots. As a result, most of the research on reconstruction from sparse imagery is now focused in this direction.

Zhang et al. [ZSTX14] propose a whole-room 3D context model that takes a full-view panorama (e.g., $360^\circ \times 180^\circ$ coverage) as input and outputs a 3D bounding box of the room, also detecting all major objects inside (e.g, *PanoContext*). Their work provides a useful evaluation of how FOV affects room layout recovery. They

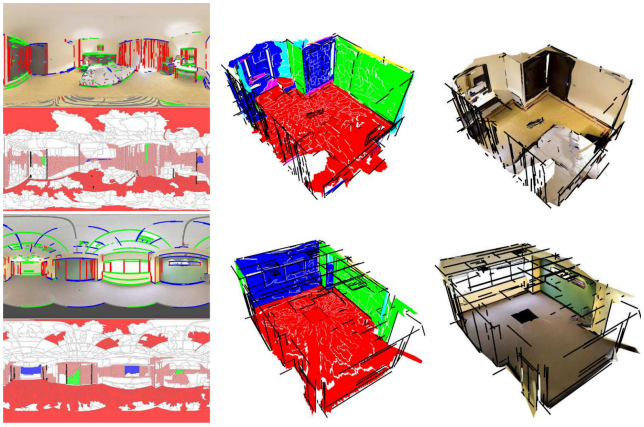


Figure 5: ***3D shape from oriented super-pixel facets and line segments.*** *The leftmost two images show the input panorama (e.g., from the SUN360 dataset [Mas12]), the extracted lines and the super-pixels. Red, green, and blue indicate Manhattan directions that are assigned on lines in the preprocessing stage. Pure colors indicate that the surface normals are restricted to certain Manhattan directions, striped colors suggest that the surface plane should be parallel with a Manhattan direction in 3D space. On the right the resulting 3D bounding surface [YZ16].*

apply respectively orientation map (OM) [LHK09], geometric context (GC) [HHF09] and both of them, observing that GC provides better normal estimation at the bottom, and OM works better at the top half of an image (less cluttered). Combining OM and GC (e.g, OM for the top part and GC for the bottom part) to evaluate the room layout, they demonstrate that by using panoramas, their algorithm significantly outperforms results on regular FOV images. More recently Xu et al. [XSKT17] extend the approach of Zhang et al. [ZSTX14] by assuming IWM instead of a box-shaped room, thus obtaining a more accurate shape of the room.

It should be noted that almost all panoramic approaches based on GC [HHF09] and OM [LHK09] convert the panoramic image into a series of perspective images (e.g, cubemaps), covering the entire sphere with some overlap. The results of GC and OM are then re-projected on the original panorama.

Most of the mentioned approaches work in a discrete manner, that is, the results are selected from a set of candidates based on certain scoring functions. The generation rules of the candidates are often based on the restrictive IWM to ensure a robust reconstruction from all types of images. This however, limits the scope of these algorithms, as in case of different height levels on the ceiling or walls that do not form right angles. Some recent works show that some of these assumptions can be relaxed in the case of panoramic images.

Yang et al. [YZ16] propose an algorithm that, starting from a single full-view panorama, automatically infers a 3D shape from a collection of partially oriented super-pixel facets and line segments. The *Manhattan World* constraint is locally applied to oriented facets and line segments (GC+OM). The core part of the algorithm is a constraint graph, which includes lines and super-pixels as vertices, and encodes their geometric relations as edges. Pintore et al. [PPG*18] tackle the problem of recovering room boundaries in a *top-down 2D* domain, in a manner conceptually similar to that of dense approaches (see Sec. 6.3). They assume indoor structures follow the *Atlanta World* (AW) assumption [SD04], i.e., scenes which can be described by vertical and horizontal planes in 3D. Note that this assumption does not require vertical planes to be orthogonal with respect to each other.

To recover the shape of the room from the single images they combine the ceiling-floor homography [FMMR10] to a spatial transform (E2P - i.e., *equirectangular to perspective*) [PGG*16], based on the *Unified projection model* for spherical images [GD00]. Such E2P transform highlights the shape of the room projected on a 2D floorplan, as illustrated in the example of Fig. 6. Specifically, E2P generates two projections, respectively for the floor and for the ceiling edges. Applying ceiling-floor homography, they recover the height of the walls and enforce the 2D shape estimation from the projected contours.

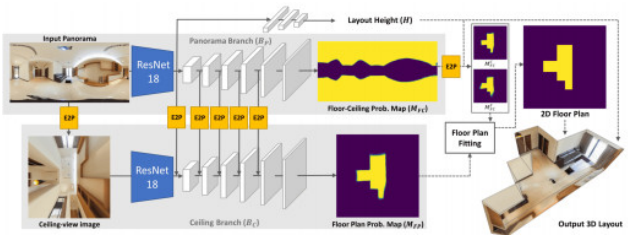


Figure 6: ***E2P transform****. Recent data-driven approaches [YWP*19] exploit the E2P transform to predict 3D room layouts. Such E2P transform highlights the shape of the room projected on a 2D floorplan [PPG*18].*

Recent data-driven approaches [ZCSH18, SHSC19, YWP*19] have also demonstrated success in recovering the 3D boundary of a single uncluttered room meeting the Manhattan World constraint. Yang et al. [YWP*19] exploit the E2P transform and ceiling-floor homography within a deep learning framework, called *DuLa-Net*, to predict IWM 3D room layouts from a single RGB panorama, outperforming other similar approaches, especially in rooms with non-cuboid layouts. Also with the goal of predicting a 3D room layout under IWM, Sun et al. [SHSC19] represent room layout as three 1D vectors that encode, at each image column, the boundary positions of floor-wall and ceiling-wall, and the existence of wall-wall boundary, with a conceptually similar representation to that of Lee et al. [LHK09]. Recently Zou et al. [ZSP*19] have presented an extensive evaluation of these single-view methods, together with their source code and data, detailing the common framework, the variants, and the impact of the design decisions.

These 360° approaches, characterized by an extreme ease of use and capture speed, have evolved over the years, achieving impressive results, and under particular conditions their accuracy competes with measuring instruments [YJL*18]. However, one of the main limitations is that all the corners of the room must be visible from a single point of view; moreover, their application is constrained by heavy priors. These issues make this approach ineffective in many common indoor environments (e.g., hidden corners, multi-room scenes, corridors, sloped ceilings).

6.2 Reconstruction from sparse geometric measures as input sources

Several approaches propose a trade-off between capture simplification and reconstruction scalability. In particular, by acquiring a few

overlapping images, it is possible to exploit SfM to register images in a single reference frame and extract at least a sparse amount of 3D geometric cues. In this context, several authors have introduced methods to obtain labeled planar structures (ceiling, floor, wall) under the IWM prior for scenes larger than one room [FMR11], or to improve accuracy and completeness for highly cluttered small-scale scenes [BFFFS14]. As for the pure single-view methods, the main limitations lie in the limited field-of-view of conventional camera images, which complicates geometric reasoning.

As for single-view analysis, research has recently focused mostly on panoramic images. In general, these techniques exploit multi-view registration to automatically align the reconstruction of multiple rooms, as well as to jointly analyze 2D and 3D data for room reconstruction. Such an analysis can be exploited either to enrich the point cloud coming from triangulation with 3D points inferred from 2D analysis, and then performing the final reconstruction in 3D [FCSS09b, CF14], or by exploiting sparse 3D information coming from feature triangulation to propagate known geometric information to homogeneous areas coming from geometric context analysis of 2D images [PGP*18].

Exploiting the *Manhattan World* assumption, Furukawa et al. [FCSS09b] propose a fully automatic 3D reconstruction and visualization system from RGB images. Given a set of images of a scene, they use Structure from Motion (SfM) and multi-view stereo (MVS) to reconstruct a first sparse set of 3D oriented points. Due to lack of texture, MVS produces incomplete models for most architectural scenes. Therefore, a constrained stereo algorithm [FCSS09a] which exploits the Manhattan-world assumption is then applied. Knowing that surfaces are piece-wise planar and aligned with three dominant directions, a dense depth map is inferred for each input image. Such depth maps are then merged in a 3D mesh of the environment through a volumetric graph-cut approach, exploiting evidence for the emptiness or fullness of voxels based on visibility from camera position.

The approach has been extended by Cabral et al. [CF14], who employ a single-view structure classification method to infer 3D cues from panoramic images (see Fig. 7). The method couples a sparse panoramic coverage of the scene (i.e., one stitching per room) with the semi-dense point cloud computed with the approach of Furukawa et al. [FCSS09b]. A single-view classification labels image super-pixels into three classes (floor,ceiling and wall). For each panoramic image, the top-most row of super-pixels is labeled as *ceiling*, the bottom ones as *floor*, and the ones lying on the image horizon – i.e., middle of the image – are labeled as *wall* (see Fig. 6 top right). Structure labels are propagated by enforcing the label order (i.e., ceiling, wall and floor from top to bottom in each column). Moreover, a homography mapping [FMMR10, FMR11] is exploited: for each pixel with a floor label, the corresponding pixel through homography is labeled as ceiling, if it does not already have a label. In practice, the analysis is focused on regions rather than on edges, under the assumption that this strategy is more robust. Finally, labeled points are converted into a point cloud and merged with the other 3D data, by assuming that an indoor scene is composed of vertical facades and horizontal floor and ceiling (Indoor World Model). From the merged 3D point cloud, they reconstruct the 2D shape of the room by solving a shortest path problem on a specially crafted graph, enforcing it through *piece-wise planarity* assumption. The final 3D mesh is extruded from the 2D plane and textured.

More recently, Pintore et al. [PGP*18] propose a method for geometric context extraction based on a 3D facets representation, which combines color distribution analysis of individual panoramic images with sparse 3D multi-view clues. Their approach works with a small set of partially overlapping 360° images, where 3D cues are calculated from the images registration, without actually involving any external geometric information. Such method imposes *vertical walls* and *piece-wise planarity* for the floor, instead of Manhattan World constrains on vanishing lines [YZ16]. They apply the E2P locally (i.e., piece-wise planarity assumption) to individual super-pixels, allowing to combine facets from different images and enabling the reconstruction of structured and complex environments (e.g., L-shapes, sloped ceilings).

6.3 Reconstruction from dense geometric measures as input sources

Several measurement devices, including LiDAR scanners and RGB-D cameras, are capable of providing dense and reasonably accurate measurements of the seen portion of indoor environments in the form of point clouds (see Sec. 3.1). The size, redundancy and lack of structure of such dense 3D models are aspects that all indoor reconstruction pipelines based on dense three-dimensional input data must deal with. While several solutions have been presented with a specific input device use case (e.g, LiDAR scanner or RGB-D camera), reconstruction methods in the dense case are not necessarily input-dependent, but instead use general data management approaches, such as converting the input point cloud into higher level and more compact geometric primitives that describe the main surfaces of the environment.

Based on the observation that man-made structures are mostly composed of planar parts [BSRVG14], most approaches extract planar patches from the input points. An efficient and robust way of doing this is through the well-known RANSAC algorithm [SWK07], which generates plane hypotheses in a randomized manner and tests how well each of them describes the input data. Typically, the number of hypotheses generated is based on the maximum allowed probability to miss a planar structure. Many indoor modeling pipelines effectively use RANSAC in a pre-processing step, either directly in 3D space [JHS09, OVWK16, MSOP17, OVK19] or in a simplified 2D view of the environment [OLA14], although this can result in missing regions and in non-deterministic results due to its randomized nature. To overcome these issues, some approaches rather opt for a region-growing formulation, in which planar patches are expanded from a set of seed points based on normal deviation and plane offset [MMJV*14, MMP16]. This less robust yet more systematic way of detecting planar patches is common when the input comes from high-quality laser-scanned data [CLP10, BdLGM14].

Regardless of the specific approach used, the detected primitives are often arranged in an *adjacency graph* based on their spatial proximity [JHS09, MMP16, MSOP17]. More advanced pipelines for planar model fitting [OLA16, MMBM15] can potentially be used, although this option is generally discarded in the context of indoor modeling. For more details on methods for primitive

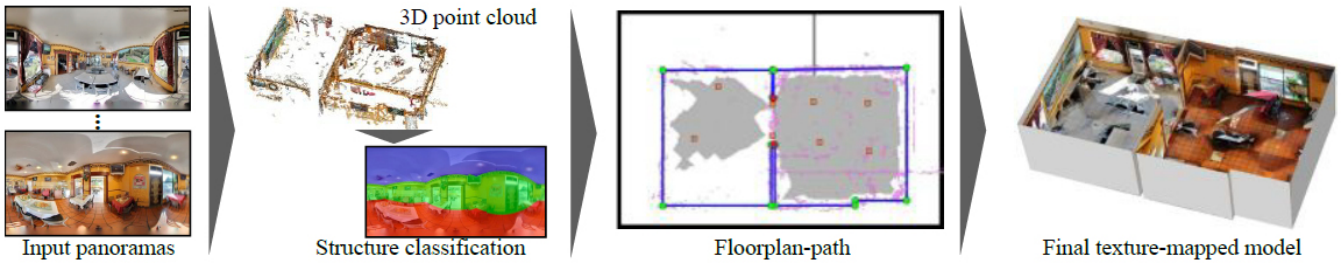


Figure 7: ***Combination of single-view analysis with multi-view cues***. *Cabral et al. [CF14] propose a single-view classification that labels image super-pixels into three classes (floor,ceiling and wall). Labeled points are converted into 3D anchor points to integrate the input point cloud. From the merged 3D point cloud, they reconstruct the 2D shape of the room by solving a shortest path problem on a specially crafted graph, enforcing it through* piece-wise planarity *assumption.*

extraction, we refer the reader to the recent survey by Kaiser and colleagues [KYZB19].

Given a decomposition of the input into simpler geometric primitives, existing approaches differ considerably in the way such primitives are used to extract the boundary surface of the rooms. In general, the complexity of the technical solutions adopted depends on the assumptions made on the quality of the input data and on the shape of the rooms. For relatively clean and complete inputs and under the *Manhattan-World* (MW) prior, rooms can be reconstructed as the union of one or more cuboids, each obtained by intersecting a group of six adjacent wall planes detected as a special configuration in the adjacency graph of primitives [JHS09, MSOP17]. Restricting the cuboids to have equal sides of a pre-determined length results in a voxel-based reconstruction. In this case, it is possible to devise more specialized approaches that work directly on a voxelization of the input scene, for instance by extracting the internal volume of a room by carving out voxels that intersect the lines of sight from scanned points to the positions of the acquisition device (if available in the input model) [TZ13]. This scheme, however, results in a blocky reconstruction of all structures that are not perfectly aligned with the three axes of the Cartesian grid.

Allowing for a more precise extraction of indoor architecture requires the boundaries of the rooms to align with more general directions than those of the three Cartesian axes. Over the last decade, a general scheme based on building a discrete subdivision of the input domain has emerged. Under this scheme, the space surrounding the input model is partitioned into convex regions by computing an *arrangement of hyperplanes* [EOS86] and such regions are organized according to their adjacency relations in a *cell complex* (see Fig. 8). Given the discretized solution space represented by the cell complex, the shape of each room (or of the entire environment) can be obtained as the union of a specific set of cells – those that are contained inside the room bounding walls. Under this general framework, the fundamental questions to be addressed are 1) how to select the hyperplanes so that they correspond to the main architectural surfaces of the environment and 2) how to select the correct set of cells for each room. Existing approaches differ in the technical solutions to these questions, as well as in the domain considered for the construction of the cell complex.

Many researchers assume that the interiors considered have a *2.5D structure*, that is, that they have horizontal floors and ceilings and vertical walls (the *Atlanta World* prior, see Sec. 4). Under this assumption, the 3D model of each room can be easily obtained by vertical extrusion of its 2D footprint and the cell complex can be simply defined as an arrangement of 2D lines on the ground plane of the environment. A number of methods [MMJV*14, OVWK16, ACW17] use such a structure to decrease the computational cost of the reconstruction while keeping the implementation complexity low. One critical aspect that needs to be considered is how to select the 2D lines that define the 2D cells boundaries: since rooms are constructed by aggregation of 2D cells, their boundaries should align to the real-world bounding walls of the rooms. The lines of the arrangement can be defined by projecting downwards all vertical planes (discovered in 3D space) that have a minimum surface area [OVWK16]. To reduce false positives, it is possible to restrict the selection to those vertical structures that span a vertical extent comparable to the ceiling height; in cluttered environments, robustness to view occlusions can be achieved by explicitly analyzing the shadows cast by potential occluders onto the vertical structures [MMJV*14]. Note that this requires that the scan positions are available in the input model. Working fully in 2D, Oesau et al. [OLA14] project the point cloud onto the horizontal plane and detect the points that belong to walls by analyzing their normal vector (which should be orthogonal to the up direction) and their spatial neighborhood in 2D (which should exhibit one clear dominant direction). Regardless of whether walls are detected directly in 2D or in 3D, the corresponding line segments on the horizontal plane are often clustered to obtain more representative lines for the construction of the arrangement [MMJV*14, OLA14].

The vast majority of these methods are based on the detection of planar walls, and handle curved walls as piece-wise linear approximations, typically by inserting each individual representative line segment surviving clustering into the cell complex [ACW17]. However, the use of too many linear segments will lead to too many small cells in the cell complex, increasing the computational and storage burden. Recently, Yang et al. [YZS*19] have proposed, in addition to clustering linear segments, to reduce the number of primitives by using a curved-line fitter on the horizontal plane, so as to decompose the 2D projection into spheres, ellipses, and straight lines, showing significant reduction in cell complex complexity.

While still working in 2D, Turner and colleagues [TZ12, TZ14] restrict the cells to have triangular shapes and construct the complex as a 2D Delaunay triangulation. Only the 2D locations that correspond to wall structures are used to compute the triangulation. To do so, a grid is built on the horizontal plane and at each location of the grid a histogram is built using the heights values of the scanned points that project onto that location. The locations for which the histogram covers a sufficiently high vertical extent (defined by a threshold) form the input of the triangulation.

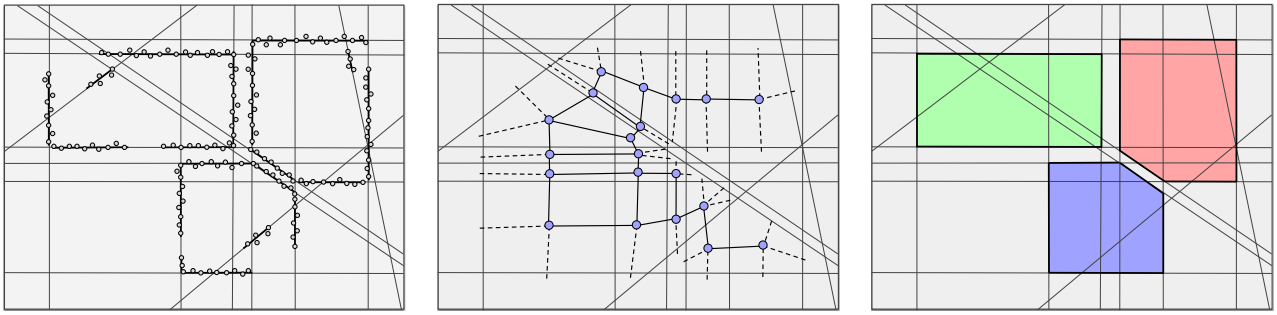


Figure 8: ***Multi-room reconstruction by labeling a 2D cell complex.*** *Working on a top-down view of the environment, a set of 2D line segments that correspond to vertical 3D structures can be extracted from the input measurements (grey dots) and used to define an* arrangement of lines *(left). The polyhedral cells defined by the lines and their adjacency relations define a* 2D cell complex*, which can be conveniently represented through its dual graph (middle). Labeling the cells of the complex based on the room they fall into yields the 2D footprint of the rooms and their boundaries (right, where room labels are encoded by colors).*

A similar 2D reconstruction approach is also employed in recent works exploiting RGB-D input. A prominent example is the work of Ikehata et al. [IYF15], which reconstructs a structured indoor scene by applying grammar rules to a set of registered RGB-D panoramas. Room reconstruction starts with the generation of free-space and point evidence, voxelizing the input space according to Manhattan directions (with the Z-axis denoting the up direction) [FCSS09a], and accumulating for each voxel the number of 3D points inside it (surface evidence) and the number of times it is traversed by a view ray (free space evidence). Room segmentation projects these indicator to the XY plane and applies clustering to find the room floors (Sec. 5.2). The 2D room outline is then found by applying the shortest path algorithm to the pixels contained in the room [CF14], leading to a piecewise planar reconstruction with very few vertices. The 2D outline is then extruded to 3D by estimating floor and ceiling by horizontal plane fitting via RANSAC, below and above the average camera height, respectively.

Such a 2.5D approach is also exploited in recent data-driven methods for room reconstruction. In this context, the FloorNet approach [LWF18a] first converts a set of registered panoramic RGB-D scans into a rasterized floor-plan representation, and then uses a state-of-the-art raster-to-vector transformation [LWKF17] to reconstruct the 3D shape under the Manhattan-world prior. The intermediate rasterized representation consists in a top-down 2D view of the environment, which provides pixel-wise predictions on floorplan geometry and semantics in the form of features computed by parallel neural networks. In particular, for room reconstruction, wall structure is represented by a set of junctions where wall segments meet. There are four wall junction types, I-, L-, T-, and X-shaped, depending on the degrees of incident wall segments. The locations of each junction type are estimated by a 256×256 heatmap in the 2D floorplan image domain. A raster-to-vector converter aggregates detected junctions into a set of simple primitives (e.g., wall lines, door lines, or icon boxes) to produce a vectorized floorplan, while ensuring a topologically and geometrically consistent result. The resulting 2D plan is then extruded to 3D. Excellent results have been demonstrated for this approach, which is capable to combine visual and depth information for joint estimation. However, the method is only applicable to Manhattan World scenes.

The Manhattan World constraint is removed by the approach of Floor-SP [CLWF19], which uses a data-driven approach for room segmentation starting from a top-down point-density/normal map for room segmentation (Sec. 5.2) and, given room segments and the input point-density/normal map, formulates an optimization problem that reconstructs a floorplan graph as multiple polygonal loops, one for each room. The objective function is formed by a data term measuring the discrepancy with the input sensor data over the set of loops, a consistency term encouraging loops to be consistent at the room boundaries (i.e., sharing corners and edges), and a model complexity term, which penalizes long loops. Deep neural networks derive data terms in the objective function from the input point-density/normal map. As presented, the method only produces a 2D floorplan, which can however be extruded to 3D by estimating a floor and ceiling plane from the input point cloud [IYF15].

In order to drop the 2.5D assumption, Mura et al. [MMP16] follow the path taken in urban reconstruction [CLP10] and use a 3D cell complex, thus allowing for arbitrary wall alignments in the output models. In their approach, the bounding box of the input scanned model is repeatedly subdivided by a set of 3D *dominant planes*. Similarly to the 2D lines under the assumption of vertical walls, these planes must snap to the main architectural surfaces of the environment and are obtained by clustering a set of planar patches corresponding to candidate permanent structures. Given the high computational cost for the construction of the 3D complex, particular care is taken to reduce the number of dominant planes to the minimum. To do so, only the planar patches that conform to a set of pre-determined spatial configurations (called *structural patterns*) are used to compute the dominant planes. It is worth noticing that the cells of the complex correspond to generic convex *polyhedra*. Interestingly enough, in contrast to the 2D setting, the use of simplicial complexes has been widely disregarded in the 3D case and has only been object of preliminary exploration [MJM*14].

As a compromise between the 2D and the 3D case, some researchers [OLA14,OVK19] opt for an intermediate solution and construct a 3D complex by vertically stacking multiple 2D complexes, one for each interval between two peaks in the height histogram of the input model. Under the rationale that a peak in this histogram is likely to correspond to a floor or a ceiling, these approaches naturally allow the reconstruction of rooms that have multiple floor/ceiling levels, without incurring in the overhead of building a fully three-dimensional space partitioning. Clearly, this approach can not model structures that have non-vertical and non-horizontal orientations.

Once the cell complex has been constructed, a fundamental point

is how to select the groups of cells that correspond to each room. A natural option is to apply a clustering algorithm to the set of cells, based on a metric that ensures that cells belonging to a same room have low distance. To this purpose, Mura et al. [MMJV*14] (as discussed in Sec. 5.2 for room segmentation), use *diffusion maps* to simulate the diffusion of heat inside the cells of the complex; in this metaphor, heat propagation is slowed down by wall structures, so that it diffuses fast inside the environment and even faster within individual rooms. The distances obtained in this way are used in an iterative clustering process in which room detection and reconstruction are performed jointly: to avoid the need to know the target number of room clusters beforehand, a binary k-medoids clustering is repeatedly performed to separate one room cluster from the rest of the cells not yet assigned.

A major drawback of clustering algorithms is their lack of a controllable regularization term. For this reason, the vast majority of the approaches cast the reconstruction of the rooms from the complex as a labeling problem based on energy minimization. The energy function normally consists of a *data term*, which quantifies the error for assigning a certain label to a cell, and of a *smoothness term*, which penalizes certain labelings of groups of cells (e.g., of pairs of adjacent cells) to favor more regular label assignments. Normally, the data term is based on an "initial guess" on the most likely label, while the smoothness terms is used to avoid jumps in the labeling of neighboring cells, which results in jagged or implausible boundaries.

If one is only interested in reconstructing the environment as a whole (i.e., without extracting the individual rooms), only two labels are needed, one for the inner and one for the outer space. Working in 2D on the top-down view of the environment, the data term for a cell can simply correspond to the fraction of its area covered by input samples (i.e., its *coverage*, see Fig. 9a), as done by Boudroni and Böhm [BB10]. In a similar setting, Oesau et al. [OLA14] use a different approach and compute the data term for each cell by casting visibility rays from its center: the penalty for labeling a cell as outside is then proportional to the fraction of rays that are blocked by scanned geometry (see Fig. 9b). In their pipeline, the smoothness term penalizes assigning different labels to adjacent cells by a factor inversely proportional to the coverage of the facet that separates the two cells - i.e., the proportion of the facet area that is covered by scanned points. The intuition behind this is that the assignment of different labels to adjacent cells should only happen if a physical separation occurs between the two; the presence of scanned data is used as an indication of such a separation. Note that no penalty is set for the assignment of the same label: this results in an energy function that can be optimized efficiently (and exactly) with combinatorial algorithms [BVZ01].

The approaches that reconstruct multiple rooms from the cell complex proceed in a similar fashion, using however a multi-label optimization [OVWK16, MMP16, ACW17, OVK19]. Typically, the number of labels used is an estimate of the number of rooms in the environment, with one additional label reserved for the outer space, as shown in Fig. 8. The data term for each cell is often defined in terms of the parts of the scene that are visible from that cell, either with respect to some representative view positions for each room [MMP16] or based on a visibility-based pre-segmentation of

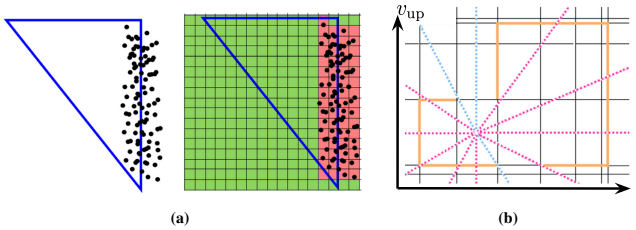


Figure 9: ***Coverage and visibility.*** *Fig.(a): Given a region (e.g., a triangle) in the top-down view of the environment, if only a small proportion of its surface area is occupied (i.e.,* covered*) by input measurements, this region is likely to have been out of the sight of the acquisition device and hence outside of the environment. Fig.(b): For a given location in 3D space, if a high proportion of the visibility rays (shown here as dotted segments) shot from that location hits dense clusters of input measurements (shown as orange segments), the location is likely to be inside the environment. Figures adapted from Oesau et al. [OLA14]*

the input geometry [OVK19]. In some cases, however, an approximate assignment of the input points to each room is performed beforehand, and the proportion of points of each room that fall inside the cell is considered [OVWK16, ACW17]. The definition of the smoothness term follows closely the one proposed by Oesau et al. [OLA14] for the binary labeling; nevertheless, some approaches [MMP16, OVK19] have extended this term to favor structurally consistent room shapes and well-defined boundaries between different rooms. Note that, for the case of multi-label optimization, combinatorial techniques only yield an approximation of the globally optimal solution. For the sake of efficiency, most methods use these schemes, without reporting a noticeable decrease in the quality of the output models [OVWK16, MMP16, ACW17]. A notable exception is the recent work of Ochmann and colleagues [OVK19], who solve the labeling problem using integer programming.

6.4 Discussion

Modeling room boundaries proves to be one of the most challenging aspects in structured indoor reconstruction, since it has to deal with very high levels of noise, clutter, and missing data. For this reason, the proposed solutions make, in general, heavy use of priors to constrain the solution space. In terms of geometry, with very few exceptions, almost all the techniques work totally or partially in 2D or 2.5D, extracting a floor-plan and then extending it to 3D using several techniques, from single-floor/single-ceiling approaches to less constraining works that handle walls of variable elevations. With very few exceptions, planarity of bounding surfaces is also assumed, and curved walls or ceiling are handled as piecewise planar surfaces. Reconstruction of free-form interiors using fully-3D solutions and solutions that handle general curved surfaces is an open research problem, and only few partial solutions have been proposed.

While in the past there was a strict distinction between methods based on purely visual methods, stemming from computer vision, and methods based on dense point clouds, stemming from computer graphics, the boundary between these two areas is more blurred now, due to the emergence of RGB-D cameras. With few exceptions, however, techniques using both RGB and 3D data tend to exploit the two measures to densify the 3D information, and then use methods based on point cloud processing to perform the reconstruction. Techniques that perform full data fusion to exploit reconstruction are just emerging, especially in the context of data-driven techniques that

combine visual and depth hints to extract robust descriptors for further processing (e.g., [LWF18a]). Early results are very encouraging. These techniques, however, are currently limited to 2D solutions; extending them to 3D is a promising avenue of future work.

Besides the advances in data fusion, the recent breakthroughs in data-driven 3D plane estimation from single images [LKG*19, YZ18, LYC*18] promise to further increase the convergence between purely image-based and 3D reconstruction pipelines. In future research, one could combine these techniques with semantic segmentation methods [HGDG17] to extract only the planes corresponding to architectural structures and use such planes as input for a volumetric reconstruction of the room shapes (see Sec. 6.3). This has the potential to allow for piecewise planar, yet fully-3D reconstruction from a sparse set of registered RGB images.

7 Indoor object detection and reconstruction

Modeling objects that occur in indoor scenes is a recurrent problem in computer graphics and computer vision research. In this context, the term *object* refers to a part of the environment that is movable (typically, furniture) and thus does not belong to the architectural structure.

While objects and structures are typically different in the context of 3D reconstruction of structured indoor environments, it is not uncommon to see approaches that have addressed many aspects of indoor scene reconstruction without distinguishing between objects and architectural structures, especially if the target is primarily visual inspection (Sec. 9). This is true both for methods that work directly on 3D input data [TZ13, BdLGM14] and for approaches that extract accurate metric reconstructions from multiple images [BSFC08, SSS08b, FCSS09a, FCSS09b]. Furukawa et al. [FCSS09b], for example, introduce a fully automatic pipeline that reconstructs a whole indoor scene, including objects inside, from a dense set of images (see Sec. 6). In this case objects are modelled by means of full voxels, without however distinguishing between them and the underlying architectural structure. Working on 3D

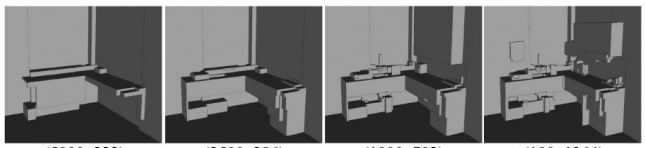


Figure 10: ***Reconstruction without distinguishing between objects and architectural structures****. In this example, objects are modelled by means of full voxels, without however distinguishing between them and the underlying architectural structure [FCSS09b]. Model detail depends on a global parameter (first value in the round brackets), which affects the number of faces of the whole scene (second value between commas) without distinguish between boundaries or clutter detail.*

point clouds as input, Boulch and colleagues [BdLGM14] extract more general piecewise-planar 3D meshes and aim specifically to reduce the geometric complexity of the output model by minimizing the number of edges and corners in it. Even this pipeline, however, includes without distinction movable and permanent structures in the output model.

Many of the approaches, however, distinguish movable and permanent structures. A prominent example is related to techniques that analyze objects as entities detached from the rest of the scene. A typical problem is to detect and classify the parts of sampled 3D data that correspond to objects, for instance using co-occurrences found in large object databases [KMYG12] or in the data themselves [MPM*14]. In this task, the data that is likely to correspond to permanent structure is often ignored or actively discarded [MPM*14]. Even when input sources are purely visual, researchers often focus on extracting objects semantics under the assumption that the underlying geometry is already available [GSEH11], as opposed to leveraging the semantics to improve the reconstruction of the whole scene [IYF15].

In the context of this survey, we are interested in those aspects of indoor object modeling that are integrated in the reconstruction of the entire indoor scene.

In many pipelines, objects are regarded mainly as clutter that should be discarded when reconstructing the architectural shape of the environment (see Sec. 7.1). Other methods, on the other hand, insert the modeling of indoor objects in a *whole indoor geometric context* [LGHK10, ZSTX14] and analyze the geometric properties of the objects and their relationship to the rest of the model in the light of the typical real-world applications (e.g. guidance, energy management, location, routing or content creation for security). For these approaches, we base our discussion on the indoor model paradigm described in Sec. 4 and further distinguish between *3D indoor objects* and *flat indoor objects*.

We regard as 3D indoor object every solid, contained in the free space defined by the walls of the rooms, that has a non-zero finite volume. A prominent example in this class is furniture. The 3D solutions for object detection and reconstruction are summarized in Sec. 7.2. Many fundamental indoor objects, however, are approximately flat and placed on walls, ceilings or floors (e.g., outlets, air-vents, and a wide variety of integrated lighting fixtures). The shape, location, and placement of objects are an integral part of an indoor model; however, due to their placement and flatness, 3D solutions for object detection are generally ineffective in automatically identifying these elements. To suitably augment the whole indoor model, researchers have therefore introduced specialized techniques to detect flat objects in the acquired images (see Sec. 7.3).

7.1 Object detection as clutter removal

Several indoor modeling approaches focus specifically on the accurate modeling of the permanent structures, regarding movable objects as cluttering elements that should be ignored or discarded as early as possible. Arguably, this can be explained with the fact that in the Architecture, Engineering and Construction (AEC) domains – a main driver for the development of structured indoor reconstruction techniques – the focus of the analysis is on the permanent, built structures and a very precise modeling of their *as-built* condition is expected.

Moving from these premises, several researchers – especially those that work on 3D input data – include a first structuring step in their pipeline to discard the input data that is likely to correspond to movable objects. As explained in the discussion on bounding surfaces reconstruction (Sec. 6), this is often done by first detecting

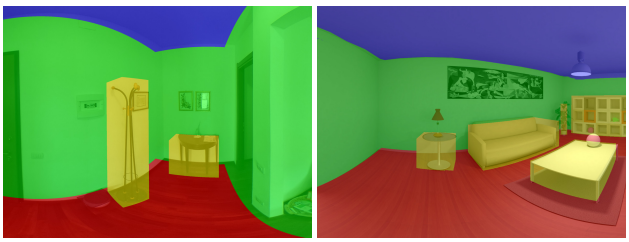
Method	Input type	Input requirements	Output	Priors	Features
Furukawa et al. [FCSS09b]	Dense RGB	Multi pinhole	3D mesh	PW planarity+MW	VF; FP evidence
Lee et al. [LGHK10]	Sparse RGB	Single pinhole	3D boxes	MWL cuboid; LF	3D-3D volumetric reasoning
Hedau et al. [HHF10]	Sparse RGB	Single pinhole	3D boxes	MWL cuboid; LF	3D-2D context reasoning
Shao et al. [SXZ*12]	Sparse RGB-D	Multi pinhole	Models from database	Data-driven; VW	Virtual scans
Nan et al. [NXS12]	Dense PC	Multi pinhole	Models from database	Data-driven	Deform-to-fit reconstruction
Kim et al. [KMYG12]	Dense RGB-D	Multi pinhole	SP composition	Data driven; SP	Local deformation; part relationships
Shen et al. [SFCH12]	Sparse RGB-D	Multi pinhole	Parts connections	Data-driven	Visual/3D data fusion
Del Pero et al. [DBK*13]	Sparse RGB	Single pinhole	Composed boxes	MWL cuboid; SR	Generative models
Schwing et al. [SFPU13]	Sparse RGB	Single pinhole	3D boxes	MWL cuboid; LF	Branch-and-bound strategy
Satkin et al. [SRLH15]	Sparse RGB	Single pinhole	Model from database	MWL room	Data-driven; 3D-2D rendering
Bao et al. [BFFFS14]	Dense RGB	Multi pinhole	Oriented planes	MWL cuboid	GC+OM+MV features
PanoContext [ZSTX14]	Sparse RGB	Single panorama	3D boxes	MWL cuboid; LF	GC+OM; bottom-up strategy
Pano2CAD [XSKT17]	Sparse RGB	Single panorama	Models from database	IWM; LF	Data-driven;top-down strategy
Pintore et al. [PPG*18]	Sparse RGB	Single panorama	Flat objects	AMW	Rectification;query-by-example
Pintore et al. [PGJG19]	Sparse RGB	Multi panorama	3D boxes	PW planarity; LF	PW plane sweeping
3D-SIS [JDN19]	Dense RGB-D	Multi pinhole	3D boxes	Data-driven	Data fusion; CNN

Table 4: ***Reconstruction of 3D objects methods****. We summarize in this table the methods that explicitly return a clutter model. OR: object reasoning; MWL cuboid: object aligned to a Manhattan World layout (recovered from vanishing dominant lines [LHK09, HHF09]); LF: object lying on the floor; MCMC: Markov chain Monte Carlo; SR: fixed size ratio; MWL scene: camera calibrated on a Manahattan World layout; PW: piece-wise; VW: vertical walls; SP: simple primitives (e.g., plane, box, cylinder).*

the main architectural structures and then discarding the remaining data [OVWK16, MSOP17, OVK19] or by analyzing the properties of the 2D neighborhood of the input points projected on the horizontal plane [OLA14, TZ12, TZ14].

A more specific technique is used in the work of Mura and colleagues [MJM*14]. Under the *Atlanta World* (AW) assumption (i.e. vertical walls, horizontal floor and ceiling, see Sec. 4.2), they partition the input data into oriented fitting rectangles and analyze their vertical extent to decide whether they belong to cluttering objects or to permanent structures (in this case, vertical walls). The rationale is that rectangles belonging to walls should span a vertical extent approximately equal to the height of the ceiling. In their work, this step is performed in an occlusion-aware manner, that is, augmenting the measurable vertical extent of a rectangle with an additional height range that was potentially hidden from the line-of-sight of the scanner. Clearly, this approach relies on the availability of the scan positions as input. In a later work [MMP16], the same authors propose another technique for early clutter removal that does not depend on viewpoint information. In this case, the fitting rectangles are arranged in an adjacency graph based on their spatial proximity (see Sec. 6). After detecting the nodes of this graph that belong to floor and ceiling structures, a number of *structural paths* are computed in this graph: these correspond to sequences of nodes that start from a ceiling node and reach a floor node by moving across *structurally sound* edges, i.e. edges whose endpoints correspond to rectangles that reflect one of six valid spatial configurations. These are chosen so that a transition from a permanent structure to an object standing next to it is not allowed. As a result, the rectangles corresponding to the nodes on structural paths are likely to belong to permanent structures and are used in the remaining of the modeling pipeline to reconstruct the final indoor model.

The removal of clutter before reconstruction of the permanent structure is not only done on 3D data, but has proven to also be beneficial in image-based pipelines. For instance, Pintore et al. [PGJG19] propose a pipeline for reconstruction starting from a collection of registered panoramic images. Before performing room boundary reconstruction (Sec. 6), they classify image pixels in each panoramic image into foreground (clutter) and background (wall, ceiling, and floor layout) exploiting a saliency-based approach for single-panorama analysis [YJL*18]. It should be noted that the object masks in individual images do not necessarily describe a complete object shape: since they are obtained by automatic segmentation, they typically contain only salient parts of the object. As a result, for each panoramic image a *mask* containing the pixels from the foreground is extracted, and only pixels part of the background participate in the reconstruction of permanent structures. Image classification and labeling of multi-view 3D features is then enhanced exploiting the recovered clutter models (see Fig. 11). Results demonstrate that early clutter removal is beneficial for the accuracy and stability of the method [PGJG19].



(a) (b)

Figure 11: ***Re-projection of recovered models for clutter removal.*** *Bounding surfaces labeling and reconstruction is improved by exploiting 3D clutter models [PGJG19].*

It is worth stressing that all these methods move from the rationale that object parts are potential *false positives* in the process of extracting permanent structures; in other words, these techniques are designed to maximize the recall of the permanent components detection and not its precision. For this reason, care has to be taken when attempting to reconstruct the objects of the scene from the sole data classified as clutter, as several parts of the objects that do

appear in the input data might be included in the potential evidence for the permanent structures.

7.2 Reconstruction of 3D objects

In many pipelines, objects are regarded as an integral part of the scene to be reconstructed rather than cluttering elements that potentially hinder the modeling process. This is because information on the interior objects (at least in terms of 3D footprint of major indoor objects) is also required for many other use cases, such as guidance, energy management, security, evacuation planning, location awareness and routing [IYF15].

Most of the pipelines dealing with interior objects work on purely visual or mixed geometric and visual input sources.

When modeling indoor scenes from single images, many researchers [LGHK10, HHF10, DBF*12, SFPU13] have looked at the geometric and semantic properties of the entire scene, including the objects located within. Their approaches are based on two strong assumptions: first, the object planes are parallel to the walls; second, the object base touches the floor, as usually true for most furniture in a room [HHF10]).

Lee et al. [LGHK10] propose a method to jointly extract the spatial layout of the room and the configuration of objects in the scene. They model the spatial layout of the room by a 3D box [LHK09] by estimating Manhattan World *dominant directions* from vanishing points [LHK09], then they search inside the room space for solid objects aligned with the room layout. Interactions between clutter and spatial-layout are modeled through a volumetric (3D-3D) approach. Hedau et al. [HHF10] use an appearance-based clutter classifier computing visual features only from the regions classified as *non-clutter*, while they parameterize the 3D structure of the scene by a box (see Fig. 12). As for [LGHK10], they use structured approaches to estimate the best box fitting to the image, however, the modeling of interactions between clutter and spatial-layout of the room is only done in the image plane (3D-2D) and the 3D interactions between room and clutter are not considered.

These early approaches to 3D layout estimation reduce the complexity of the problem by utilizing a small set of candidates. Performance is however limited, as only a small number of hypotheses is considered.

Del Pero et al. [DBF*12, DBK*13] explored generative models and performed inference through Markov chain Monte Carlo (MCMC) to extend the set of candidates. They simultaneously infer the 3D room layout [DBF*12] and integrate composite 3D geometry for objects [DBK*13] in order to achieve more accurate recovery of fine structures. However, such an approach imposes very restrictive priors on camera position (ratio between camera height and room height), room size ratio and objects size ratio (e.g., ratio between object width and length).

Schwing et al. [SFPU13], instead, propose a more accurate solution which reasons about the exponentially many layouts as well as the exponentially many object locations and sizes, exploiting an efficient branch-and-bound strategy.

Satkin et al. [SRLH15] proposed a top-down matching approach to align 3D models from a database with an image. In order to

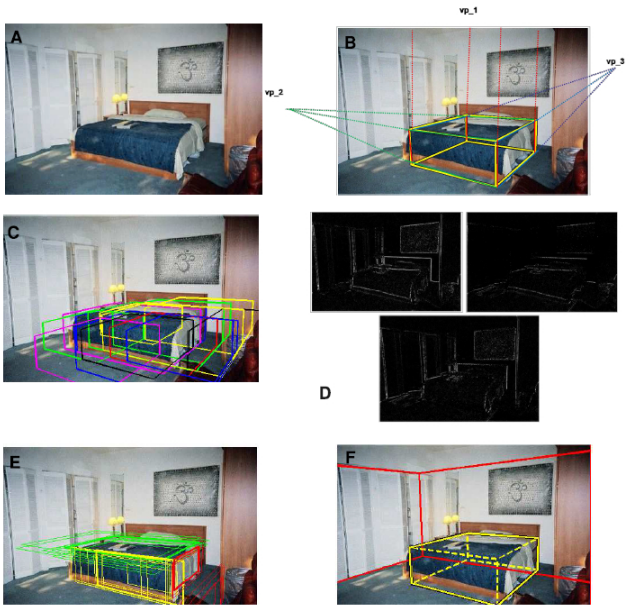


Figure 12: ***Objects aligned to the dominant indoor structure***. Thinking Inside The Box *approach [HHF10] builds tightly constrained models of appearance and interactions of objects in a way that reflects the dominant structure of the indoor scene. The main assumption is that the objects are aligned with the dominant directions of the scene.*

compare models to the scene, they need to render the 3D models on the room layout. To do so they perform camera calibration assuming a box-shaped room aligned to the main Manhattan World directions [LHK09, HHF09]. To match 3D models to images, they exploit pixel-wise object probability, surface normals and image edges as descriptors [LHK09, HHF09]. CNNs have also been used for the same purpose, as in the work by Su et al. [SQLG15], who trained a CNN for pose estimation using rendered models of 12 object categories from the PASCAL 3D dataset [XMS14].

Both the dense multi-view and the single-view methods presented have obvious limitations. The dense methods are based on a predominantly geometric approach that does not integrate the semantic information present in the images [BFFFS14], while the single-view methods suffer from a limited visual context [ZSTX14].

Parallel to the evolution of boundary reconstruction methods (see Sec. 6), indoor object reconstruction has therefore evolved along two main directions: integrating multi-view clues with analysis of individual images, and extending the visual context of images with panoramic captures.

Bao et al. [BFFFS14] apply both single-view and multi-view reasoning to robustly understanding the geometrical and semantic structure of a cluttered room. Their approach is focused on small scenes (i.e., room corners) and requires using a large number of pin-hole images (at least 10 images). As a result they provide 2D image labeling including foreground objects, imposing Manhattan World box prior to generate room hypotheses [LHK09]. The regions labeled as objects can be back-projected into 3D space if they carry sufficient SFM points, thus obtaining a bunch of representative 3D planes for each object.

The work by Zhang et al. [ZSTX14] (i.e., *Panocontext*) shows that

context evidence of an entire room can be captured from panoramic images (see Sec. 6). They learn pairwise object displacements to score their bottom-up object hypotheses from image edges. However, their box-shaped room model does not take relative orientation or distance to walls into account. Xu et al. [XSKT17] extend the approach of Zhang et al. [ZSTX14] by leveraging a *Manhattan World* room layout instead of a box-shaped layout. Object location and pose are estimated using top-down object detection and 3D pose estimation using a public library of 3D models.

Recently Yang et al. [YJL*18] combine geometric cues (lines, vanishing points, orientation map, and surface normals) and semantic cues (saliency and object detection information) to recover a room layout and typical clutter objects from a single panoramic image. As in previous pinhole-image understanding approaches [WGK10], they segment the panorama into layout (background) and clutter (foreground). However, also this method is limited to Manhattan-World environments and does not return object models (e.g., returns the image depth map).

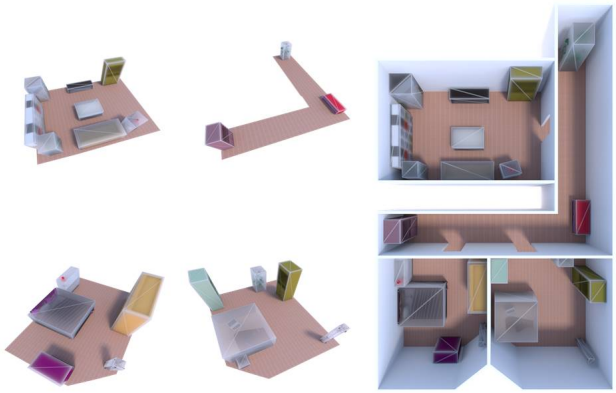


Figure 13: ***Clutter models using virtual plane-sweep reconstruction****. The E2P transform [PPG*18] is applied locally for each single object segmentation and then merged in a plane-sweep fashion. Virtual plane-sweep camera is assumed to be a top-down view [PGJG19].*

Pintore et al. [PGJG19] present an approach to reconstruct structured 3D floorplans with clutter from a small set of panoramic images, by combining sparse multi-view features from images registration and single image analysis (see Fig. 13). At the image level, they follow the same approach of Yang et al. [YJL*18], segmenting the panorama into room structure (background) and objects (foreground). To segment objects in the image they fuse results of state-of-the-art saliency [ZSL*15] and object detection [RF17] algorithms to recover candidate object positions also when objects have unusual shapes or are partially visible. As a result, each panorama image is enriched with the pixel mask of candidate foreground objects. In addition to exploiting these masks for clutter removal (Sec. 7.1), pose and size of each object are recovered using a *virtual* plane-sweeping approach to jointly reason about the content of multiple images, based on piecewise planarity. The particularity of this approach is that it is applied using a virtual camera based on E2P; contrary to the case of rooms boundaries, however, this is applied locally for each portion of segmented object.

While the above approaches perform object recognition on images or inside the reconstruction step, 3D object segmentation can also be performed over the 3D reconstruction of scene geometry, in order to facilitate 3D spatial and structural reasoning [ZXTZ15, XHS*15], at least when a dense input is available. In this context, Hou et al. [JDN19] and Zheng et al. [ZZZ*19] have recently proposed methods for active scene understanding based on online RGB-D reconstruction with volumetric segmentation. In those approaches, a deep neural network is leveraged to perform real-time voxel-based semantic labeling. The network contains a 2D feature extraction module used for extracting 2D features from multi-view RGB images as well as an incremental 3D feature aggregation module specifically designed for real-time inference. The 3D feature fusion and spatial reasoning based on the online updated volume leads to reliable online semantic segmentation, which can be performed at interactive rates.

Figure 14: ***Data-driven object reconstruction****. The* search-classify *approach [NXS12] initially over-segments the scene (top-right) and search-classifies meaningful objects in the scene (bottom-left); these are reconstructed by templates (bottom-right) overcoming the high clutter present in the input scans (top-left). Source images courtesy of Nan et al. [NXS12].*

All the approaches mentioned above mostly strive to identify and locate the objects, and not to reconstruct their precise shape, as this is sufficient for a variety of applications, such as semantic labeling or walkability computation. The solutions that strive, instead, to reconstruct the shape of indoor objects generally exploit data-driven priors, typically in the form of a collection of known shapes to help perform reconstruction (e.g., furniture databases).

Most data-driven solutions look for rigid transformation of objects or object parts retrieved in a database. As a prominent example, Shao et al. [SXZ*12] first semantically segment an input point cloud into potential scans of a single object, which are matched to virtual range scans of the objects in the database in various poses, making the method robust to missing data. Nan et al. [NXS12] extend this approach to retrieval under non-rigid transformation. In their approach (see Fig. 14) the point cloud resulting from a scan of a cluttered indoor environment is first over-segmented into a set of patches, which are iteratively merged together if they have a high confidence on their class label. The models in the database are then non-rigidly deformed to align them with the patches, selecting the best match as the model with the smallest registration residual. The approach is very effective under high amounts of clutter, noise, and occlusion, but tends to work only with small-sized databases, since registration is done in post-processing for each tentative model. At

the same time, since the models are matched in their entirety, only few object types can be supported. Kim et al. [KMYG12] note that in indoor environment it is common to have the same object in multiple poses. For this reason, they learn a deformation model over multiple incomplete scans for a given object, with the goal of identifying objects by incomplete observations of their parts. These parts are detected by following the same over-segmentation, merging and matching approach of previous work [SXZ*12]: they are matched against the learned local deformation modes, and part relationships are used to verify the suitability of a match. While this method introduces a part-by-part matching, each part still originates from a whole model, and combining parts from different models is not possible. Shen et al. [SFCH12] overcome this limitation by starting from a database of segmented models. It is interesting to note that, in this case, a mixed visual/3D input source is exploited, using RGB data to find parts that are completely missing in the 3D scan. The final reconstructed model consists in the union of the matched parts that best cover the input geometry while minimizing the overlap.

Such approaches assume either exact or partial database matches for objects or object parts, which requires, at some point, the creation of a reasonably representative database. A notable exception is the work of Mattausch et al. [MPM*14]: moving from the same premises of Kim et al. [KMYG12], they exploit repeated occurrences of a same object in a large-scale input scene to detect object clusters without the need for an external database. Obviously, this approach is bound to fail for small scenes that do not contain multiple instances of the same objects. To generalize to entirely new shapes in the absence of repeated occurrences, general data-driven structured prediction methods have shown promising results. One of the early successful examples in this area is *Voxlets* [FMAJB16], which uses a random decision forest to predict unknown voxel neighborhoods based on a supervised model trained on a database of volumetric elements. Such a voxelized representation has also been the basis of recent work on dense object reconstruction from sparse or single views [YRM*18].

7.3 Reconstruction of flat objects

The above approaches, which assume that objects have a non-negligible volume, can not work directly on totally or approximately flat objects. For this reason, pipelines that aim to recover them exploit purely visual sources.

The typical approach used when looking for approximately flat objects in sets of images is to apply a pure image-based method. Image-based object localization is a very active area of research, and we refer the reader to a recent survey for a wide coverage [ZYT17]. In our case we are not only interested in the instance detection but especially in mapping the objects on the room boundaries, in terms of location and size with respect to the entire indoor model. In this context standard solutions using images do not work because not integrated with 3D geometric reasoning.

For standard narrow-FOV perspectives, this is done by first detecting *scale-invariant features* (e.g., SIFT) and checking the consistency of the geometric relations of detected 2D points with respect to their supposed 3D position [VZ18]. However, as already discussed in the previous paragraphs, the limited context captured by a perspective image makes it difficult to spatially map a flat object.

In this context, adopting wide-FOV images offers a promising evolution, providing a whole context with very few shots [ZSTX14]. Object detection in spherical images is, however, difficult: since angles and straight lines are not globally kept, objects appear variably stretched, and, as for any image-based approach, the missing metric information leads to the need for variable scale detectors, which increase the number of localization errors. Several works deal with the general problem of object detection in spherical images in different ways, either by converting images with arbitrary projections to standard perspectives [WL09, IDB*10, KRNH11] (e.g., cube maps), or, more specifically, by modifying feature computation to work directly on catadioptric camera cases [CB16].

Pintore et al. [PPG*18] propose an approach that exploits the underlying 3D structure recovered under an Atlanta World model and Structure-from-Motion, first projecting the original panoramic images on the planar patches comprising the room boundaries to remove panoramic distortion, and then performing object detection and localization on the undistorted images using a technique bases on HOG (Histogram of Oriented Gradient) descriptors combined with an Exemplar Support Vector Machine (E-SVM). The recovered object instances are also automatically mapped on the 3D structure (see Fig. 15). The approach provides solid performance on a variety of indoor scenes, such as real-world multiroom environments and various scenes from the publicly available SUN360 dataset [PPG*18].

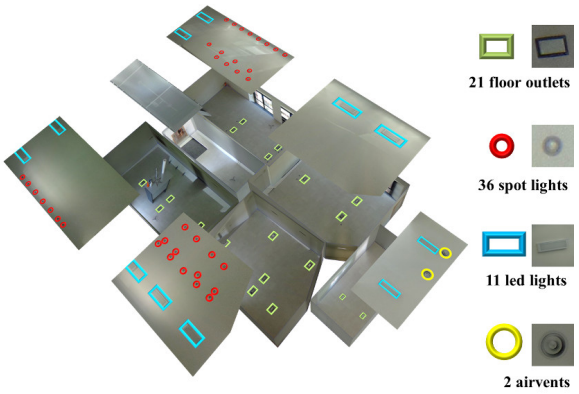


Figure 15: ***Flat objects mapping on the 3D structure****. Multi-room model reconstructed with the detected objects mapped on it [PPG*18]. Ceilings have been moved in the examples to enhance illustration. Beside the models we show the query images (i.e., patches extracted from the processed images), and the number of detected occurrences.*

7.4 Discussion

Existing approaches for the reconstruction of indoor environments vary significantly in the way they approach the modeling of objects. A major distinction can be done based on the nature of the source data considered. When the input is represented by dense, low-noise 3D measurements, most approaches focus on the accurate reconstruction of the architectural structures and conservatively discard all data that can potentially hinder this task, in an attempt to maximize the precision of the modeling process. For instance, including some measurements that stem from movable objects when fitting the geometric primitives of the permanent structures can result in a

final 3D model in which walls are significantly tilted compared to their real-world orientation. In this case, the potential advantages of having highly accurate input data would be lost.

Conversely, pipelines working on purely visual or mixed data have a broader focus. While there have been some attempts to work on object detection to improve accuracy, the goal is not to maximize accuracy, but rather to make pipelines more robust to clutter. This is partly explained by the fact that the reconstruction of a fully three-dimensional model from a (more or less) sparse set of 2D images is an inherently under-constrained problem. As of such, such methods try to encompass a broader scope of use cases for which the reconstruction and correct placement of objects also plays an important role. In this context, prior knowledge is used to help the modeling process, either in the form of assumptions on the shape and positioning of objects (e.g., box-like objects that lie on the floor) or as pre-trained networks that can detect and/or classify regions of the input images, as well as reconstruct detailed object shapes. These priors make it possible to create holistic models of indoor scenes that include both permanent structures and movable or attached objects. Moreover, when properly augmented with visual data (see Sec. 9), such models can be effectively used in a wide variety of applications, including interactive applications such as navigation and virtual exploration.

8 Integrated model computation

The structured reconstruction of a complex environment requires not only the analysis of isolated structures, permanent or not, but also ensure their integration into a coherent structured model.

First of all, the boundary models of the different rooms should be made geometrically and structurally consistent, ensuring for instance that the separating wall boundaries between adjacent rooms are correctly modeled based on the specific output representation of choice (Sec. 8.1). Secondly, as described in Sec. 8.2, most adjacent rooms are connected by doors or large passages that directly reflect the intended functionality of the environment and that should therefore be integrated in its structured representation. Moreover, the structure of a multi-room environment goes beyond the plain geometric description of its rooms and is strongly related to the way such rooms are connected. For this reason, the extraction of a graph that encodes the room interconnections is an important aspect in the modeling of multi-room environments (Sec. 8.3). Last, but not least, most real-world buildings (as well as their interiors) have a multi-story structure; as discussed in Sec. 8.4, this should be correctly captured by the output of the structured modeling process.

8.1 Room consistency

In the context of structured indoor modeling, the exact notion of consistency for a multi-room output varies based on the specific representation adopted: in most pipelines, each room is represented as a polyhedron that adheres to the inner boundary walls, so that adjacent rooms are separated by an explicit representation of the empty space; other approaches ignore this aspect and use "paper-thin" walls with no volumetric extent.

Methods belonging to the first group should ensure that no two polyhedra touch at the boundary. To this purpose, many pipelines that reconstruct all room boundaries in a global, multi-label optimization step require that the labels assigned to any two adjacent cells do not both correspond to room labels - that is, one of them must be the label of the outer space. This is obtained either by using penalty terms [MMP16] or by enforcing hard constraints [OVK19]. An alternative solution is to model walls as regular, thin planes and extrude them along their normal direction to recover their original thickness [OVWK16]. Still, before the extrusion it is necessary to verify that the border separating two rooms corresponds to the presence of scanned physical structures; if this is not the case, the separating border is rather considered an artifact and the corresponding adjacent rooms are merged [OVWK16]. The consistent presence of scanned data on a room boundary is also considered in pipelines that *expect* but do not *enforce* thick dividing walls between rooms [MJM*14].

For approaches that use non-volumetric walls, consistency implies that the thin boundary between adjacent rooms is represented by the same primitive. When room boundaries are defined on a top-down view of the environment, this means that each wall surface that acts as a divider between two adjacent rooms must be represented by a single 2D line segment. In this setting, such properties are often obtained by pairing the solution of a suitable optimization problem with some simple post-processing operations. For instance, Liu et al. [LKF16, LWF18a] extract the corners of wall segments using a data-driven approach and reconstruct room boundaries as 2D loops via integer programming, including specific constraints to enforce consistency. Chen et al. [CLWF19] also extract rooms as 2D loops, using a formulation based on dynamic programming that penalizes the use of separate loop segments to describe shared wall segments. Both pipelines include a rule-based refinement step to snap together corners that are closer than a pre-defined distance.

The need for an explicit room assembly step is even stronger when the shapes of the individual rooms are reconstructed independently. A straightforward solution is to merge rooms that exhibit some spatial overlap, as suggested in a preliminary study for a full-3D setting [MP17]. A similar yet more complete scheme is applied by Pintore and colleagues [PPG*18] to merge the 2D room polygons obtained by processing separately a set of input panoramic images. In particular, they first cluster polygons so that elements of a same cluster overlap by more than a given threshold of their area; then, they compute the union of the polygons in each cluster, regularizing the corners after each merge operation.

8.2 Portals extraction

Doors and windows are important elements of building interiors and their detection has been tackled by many researchers in the context of indoor modeling, often using machine learning and visibility-based cues [XAAH13, OVWK16]. While early work mainly focuses on enriching the geometric detail of the output model, more recent approaches use the presence of a portal - i.e. a door or a large passage embedded in a wall - to assess the correctness of previously detected room separations.

In particular, Ochmann et al. [OVWK16] compute the intersections between a candidate room wall boundary and a set of view rays cast from the scan positions and then analyze clusters of intersecting points on such candidate wall using a SVM classifier. This way, they

reliably distinguish "virtual" walls (i.e. artifacts) from solid walls and from portals, allowing for a correct extraction of the latter. A similar approach is used by Ambrus et al. [ACW17], who do not use visibility information and simply detect openings as clusters of empty regions in vertical planes. In their pipeline, detected openings define separating elements in a top-down view of the environment from which the space occupied by the rooms is extracted.

Ikehata et al. [IYF15] also extract portals as empty regions on wall planes, although in a more principled manner that matches their structured modeling approach. In particular, they analyze pairs of parallel wall planes between two adjacent rooms; for each pair, they extract the rectangle whose projection on such planes encloses the most empty space and the least scanned data. Rectangles that have a minimum area are selected as portals. Interestingly, their detection does not simply increase the robustness of the room detection, but rather serves the purpose of extracting the higher-level structure of the environment, allowing to encode its elements as a graph.

8.3 Room graph computation

For a complex indoor environment, the correct modeling of each room as a separate sub-space represents fundamental information about its structure - in fact, the decomposition of a building into rooms has been proposed as the most basic hierarchical relation for a full top-down parsing of interior scenes [IYF15, AHG*19]. Some of the approaches targeting multi-room environments are restricted to performing such a room-aware reconstruction and either do not determine at all whether and how these are connected [MJM*14, MMP16] or analyze the presence of connecting elements with the specific goal of improving the extraction of their boundaries [OVWK16, ACW17].

The recent work by Ochmann et al. [OVK19] takes one step further and presents an optimization approach for recovering individual rooms by enforcing the presence of volumetric separators between them. While their pipeline does not explicitly reason on the room interconnections, each reconstructed separating element carries information on the two adjacent rooms. From this, one could easily compute a graph of room adjacencies.

Nevertheless, a more complete characterization of the environment can be given by computing the room connection graph explicitly, including also the type of the connections encoded. The work by Ikehata and colleagues [IYF15] is exemplary in this regard. They model an indoor scene as a structure graph, whose nodes correspond to the structural elements of the scene, and propose an associated structure grammar, whose rules define transformations between nodes and come with pre-conditions that must be satisfied in order for the rules to be applied. As illustrated in Fig. 16, the type of the connection between two rooms is classified based on the presence of a portal (see Sec. 8.2) and determines the pre-conditions for two specific rules of the grammar: (1) a "door-addition" rule, which is conditioned on the successful detection of a portal between two rooms and adds a connection edge between the corresponding room nodes in the structure graph; (2) a "room-merging" rule, which is applied if the open passage detected between two rooms is too large to be a portal and which collapses the two room nodes into a single one (also triggering the merging of their geometric representations).

The structure graph proposed by Ikehata et al. [IYF15] models

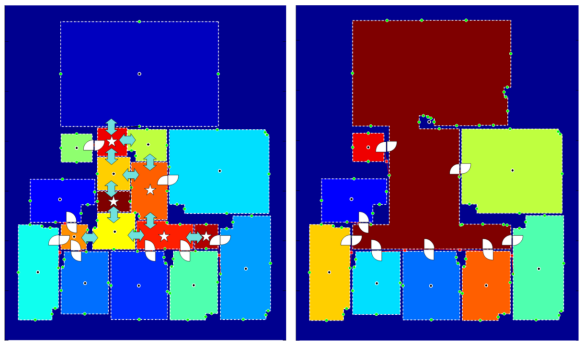


Figure 16: ***Room graph computation using portals.*** *Doors and overly large passages are detected between an initial set of candidate rooms (left). The separation between two rooms is kept in the presence of a door, while rooms that are connected by large passages are merged (right), yielding the correct room connectivity. Figures from [IYF15].*

the room graph in a systematic and semantically rich manner, as it includes only the connections that correspond to actual passages between rooms. However, the creation of these connections is rule-based and does not stem from a globally optimal analysis of the multi-room structure. Modeling the room connections simply in terms of shared walls rather than portals, Chen et al. [CLWF19] propose a formulation that optimizes for a room-based floorplan graph, in which each room is modeled as a 2D boundary loop on a top-down view of the environment and two rooms are considered to be adjacent if they share a portion of their boundary loop. In this approach, the reconstruction of the polygonal loops is based on a preliminary, data-driven detection of wall corner points, which are then connected by 2D boundary segments by solving a sequence of shortest path problems. The solution obtained with this procedure minimizes an objective function that encourages adjacent rooms to share corners and boundary segments, yielding optimal room shapes and connections while at the same time favoring multi-room consistency (see Sec. 8.1).

8.4 Multi-story structure

Besides the structuring into different rooms, another aspect that significantly increases the complexity of an indoor environment is its subdivision into multiple stories.

Most approaches do not consider this problem [XAAH13, TZ13, TZ14, MJM*14, ACW17] or explicitly defer it to future work [IYF15, OVWK16, MSOP17]. However, some pipelines explicitly include a dedicated step to detect individual floor levels. Typically, this is done by analyzing the distribution of the input measurements along the vertical axis of the scene [AH11, TZ12, OLA14]: the peaks of this distribution, which denote large horizontal structures, are likely to correspond to floors and ceilings and can be easily extracted using a 1D mode finding algorithm such as mean-shift [OLA14].

While this technique does not force all rooms in the same story to have the same floor and ceiling height, it does assume that all stories of a building can be completely separated by horizontal cuts; moreover, it fails to detect ceilings with non-horizontal orientation. Aiming at a full-3D modeling, Mura et al. [MMP16] extend this technique to also extract slanted roofs; in particular, they also consider as ceilings those slanted flat structures that do not have other structures above them. However, with the exception of the top story

of a building, all other floor levels are required to not overlap along the vertical direction.

In the recent approach by Ochmann et al. [OVK19] this limitation is overcome by making the reconstruction of the room shapes oblivious of the explicit definition of the floor and ceiling levels. Pairs of parallel and spatially close planes are used to define slices of the 3D space that can correspond to volumetric separations between rooms. The volumetric room shapes are simply obtained by computing an optimal and globally consistent configuration of these separating slices that ensures that each room is fully bounded. This process does not require any notion of floor levels, thus allowing for the extraction of rooms that overlap along the vertical direction; however, all structures are restricted to being either horizontal or vertical, preventing the reconstruction of more generic piecewise-planar structures.

In the case of multi-room environments that span different stories, a promising solution is to perform an early room detection only based on the input data (Sec. 5.1). This makes it possible to reconstruct each room model separately, allowing to handle the case of arbitrary room arrangements in a seamless way. This strategy has been recently proposed for the efficient full-3D reconstruction of large-scale interiors [MP17] and has been effectively exploited to model multi-room environments from sparse visual input under the *Atlanta World* (AW) assumption [PGJG19].

While it allows to recover more flexible spatial arrangements, treating rooms as separate sub-spaces - possibly spread across different stories - opens other important questions. First of all, as already discussed in Sec. 8.1, additional measures must be taken to ensure the consistency of the resulting multi-room assembly. Secondly, it increases the complexity of extracting the room connections, as these can not be assumed to be fully captured by the presence of either a portal or a separating wall. In particular, stairs become an important connecting structure between rooms placed at different height levels and should be therefore appropriately modeled. This poses a non-trivial and largely ill-posed problem, since stairs exhibit a huge variability in structure: in the simplest case, they consist of uniformly sized, aligned steps completely contained within a room; in other cases, they can define entire sub-environments between rooms, as in the case of stairwells. In the first case, the geometric modeling of stairs can be seamlessly included in a piecewise-planar surface reconstruction formulation [BdLGM14] or handled with a dedicated technique [SZ12], possibly defined based on the rules of a grammar for the semantic parsing of building models [BHMT13]. More complex cases, however, have not been addressed systematically by researchers: for this reason, the modeling of stairs both as geometric structures and as connecting elements represents an open question in the context of structured modeling of interiors.

8.5 Discussion

The reconstruction of the different rooms as individual entities has recently emerged as a way of adding basic structure to the model of a complex, multi-room indoor environment. This trait is almost ubiquitous in modern pipelines that consider 3D input data and is becoming increasingly popular among those approaches that operate on pure visual inputs [PPG*18, PGJG19].

Only few of these approaches try to enforce a consistent assembly of the multiple rooms, so that both the separating walls and the connecting entities (i.e. doors) between them are represented in a coherent and complete way. Often, this goal conflicts with the need for a more faithful representation of the geometric structures: for instance, the presence of a volumetric separator between rooms is only favored by a specific term in the full-3D approach by Mura et al. [MMP16], while the pipeline by Ochmann and colleagues [OVK19], which enforces the presence of volumetric walls between rooms via hard constraints, can only reconstruct 2.5D room shapes. A similar argument holds for the work by Chen et al. [CLWF19], which jointly optimizes for room shapes and connections, but does so on a 2D top-down view of the scene. How to enforce room consistency in a globally optimal sense without resorting to restrictive structural priors is an open and very important research question.

Interestingly, while doors and portals represent fundamental evidence of room connections, very few pipelines [IYF15] have exploited their detection in a systematic way to recover the structure of the environment. Another unexplored aspect is how to consistently reconstruct multi-story interiors that do not exhibit a clear-cut separation between different levels. In such settings, recognizing and modeling the connections between rooms at different height levels represents a particularly ambiguous problem, which could be solved effectively using data-driven approaches.

9 Visual representation generation

As discussed in the previous sections, the outcome of reconstruction algorithms combines in a single consistent structured model an approximate topological and geometrical representation derived from an input source, be it visual, purely geometric, or multi-modal. However, a geometric and topological description may not be enough for the applications that should ultimately visualize the reconstructed model. It is therefore necessary to enrich the structured representation with information geared towards visual representation.

In our context, generating visual representations translates into two different problems: the improvement of appearance of reconstructed models with additional geometric and visual data, and the generation of structures to support exploration and navigation.

Improving the appearance of reconstructed models can be achieved either by refining the color or by refining the geometry. In particular, one way this goal has been pursued is by exploiting visual sources by trying to map color and/or reflectance information to the approximate geometry coming out of structured reconstruction. These *texturing* approaches are discussed in Sec. 9.1 and summarized in Table 5. Another way has been to use alternative representations in place of poorly sampled geometry, which is particularly suitable when, like for indoor environments, the domain of shapes is limited in number and types. These *geometry refinement* approaches are discussed in Sec. 9.2 and summarized in Table 6.

Providing support for visualizing/exploring the dataset, instead, has especially been tackled in the context of applications that link the structured reconstruction to the original data. This problem has two important characteristics that make it different from a generic 3D visualization task. First, the data to be visualized may have several forms. Second, the representation of an indoor environment

Method	Output Type	Technique	Occlusions	Target
Cabral et al. [CF14]	textured geometry	closest image projection	NO	generic
Pintore et al. [PGGS16b]	VD textured geometry	image projection	NO	generic
Turner et al [TCZ15]	textured geometry	projection and blending optimization	NO	generic
Xiao et al. [XF14]	textured geometry	confines seams at the natural borders	NO	large scale
Lee et al. [LRY*16]	textured geometry	projection and blending optimization	NO	generic
Agarwala et al. [ACH*13]	VD textured geometry	projection and blending optimization	YES	editing
Zhang et al. [ZCC16]	reflectance mapped geometry	geometry segmentation	YES	editing
Huang et al. [HDGN17]	textured geometry	blending optimization and sharpness enhancement	YES	VR/games
Chen et al [CXY*15]	textured geometry	data driven colorization	YES	generic
Zhu et al. [ZGM18]	textured geometry	data driven colorization	YES	generic

Table 5: ***Visual representation generation by texturing***. *Summary of the approaches to texturing presented in Sec. 9.1*

is typically intended as the view from a walking observer inside the environment, hence it is limited both in terms of viewpoints and, consequently, of human-computer interaction required. These approaches are discussed in Sec. 9.3.

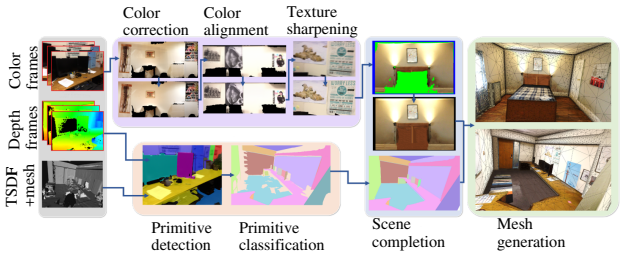


Figure 17: ***The 3Dlite approach to visual model generation***. *The method takes a set of RGB-D frames as input, which contain holes, noisy geometry, and blurry surface textures. From this set, it computes a primitive abstraction that is used to optimize for sharp surface textures and infer missing scene parts. In the end, a low-polygonal lightweight 3D reconstruction is obtained. Pipeline from [HDGN17]*

9.1 Texturing

Textured geometry is a powerful way to obtain a compelling visualization of 3D data. This is especially true in the case of indoor environments, where the geometric detail is often poor. However, texture mapping an indoor environment presents a few peculiarities with respect to a generic 3D object. First, the images are typically photographs in an uncontrolled light environment taken during the 3D acquisition, and may in fact be the only source of data. Second, image-to-geometry correspondences may be scarce due to the lack of detail (both in the image and in the geometry). Finally, many methods only aim to reconstruct the bounding surfaces, that is, excluding what is contained in the rooms and hence in the images. Cabral et al. [CF14] perform bounding surface reconstruction from panoramic images and assign a texture quad to each room facade. The color of each texel of the texture is taken by the projection of the closest panorama image. The same approach is used by Pintore et al. [PGGS16b], with the only difference that a single texture is created for the entire vertical boundary (plus two others for the floor and the ceiling) and that they use a single panorama for each room. Turner et al. [TCZ15] account for the fact that pose estimation and surface reconstruction are both prone to inaccuracies. In a first step, they look for matching edges between geometry and images and use them to refine the camera poses. Then, feature matching among images is used to formulate a minimization problem over the SIFT features reprojection error. Xiao et al. [XF14] also formulate a minimization problem but in a discrete fashion, by partitioning the geometry in floors, ceilings and walls and solving the texture mapping separately. The rationale behind their choice is that mapping artifacts are less noticeable on the natural seams of the environment (that is, the corners between adjacent walls) and that the approach is local and thus easily scalable to large datasets. The actual blending of images relies on the graph cut optimization proposed in [SSS*08a]. A similar approach is taken by Lee et al. [LRY*16]. In this case the geometry is partitioned in generic *surfaces*. The problem is cast as a labelling problem where each surface is associated with a single image. Their cost function incorporates surface visibility from the camera position, distance and inclination, plus a smoothing term to favor image continuity between neighbor surfaces.

Zhang et al. [ZCC16] proposed a system to virtually empty and refurnish a room. Starting from a generic dense mesh of the environment and a large set of aligned images obtained by a Google Tango enabled device, they used the LDR input images to create an HDR radiance map stored *per vertex* on the mesh, under the assumption of diffuse materials. The mesh is analyzed to recover walls, ceilings and furniture and the latter is removed from the dataset; then, morphological dilation/erosion operations are used to define the radiance value for the regions occluded in the input images. Finding the position of the emitters in a semi-automatic fashion and modelling a set of emitter types, they can solve the inverse rendering problem and finally relight the input scene with modified furniture. Huang et al. [HDGN17] start with the same input but favor the production of high quality texture mapped geometry over faithful correspondence to the real environment (see Fig. 17). They use a plane detection strategy over a selection of frame depths. The planes are then merged hierarchically minimizing a quadric error metric. A further refinement step encourages a Manhattan World arrangement of the found planes. Texture mapping is tackled in two main phases. The first phase is devoted to obtaining a color- and geometry-consistent projection of the image onto the segmented geometry. This phase requires the refinement of the geometry segmentation based on the original RGB-D images and the refinement of camera poses to minimize image features correspondences and color consistency across images (obtained in a way similar to the approach of Zhang et al. [ZCC16]). The second phase consists in computing the texture. This is achieved by formulating a graph-cut [BVZ01], based energy optimization that aims associating to each point the sharpest images, trying to preserve continuity in image association (that is, discouraging the association of neighboring points to differ-

ent images). In order to ensure continuity at region transitions, the final sharp solution is blended with the color coming from averaging all views. Finally, unseen regions (i.e., points that do not correspond to any image), are completed through image inpainting, following *Image Melding* [DSB*12].

When the goal of the reconstruction is to virtually edit the model, static image to geometry projection may be insufficient. Agarwala et al. [ACH*13] developed a system for making virtual modifications to an indoor environment, for example enlarging a door or removing a column. The editing is image-based, meaning that the 3D environment is modified from the original views. Even so, if a column is removed, the geometry previously occluded needs to be shown with color. The authors create a view-dependent texture atlas (VDTA) using the available images so that all the geometry in the frustum of the current view is textured. The advantage of view-dependent texture is that the process of source selection and color blending is optimized for the single view. The authors also use a global view-independent texture atlas to show the geometry during transitions among different views.

An alternative to using actual images of the scene consists of just coloring the scene components in a sensible way. Chen et al [CXY*15] proposed a data-driven approach to assign color to the 3D elements of the scene by learning a set of *aesthetic rules* from a database of labelled images of interiors. Zhu et al. [ZGM18] extended this approach by including 3D models of furniture in the database and established hierarchical correspondences between each model and the images of the same sort of furniture. This association is then used to aid 3D input model segmentation. These and similar solutions assume a scene-object-part 3-level hierarchy and are presented in the context of modelled scenes.

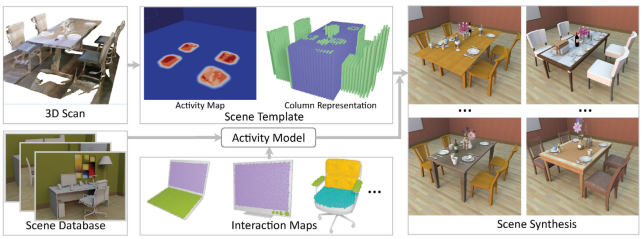


Figure 18: ***Activity-centric scene synthesis for functional 3D scene modeling.*** *A scene consistent with the rough geometry and the detected activity is generated using a mapping learned from examples. Pipeline from [FSL*15]*

9.2 Geometric refinement

The reconstruction of geometric details of architectural surfaces from images or from occluded and poorly sampled scans is an established research area in outdoor modeling [MWA*13]. Typical solutions exploit prior knowledge of architectural shapes and their symmetries, e.g., through the application of shape grammars [MVG13]. In indoor modeling, however, the reconstruction of geometric details of room boundary surfaces is less developed, and often takes the form of the computation of displacement maps on top of the planar walls coming out of boundary reconstruction [IYF15]. The solutions in the presence of poorly sampled surfaces are typically obtained by piecewise planar depthmap algorithms, which strive to fit sparse measures with a low number of different planes [SSS09, FCSS09a, GFP10]. The limited developments in this area with respect to outdoor solutions are motivated by the fact that indoor surfaces are at a much smaller scale of building facades and have little decoration. Geometric refinement for visual display has thus mostly concentrated on indoor objects.

As discussed in Sec. 7.2, a model of the boundary surface and a rough identification of the footprint or a box-like description of major indoor objects is enough for most practical applications [IYF15]. However, when the target of the reconstruction includes a convincing visualization, using boxes in place of indoor objects is obviously unsatisfactory [JGSC13]. Given the restricted domain of indoor furniture, one popular option is to segment and label the scene and then to replace the furniture elements with their corresponding models from a database. This is done in several of the papers referred in Sec. 7.2, specifically in those methods that try to fit portions of the data to template models [SXZ*12, NXS12, SRLH15].

Another usable approach, in the same spirit of the colorization approaches mentioned in Sec. 9.1, consists in generating plausible interiors, which may reflect only partially the actual location and shape of the objects but provide a final output consistent with the type of room. The solutions in this scope revolve around the idea of exploiting the relations between objects in order to sample a domain of plausible arrangements. In the context of indoor reconstruction, such approaches are promising as a way to cope with severely undersampled environments.

A number of solutions exploit user input for generating interior arrangements. Merrel et al. [MSL*11] used predefined guidelines to turn a sketched user input into a set of configurations. Xu et al. [XCF*13] also take user input but learn the relations in an offline phase on a database of complete scenes, while Kermani et al. [KLTZ16] use the annotated Sun RGB-D dataset and ask the user to specify what objects should be included in the room, then include them and possibly add others to create an interior consistent with the type and size of the room. Several authors, instead, have proposed more automated solutions. Fisher et al. [FSL*15] use an annotated database with more than 1000 indoor scenes to learn the relation between the activities that can take place in a scene and the objects involved, within their arrangement in the room. Then, given a 3D scan, they infer an *activity map* of the room and generate a scene consistent with the rough geometry and the detected activity (see Fig. 18). Fu et al. [FCW*17] also infer object-activity relations (from a database of 2D floor plan) and generate interiors starting from user-based definition of elements to be included.

9.3 Visualization methods

The general topic of visualization methods for indoor walkthroughs is vast; many integrated solutions for the presentation of input data and/or reconstructed models have been proposed that combine massive model rendering [GKY08], specialized visibility culling [COCSD03], and interactive virtual camera controllers [CON08]. In the context of this survey, we focus solely on the aspects closely related to reconstruction from data that present some peculiar characteristics. In particular, specific approaches have been proposed to exploit the link between original data (images and/or scans) and reconstructed model.

An intuitive visualization of interior spaces is easily provided

Method	Input Type	Output Type	Technique
Jia et al. [JGSC13]	RGB-D images	3D boxes	segmentation and fitting
Sinha et al. [SSS09]	RGB images	piecewise planar surfaces	plane fitting and graph-cut
Furukawa et al. [FCSS09a]	RGB images	MW piecewise planar surfaces	MW compliant minimization
Gallup et al. [GFP10]	RGB images and depthmaps	planar & non-planar surfaces	RANSAC and graph-cut
Merrel et al. [MSL*11]	user choice	3D model	guidelines driven constraints
Xu et al. [XCF*13]	user sketch	3D model	retrieval from a database of labelled scenes
Kermani et al. [KLTZ16]	user choice	3D model	learning relations from SUN 3D RGB-D dataset
Fisher et al. [FSL*15]	RGB-D images	3D model	learning activity-object relation, querying by 3D similarity
Fu et al. [FCW*17]	user choice	3D model	learning activity-object relation

Table 6: ***Visual representation generation by geometric refinement***. *Summary of the approaches to geometric refinement presented in Sec. 9.2*

by allowing the user to wander in the virtual scene as if she was in its real counterpart. Countless examples of this approach are given by first person shooter video games. However, when the goal is only to view the model, a precomputed set of viewpoints can simplify human-computer interaction. Moreover, if the views are not synthetic but acquired, the underlying geometric model can be simplified, since location recognition is ensured by the photorealism or original images. A prominent example is TUMViewer [Nav12], which allows the user to explore the environment from a fixed set of viewpoints. Each viewpoint corresponds to a 360° panorama image, so that the user can look in every direction, and the viewpoints in the proximity of the current one are shown as spheres. Clicking on one of such spheres brings the user to the corresponding panoramic image. The transition between panoramas is obtained with a zoom-in effect centered in the simulated direction of movement, followed by a fade between current and next panorama. A clickable map of the whole floorplan with all the points of view is also shown. This approach does not use a 3D description of the environment but only the position of the viewpoints. As a consequence, viewpoints that would not be visible in the real environment because of some wall can instead be seen and reached. Stroila et al. [SYMA12] propose to overlay arrows on the panorama view, indicating the next reachable position and accounting for the geometry of the environment. Di Benedetto et al. [DBGBR*14] also use a set of panorama images and arrows that show the reachable position given a current one. In their work, the 3D model is used in a preprocessing phase to create a panorama image for each viewpoint and a panoramic video for each transition between nearby viewpoints. In this manner the visualization relies only on images and videos, which are created offline with any wanted degree of photorealism. The same approach is used by Pintore et al. [PGGS16b] on models created by few panorama images. Their approach is a mix of image-based rendering, that is, the direct use of acquired panoramas for the viewpoints, and textured geometry for the transition between views. The work by Sankar et al. [SS12] is also entirely based on images and videos, but the transition videos are the video the user creates when moving from a room to the next. Google inc. is also using panoramic images and transitions for indoor mapping in the same way as they do for outdoor environments [AB12].

Matterport [Mat17] is a prominent example of user interface that exploits both the reconstructed model and a list of selected viewpoints. The user is offered the choice of exploring the model using three different modes. In the dollhouse mode, the camera zooms out and shows the 3D model from outside, allowing rotation along any axis to see it from any perspective. In the inside view, the camera is moved inside the model, and the user interface supports a walk-through of the space, moving room-to-room and constraining the camera to the fixed set of captured viewpoints, using transitions as in previously mentioned work. In the floor-plan view, the camera looks down from above, looking at the floor plan by removing the roof. A floor selector identifies which floor is visualized.

9.4 Discussion

The reconstruction of indoor environments is nowadays a problem with many solid solutions which vary in terms of desired output and acquisition technologies. However, translating the shapes in a production-ready 3D model, with the proper color information and level of detail still seems a not entirely solved problem. One of the basic issues is the difference in precision between recovery of structure and recovery of structured representation. On the one hand, recovering structured models with an accuracy in the order of several centimeters is indeed a sensible choice, since such models typically lack geometric features and exhibit scales that ranges over several meters. On the other hand, this has important drawbacks for the generation of visual models.

First, photographs of the scene do not correspond exactly to the reconstructed geometry, which means that there is no exact camera pose to be found and classic images-to-geometry projection falls short. Most of the recent approaches to this problem, thus, propose algorithms that pick the color from the images by trying to enforce photoconsistency on the geometry and/or image edges [TCZ15, XF14]. Notably, most of the work is focused on texturing the boundary, in scenarios where the furniture is removed or ignored [CF14, PGGS16a].

Secondly, the elements of the furniture are neither sampled with enough density nor covered in all their surface to produce a detailed 3D model. Several methods, thus, include a segmentation phase where they try to understand the type of the elements in the environment (chair, table etc.) so that their partial sampling can be completed by exploiting symmetries and fitting with models in databases. Other approaches, more devoted to design than to reconstruction, propose procedural solutions to create the interior of a room from little user input [MSL*11, XCF*13] or even from scratch [FCW*17]. All of these methods harness the learned knowledge of the relations between the elements of the interior and create an environment that makes sense w.r.t. to its functionality. Among those approaches, the one by Fisher et al. [FSL*15] exemplifies an important trend in the area, as it shows the ability to cope with severely corrupted inputs, in their case RGB-D scans, by synthesizing consistent models using a mapping learned by examples.

10 Conclusion

The area of structured indoor reconstruction has witnessed substantial progress in the past decade, growing from methods handling small-scale single-room simple environments, to techniques that handle substantial artifacts and produce high-level structured models of large-scale complex multi-room buildings. Our survey has provided an integrative view into this wide array of methods, highlighting strengths and limitations that currently exist in the field. On the basis of this analysis, we provide a view of open problems and current and future works.

Less constraining priors. Capturing the huge variability of 3D shapes in real-world interiors while working on limited input information – possibly only two-dimensional – is an inherently ill-posed problem. Researchers have used a variety of priors (see Sec. 4) to constrain the architectural shapes that can be captured and allow for a robust and efficient reconstruction. In particular, the presence of planar surfaces is assumed in almost all existing approaches and is typically coupled with further limitations on the orientations of wall and ceiling structures (e.g., vertical walls, horizontal ceilings). Very few methods target curved surfaces [YZS*19] or even just planar ones with arbitrary orientations [MMP16]; this is due to the high computational cost of the technical solutions required to handle these cases, as well as to the increasing ambiguity that they bring to the reconstruction process. Nevertheless, the availability of improved acquisition systems capable of providing richer and cleaner inputs, together with the development of more powerful data processing techniques (e.g., based on data-driven approaches), promise to reduce the need for restrictive priors and open the way for effective free-form 3D modeling of interiors.

Global large-scale solutions. Over the last decades, methods for the modeling of interiors have progressively evolved from simple sequences of rule-based processing steps (largely driven by heuristics) to more complex pipelines that include elaborate optimization-based techniques. This has contributed to increasing significantly both the robustness to defects in the input data (e.g., outliers, missing data) and the correctness of the output models, for instance by enforcing consistent separations between individual rooms (see Sec. 8). Still, a unified formulation of indoor modeling that considers all the different aspects of the problem while admitting a globally optimal solution is far from reach. Some recent work goes in this direction [CLWF19], although still relying on fairly restrictive priors on the architectural shapes considered. Further research is needed to develop global solutions that faithfully model the input environment while providing the solid optimality guarantees achieved in other scenarios (e.g., the extraction of planar primitives from raw data [MMBM15]). This remains a non-trivial challenge especially for large-scale scenes, for which the sheer input size makes the use of many optimization techniques computationally unfeasible.

Data fusion. One of the major trends in recent years has been the emergence and consolidation of fast and practical multi-modal acquisition techniques in professional and consumer markets [LKN*17, CLH15]. Input sources contain both color and shape information. Such representations are thus becoming the dominant representation for reconstruction (see Sec. 6) and visualization applications (see Sec. 9). With few exceptions, however, RGB and 3D data are generally analyzed separately in the reconstruction process, most of the time exploiting RGB analysis for 3D data densification prior to the application of a pure geometric processing pipeline. Performing data fusion to combine visual and depth cues into multi-modal feature descriptors on which to base further analysis is an important avenue for future work: such a joint analysis allows to better cope with heavily cluttered and partial acquisitions, as demonstrated by early results on boundary surface reconstruction [LWF18a] and indoor object reconstruction [SFCH12, ZZZ*19, JDN19].

Reconstruction with commodity cameras. As seen in Sec. 3.1, indoor reconstruction from purely visual input is highly ambiguous; hence, photo cameras should, in principle, be the less appropriate devices for recovering accurate structured indoor models. However, in parallel with the exploitation of acquisition systems providing some form of 3D measurement, an increasing number of applications are turning towards the use of the most widespread and low-cost consumer-level visual capture devices, in particular mobile phones used by casual users [SS12, PAG14, DZZ*19]. This direction is, in particular, motivated by the fact that individual users can map and share virtual representation of their homes by using a very familiar device and without the need of providing physical access to external personnel, solving privacy issues and fueling the diffusion of important applications such as real-estate or interior design. Current visual-based solutions, however, are either heavily relying on very restrictive priors or require considerable manual intervention (see, in particular, Sec. 6.1). Improving reconstruction from visual input is definitely an important area for future research. Promising approaches in this context are the usage of data-driven priors trained on very large online collections [YXL*19], as well as the replacement of off-line optimization with smart online algorithms that guide the user towards under-sampled areas and/or exploit user input to resolve ambiguities.

Data-driven approaches. Recent years have seen an extraordinary development of data-driven methods, which was largely fostered by the increased availability of large collections of data to be used in the training process. These methods effectively learn hidden relations from the available data, thus obtaining prior knowledge that can be leveraged to increase robustness when processing corrupted inputs and to extrapolate missing information from incomplete data. These capabilities have also been exploited for the modeling of interiors, leading for instance to methods that can reconstruct the 3D layout of a room from a single image [SHSC19, YWP*19]. In general, learning-based techniques are becoming increasingly popular to model indoor environments from pure visual data, while their application to the processing of 3D inputs has not been fully explored yet. This represents a very promising direction for research, especially given the increasing availability of a wide number of open real-world and synthetic datasets (see Sec. 3.2). The generation of large-scale annotated datasets for the specific purpose of structured modeling, especially for complex multi-floor environments, also stands as an important endeavor that can lead to a significant breakthrough in the field of structured indoor modeling.

Acknowledgments. This work has received funding from Sardinian Regional Authorities under projects VIGECLAB, AMAC, and TDM (POR FESR 2014-2020 Action 1.2.2). We also acknowledge the contribution of the European Union's H2020 research and innovation programme under grant agreements 813170 (EVOCATION) and 820434 (ENCORE).

References

[3DV99] 3DVISTA: 3DVista:Professional Virtual Tour software. https://www.3dvista.com, 1999. 2

[AB12] ALY M., BOUGUET J.-Y.: Street view goes indoors: Automatic pose estimation from uncalibrated unordered spherical panoramas. In *Proc. WACV* (2012), pp. 1–8. 27

[ACH*13] AGARWALA A., COLBURN A., HERTZMANN A., CURLESS B., COHEN M. F.: Image-based remodeling. *IEEE TVCG 19*, 01 (2013), 56–66. 25, 26

[ACW17] AMBRUŞ R., CLAICI S., WENDT A.: Automatic room segmentation from unstructured 3-D data of indoor environments. *IEEE Robotics and Automation Letters 2*, 2 (2017), 749–756. 7, 8, 10, 14, 16, 23

[AH11] ADAN A., HUBER D.: 3D reconstruction of interior wall surfaces under occlusion and clutter. In *Proc. 3DIMPVT* (2011), pp. 275–281. 5, 23

[AHG*19] ARMENI I., HE Z.-Y., GWAK J., ZAMIR A. R., FISCHER M., MALIK J., SAVARESE S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proc. ICCV* (2019). 6, 23

[ASZ*16] ARMENI I., SENER O., ZAMIR A. R., JIANG H., BRILAKIS I., FISCHER M., SAVARESE S.: 3D semantic parsing of large-scale indoor spaces. In *Proc. CVPR* (2016), pp. 1534–1543. 7, 8

[ASZS17] ARMENI I., SAX A., ZAMIR A. R., SAVARESE S.: Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints* (Feb. 2017). arXiv:1702.01105. 4

[BB10] BUDRONI A., BÖHM J.: Automated 3D reconstruction of interiors from point clouds. *International Journal of Architectural Computing 8*, 1 (2010), 55–73. 8, 10, 16

[BdLGM14] BOULCH A., DE LA GORCE M., MARLET R.: Piecewise-planar 3D reconstruction with edge and corner regularization. *Computer Graphics Forum 33*, 5 (2014), 55–64. 13, 17, 24

[BFFFS14] BAO S. Y., FURLAN A., FEI-FEI L., SAVARESE S.: Understanding the 3D layout of a cluttered room from multiple images. In *Proc. IEEE WACV* (2014), pp. 690–697. 3, 10, 13, 18, 19

[BHMT13] BOULCH A., HOULLIER S., MARLET R., TOURNAIRE O.: Semantizing complex 3D scenes using constrained attribute grammars. *Computer Graphics Forum 32*, 5 (2013), 33–42. 24

[BKHS17] BOBKOV D., KIECHLE M., HILSENBECK S., STEINBACH E.: Room segmentation in 3D point clouds using anisotropic potential fields. In *Proc. ICME* (2017), pp. 727–732. 7, 9

[BSFC08] BROSTOW G. J., SHOTTON J., FAUQUEUR J., CIPOLLA R.: Segmentation and recognition using structure from motion point clouds. In *Proc, ECCV* (2008), Forsyth D., Torr P., Zisserman A., (Eds.), pp. 44–57. 17

[BSRVG14] BÓDIS-SZOMORÚ A., RIEMENSCHNEIDER H., VAN GOOL L.: Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *Proc. CVPR* (2014), pp. 469–476. 13

[BTS*17] BERGER M., TAGLIASACCHI A., SEVERSKY L. M., ALLIEZ P., GUENNEBAUD G., LEVINE J. A., SHARF A., SILVA C. T.: A survey of surface reconstruction from point clouds. *Comput. Graph. Forum 36*, 1 (2017), 301–329. 1, 2, 5, 6, 7

[BVZ01] BOYKOV Y., VEKSLER O., ZABIH R.: Fast approximate energy minimization via graph cuts. *IEEE TPAMI 23*, 11 (November 2001), 1222–1239. 16, 25

[CB16] CINAROGLU I., BASTANLAR Y.: A direct approach for object detection with catadioptric omnidirectional cameras. *Signal, Image and Video Processing 10*, 2 (2016), 413–420. 21

[CDF*17] 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. In *Proc. 3DV* (2017), pp. 667–676. 4

[CF14] CABRAL R., FURUKAWA Y.: Piecewise planar and compact floorplan reconstruction from images. In *Proc. CVPR* (2014), pp. 628–635. 10, 13, 14, 15, 25, 27

[CLH15] CHEN K., LAI Y.-K., HU S.-M.: 3D indoor scene modeling from RGB-D data: a survey. *Computational Visual Media 1*, 4 (2015), 267–278. 2, 3, 4, 28

[CLP10] CHAUVE A.-L., LABATUT P., PONS J.-P.: Robust piecewise-planar 3D reconstruction and completion from large-scale unstructured point data. In *Proc. CVPR* (2010), pp. 1261–1268. 13, 15

[CLWF19] CHEN J., LIU C., WU J., FURUKAWA Y.: Floor-SP: Inverse CAD for floorplans by sequential room-wise shortest path. *Proc. ICCV* (2019). 4, 7, 9, 10, 15, 22, 23, 24, 28

[COCSD03] COHEN-OR D., CHRYSANTHOU Y. L., SILVA C. T., DURAND F.: A survey of visibility for walkthrough applications. *IEEE TVCG 9*, 3 (2003), 412–431. 26

[CON08] CHRISTIE M., OLIVIER P., NORMAND J.-M.: Camera control in computer graphics. *Computer Graphics Forum 27*, 8 (2008), 2197–2218. 26

[Cor12] CORNELL UNIVERSITY: Cornell RGBD dataset. http://pr.cs.cornell.edu/sceneunderstanding/data/data.php, 2012. [Accessed: 2019-09-25]. 4

[CRS18] CRS4 VISUAL COMPUTING: CRS4 ViC Research Datasets. http://vic.crs4.it/download/datasets/, 2018. [Accessed: 2019-09-25]. 4, 5

[CXY*15] CHEN K., XU K., YU Y., WANG T.-Y., HU S.-M.: Magic Decorator: Automatic material suggestion for indoor digital scenes. *ACM TOG 34*, 6 (Oct. 2015), 232:1–232:11. 25, 26

[CY99] COUGHLAN J. M., YUILLE A. L.: Manhattan world: Compass direction from a single image by bayesian inference. In *Proc. ICCV* (1999), vol. 2, pp. 941–947. 3, 6

[CZK15] CHOI S., ZHOU Q.-Y., KOLTUN V.: Robust reconstruction of indoor scenes. In *Proc. CVPR* (June 2015), pp. 5828–5839. 4

[DBF*12] DEL PERO L., BOWDISH J., FRIED D., KERMGARD B., HARTLEY E., BARNARD K.: Bayesian geometric modeling of indoor scenes. In *Proc. CVPR* (2012), pp. 2719–2726. 19

[DBGBR*14] DI BENEDETTO M., GANOVELLI F., BALSA RODRIGUEZ M., JASPE VILLANUEVA A., SCOPIGNO R., GOBBETTI E.: ExploreMaps: Efficient construction and ubiquitous exploration of panoramic view graphs of complex 3D environments. *Comput. Graph. Forum 33*, 2 (2014), 459–468. 7, 27

[DBK*13] DEL PERO L., BOWDISH J., KERMGARD B., HARTLEY E., BARNARD K.: Understanding bayesian rooms using composite 3D object models. In *Proc. CVPR* (2013), pp. 153–160. 18, 19

[DCS*17a] DAI A., CHANG A. X., SAVVA M., HALBER M., FUNKHOUSER T., NIESSNER M.: ScanNet Data. http://www.scan-net.org/, 2017. [Accessed: 2019-09-25]. 4, 5

[DCS*17b] DAI A., CHANG A. X., SAVVA M., HALBER M., FUNKHOUSER T., NIESSNER M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. CVPR* (2017). 4

[DHN06] DELAGE E., HONGLAK LEE, NG A. Y.: A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image. In *Proc. CVPR* (2006), vol. 2, pp. 2418–2428. 6, 10

[DNZ*17] DAI A., NIESSNER M., ZOLLHOFER M., IZADI S., THEOBALT C.: BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM TOG 36*, 4 (2017), 24:1–24:18. 4

[DSB*12] DARABI S., SHECHTMAN E., BARNES C., GOLDMAN D. B., SEN P.: Image melding: Combining inconsistent images using patch-based synthesis. *ACM TOG 31*, 4 (2012), 82–1. 26

[DZZ*19] DING Y., ZHENG X., ZHOU Y., XIONG H., ET AL.: Low-cost and efficient indoor 3D reconstruction through annotated hierarchical structure-from-motion. *Remote Sensing 11*, 1 (2019), 58. 28

[EOS86] EDELSBRUNNER H., O'ROURKE J., SEIDEL R.: Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing 15*, 2 (May 1986), 341–363. 14

[ETH17] ETH ZURICH: ETH3D Dataset. https://www.eth3d.net/datasets, 2017. [Accessed: 2019-09-25]. 5

[FCSS09a] FURUKAWA Y., CURLESS B., SEITZ S. M., SZELISKI R.: Manhattan-world stereo. In *Proc. CVPR* (2009), pp. 1422–1429. 3, 13, 15, 17, 26, 27

[FCSS09b] FURUKAWA Y., CURLESS B., SEITZ S. M., SZELISKI R.: Reconstructing building interiors from images. In *Proc. ICCV* (2009), pp. 80–87. 3, 5, 6, 7, 10, 13, 17, 18

[FCW*17] FU Q., CHEN X., WANG X., WEN S., ZHOU B., FU H.: Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Trans. Graph. 36*, 6 (Nov. 2017), 201:1–201:13. 4, 26, 27

[FDL15] FATHI H., DAI F., LOURAKIS M.: Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges. *Advanced Engineering Informatics 29*, 2 (2015), 149–161. 2, 5

[Fir16] FIRMAN M.: RGBD Datasets: Past, Present and Future. In *Proc. CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis* (2016). 4

[FMAJB16] FIRMAN M., MAC AODHA O., JULIER S., BROSTOW G. J.: Structured prediction of unobserved voxels from a single depth image. In *Proc. CVPR* (2016), pp. 5431–5440. 21

[FMMR10] FLINT A., MEI C., MURRAY D., REID I.: A dynamic programming approach to reconstructing building interiors. In *Proc ECCV* (2010), Daniilidis K., Maragos P., Paragios N., (Eds.), pp. 394–407. 10, 11, 12, 13

[FMR11] FLINT A., MURRAY D., REID I.: Manhattan scene understanding using monocular, stereo, and 3D features. In *Proc. ICCV* (2011), pp. 2228–2235. 10, 11, 13

[FPF07] FRIEDMAN S., PASULA H., FOX D.: Voronoi random fields: Extracting topological structure of indoor environments via place labeling. In *IJCAI* (2007), vol. 7, pp. 2109–2114. 9

[FSL*15] FISHER M., SAVVA M., LI Y., HANRAHAN P., NIESSNER M.: Activity-centric scene synthesis for functional 3D scene modeling. *ACM TOG 34*, 6 (2015), 170:1–179:13. 26, 27

[GD00] GEYER C., DANIILIDIS K.: A unifying theory for central panoramic systems and practical implications. In *Proc. ECCV* (2000), pp. 445–461. 12

[GFP10] GALLUP D., FRAHM J.-M., POLLEFEYS M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proc. CVPR* (2010), pp. 1418–1425. 26, 27

[GH13] GUO R., HOIEM D.: Support surface prediction in indoor scenes. In *Proc. ICCV* (2013), pp. 2144–2151. 3

[GKY08] GOBBETTI E., KASIK D., YOON S.-E.: Technical strategies for massive model visualization. In *Proc. ACM Symp. on Solid and physical modeling* (2008), pp. 405–415. 26

[GSEH11] GUPTA A., SATKIN S., EFROS A. A., HEBERT M.: From 3D scene geometry to human workspace. In *Proc. CVPR* (2011), pp. 1961–1968. 17

[HDGN17] HUANG J., DAI A., GUIBAS L., NIESSNER M.: 3Dlite: Towards commodity 3D scanning for content creation. *ACM TOG 36*, 6 (2017), 203:1–203:14. 4, 25

[HEH07] HOIEM D., EFROS A. A., HEBERT M.: Recovering surface layout from an image. *International Journal of Computer Vision 75*, 1 (Oct 2007), 151–172. 11

[HGDG17] HE K., GKIOXARI G., DOLLÁR P., GIRSHICK R.: Mask R-CNN. In *Proc. ICCV* (2017), pp. 2961–2969. 9, 17

[HHF09] HEDAU V., HOIEM D., FORSYTH D.: Recovering the spatial layout of cluttered rooms. In *Proc. ICCV* (2009), pp. 1849–1856. 6, 10, 11, 12, 18, 19

[HHF10] HEDAU V., HOIEM D., FORSYTH D.: Thinking inside the box: Using appearance models and context based on room geometry. In *Proc. ECCV* (2010), pp. 224–237. 18, 19

[HK01] HAREL D., KOREN Y.: On clustering using random walks. In *Proc. FST TCS* (2001), pp. 18–41. 7

[IDB*10] IRAQUI A., DUPUIS Y., BOUTTEAU R., ERTAUD J. Y., SAVATIER X.: Fusion of omnidirectional and ptz cameras for face detection and tracking. In *Proc. Int. Conf. on Emerging Security Technologies* (2010), pp. 18–23. 21

[IKH*11] IZADI S., KIM D., HILLIGES O., MOLYNEAUX D., NEWCOMBE R., KOHLI P., SHOTTON J., HODGES S., FREEMAN D., DAVISON A., FITZGIBBON A.: KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proc. UIST* (2011), pp. 559–568. 3

[IYF15] IKEHATA S., YANG H., FURUKAWA Y.: Structured indoor modeling. In *Proc. ICCV* (2015), pp. 1323–1331. 1, 2, 6, 7, 9, 10, 15, 17, 19, 23, 24, 26

[JDN19] JI H., DAI A., NIESSNER M.: 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In *Proc. CVPR* (2019). 18, 20, 28

[JGSC13] JIA Z., GALLAGHER A., SAXENA A., CHEN T.: 3D-based reasoning with blocks, support, and stability. In *Proc. CVPR* (2013), pp. 1–8. 3, 26, 27

[JHS09] JENKE P., HUHLE B., STRASSER W.: Statistical reconstruction of indoor scenes. In *Proc. WSCG* (2009), pp. 17–24. 10, 13, 14

[KLTZ16] KERMANI Z. S., LIAO Z., TAN P., ZHANG H.: Learning 3D scene synthesis from annotated RGB-D images. In *Proc. SGP* (2016), pp. 197–206. 26, 27

[KMYG12] KIM Y. M., MITRA N. J., YAN D.-M., GUIBAS L.: Acquiring 3D indoor environments with variability and repetition. *ACM TOG 31*, 6 (2012), 138:1–138:10. 17, 18, 21

[KRNH11] KANG S., ROH A., NAM B., HONG H.: People detection method using graphics processing units for a mobile robot with an omnidirectional camera. *Optical Engineering 50* (2011), 50:1–50:9. 21

[KSF*12] KUSHAL A., SELF B., FURUKAWA Y., GALLUP D., HERNANDEZ C., CURLESS B., SEITZ S. M.: Photo tours. In *Proc. 3DIMPVT* (2012), pp. 57–64. 2

[KYZB19] KAISER A., YBANEZ ZEPEDA J. A., BOUBEKEUR T.: A survey of simple geometric primitives detection methods for captured 3D data. *Computer Graphics Forum 38*, 1 (2019), 167–196. 2, 14

[LGHK10] LEE D. C., GUPTA A., HEBERT M., KANADE T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *Proc. NIPS* (2010), pp. 1288–1296. 17, 18, 19

[LHK09] LEE D. C., HEBERT M., KANADE T.: Geometric reasoning for single image structure recovery. In *Proc. CVPR* (2009), pp. 2136–2143. 6, 10, 11, 12, 18, 19

[LKF16] LIU C., KOHLI P., FURUKAWA Y.: Layered scene decomposition via the Occlusion-CRF. In *Proc. CVPR* (2016), pp. 165–173. 22

[LKG*19] LIU C., KIM K., GU J., FURUKAWA Y., KAUTZ J.: Planercnn: 3D plane detection and reconstruction from a single image. In *Proc. CVPR* (2019), pp. 4450–4459. 17

[LKN*17] LEHTOLA V., KAARTINEN H., NÜCHTER A., KAIJALUOTO R., KUKKO A., LITKEY P., HONKAVAARA E., ROSNELL T., VAAJA M., VIRTANEN J.-P., ET AL.: Comparison of the selected state-of-the-art 3D indoor scanning and point cloud generation methods. *Remote sensing 9*, 8 (2017), 796. 2, 3, 5, 28

[LRY*16] LEE K., RYU S., YEON S., CHO H., JUN C., KANG J., CHOI H., HYEON J., BAEK I., JUNG W., KIM H., DOH N. L.: Accurate continuous sweeping framework in indoor spaces with backpack sensor system for applications to 3-D mapping. *IEEE Robotics and Automation Letters 1*, 1 (2016), 316–323. 25

[LWF18a] LIU C., WU J., FURUKAWA Y.: FloorNet: A unified framework for floorplan reconstruction from 3D scans. In *Proc. ECCV* (2018), Ferrari

V., Hebert M., Sminchisescu C., Weiss Y., (Eds.), pp. 203–219. 4, 10, 15, 17, 22, 28

[LWF18b] LIU C., WU J., FURUKAWA Y.: Floornet data. https://github.com/art-programmer/FloorNet, 2018. [Accessed: 2018-10-24]. 4, 5

[LWKF17] LIU C., WU J., KOHLI P., FURUKAWA Y.: Raster-to-vector: Revisiting floorplan transformation. In *Proc. ICCV* (2017), pp. 2195–2203. 10, 15

[LYC*18] LIU C., YANG J., CEYLAN D., YUMER E., FURUKAWA Y.: Planenet: Piece-wise planar reconstruction from a single RGB image. In *Proc. CVPR* (2018), pp. 2579–2588. 17

[Mas12] MASSACHUSSETS INSTITUTE OF TECHNOLOGY: SUN360 Database. http://people.csail.mit.edu/jxiao/SUN360/, 2012. [Accessed: 2019-09-25]. 4, 5, 12

[Mat17] MATTERPORT: Matterport3D. https://github.com/niessner/Matterport, 2017. [Accessed: 2019-09-25]. 2, 4, 5, 27

[MJM*14] MURA C., JASPE VILLANUEVA A., MATTAUSCH O., GOBBETTI E., PAJAROLA R.: Reconstructing complex indoor environments with arbitrary walls orientations. In *Eurographics Posters* (2014). 15, 18, 22, 23

[MMBM15] MONSZPART A., MELLADO N., BROSTOW G. J., MITRA N. J.: RAPter: Rebuilding man-made scenes with regular arrangements of planes. *ACM TOG 34*, 4 (2015), 103:1–103:12. 13, 28

[MMJV*14] MURA C., MATTAUSCH O., JASPE VILLANUEVA A., GOBBETTI E., PAJAROLA R.: Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers & Graphics 44* (2014), 20–32. 1, 3, 4, 7, 8, 10, 13, 14, 16

[MMP16] MURA C., MATTAUSCH O., PAJAROLA R.: Piecewise-planar reconstruction of multi-room interiors with arbitrary wall arrangements. *Computer Graphics Forum 35*, 7 (2016), 179–188. 3, 4, 7, 8, 10, 13, 15, 16, 18, 22, 23, 24, 28

[MP17] MURA C., PAJAROLA R.: Exploiting the room structure of buildings for scalable architectural modeling of interiors. In *ACM SIGGRAPH Posters* (2017), pp. 4:1–4:2. 7, 8, 10, 22, 24

[MPM*14] MATTAUSCH O., PANOZZO D., MURA C., SORKINE-HORNUNG O., PAJAROLA R.: Object detection and classification from large-scale cluttered indoor scans. *Computer Graphics Forum 33*, 2 (2014), 11–21. 4, 17, 21

[MSL*11] MERRELL P., SCHKUFZA E., LI Z., AGRAWALA M., KOLTUN V.: Interactive furniture layout using interior design guidelines. *ACM TOG 30*, 4 (July 2011), 87:1–87:10. 26, 27

[MSOP17] MURALI S., SPECIALE P., OSWALD M. R., POLLEFEYS M.: Indoor Scan2BIM: Building information models of house interiors. In *Proc. IROS* (2017), pp. 6126–6133. 7, 8, 10, 13, 14, 18, 23

[MVG13] MARTINOVIC A., VAN GOOL L.: Bayesian grammar learning for inverse procedural modeling. In *Proc. CVPR* (2013), pp. 201–208. 26

[MWA*13] MUSIALSKI P., WONKA P., ALIAGA D. G., WIMMER M., VAN GOOL L., PURGATHOFER W.: A survey of urban reconstruction. *Computer graphics forum 32*, 6 (2013), 146–177. 2, 3, 26

[Nav12] NAVVIS: TUMViewer. https://www.navvis.lmt.ei.tum.de/view/, 2012. [Accessed: 2019-09-25]. 27

[New12] NEW YORK UNIVERSITY: NYU-Depth V2. https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html, 2012. [Accessed: 2019-09-25]. 4

[NXS12] NAN L., XIE K., SHARF A.: A search-classify approach for cluttered indoor scene understanding. *ACM TOG 31*, 6 (2012), 137:1–137:10. 18, 20, 26

[OLA14] OESAU S., LAFARGE F., ALLIEZ P.: Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut. *ISPRS Journal of Photogrammetry and Remote Sensing 90* (2014), 68–82. 8, 10, 13, 14, 15, 16, 18, 23

[OLA16] OESAU S., LAFARGE F., ALLIEZ P.: Planar shape detection and regularization in tandem. *Computer Graphics Forum 35*, 1 (2016), 203–215. 13

[OVK19] OCHMANN S., VOCK R., , KLEIN R.: Automatic reconstruction of fully volumetric 3D building models from oriented point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing 151* (2019), 251 – 262. 6, 7, 8, 10, 13, 15, 16, 18, 22, 23, 24

[OVW*14] OCHMANN S., VOCK R., WESSEL R., TAMKE M., KLEIN R.: Automatic generation of structural building descriptions from 3D point cloud scans. In *Proc. GRAPP* (2014), pp. 1–8. 7

[OVWK16] OCHMANN S., VOCK R., WESSEL R., KLEIN R.: Automatic reconstruction of parametric building models from indoor point clouds. *Computers & Graphics 54* (February 2016), 94–103. 7, 8, 10, 13, 14, 16, 18, 22, 23

[PAD*98] PULLI K., ABI-RACHED H., DUCHAMP T., SHAPIRO L. G., STUETZLE W.: Acquisition and visualization of colored 3D objects. In *Proc. Pattern Recognition* (1998), vol. 1, pp. 11–15. 3

[PAG14] PINTORE G., AGUS M., GOBBETTI E.: Interactive mapping of indoor building structures through mobile devices. In *Proc. 3DV* (2014), vol. 2, pp. 103–110. 2, 28

[PAN*15] PĂTRĂUCEAN V., ARMENI I., NAHANGI M., YEUNG J., BRILAKIS I., HAAS C.: State of research in automatic as-built modelling. *Advanced Engineering Informatics 29*, 2 (2015), 162–171. 2

[PCJC10] PRONOBIS A., CAPUTO B., JENSFELT P., CHRISTENSEN H. I.: A realistic benchmark for visual indoor place recognition. *Robotics and autonomous systems 58*, 1 (2010), 81–96. 9

[PGCD17] PINTUS R., GOBBETTI E., CALLIERI M., DELLEPIANE M.: *Techniques for seamless color registration and mapping on dense 3D models*. Springer, 2017, pp. 355–376. 3

[PGG*16] PINTORE G., GARRO V., GANOVELLI F., AGUS M., GOBBETTI E.: Omnidirectional image capture on mobile devices for fast automatic generation of 2.5D indoor maps. In *Proc. IEEE WACV* (2016), pp. 1–9. 12

[PGGS16a] PINTORE G., GANOVELLI F., GOBBETTI E., SCOPIGNO R.: Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness. In *Proc. ECCV Workshops* (2016), pp. 130–145. 3, 27

[PGGS16b] PINTORE G., GANOVELLI F., GOBBETTI E., SCOPIGNO R.: Mobile reconstruction and exploration of indoor structures exploiting omnidirectional images. In *Proc. SIGGRAPH Asia Symposium on Mobile Graphics and Interactive Applications* (2016). 25, 27

[PGJG19] PINTORE G., GANOVELLI F., JASPE VILLANUEVA A., GOBBETTI E.: Automatic modeling of cluttered floorplans from panoramic images. *Computer Graphics Forum 38*, 7 (2019). To appear. 3, 4, 6, 7, 9, 18, 20, 24

[PGP*18] PINTORE G., GANOVELLI F., PINTUS R., SCOPIGNO R., GOBBETTI E.: 3D floor plan recovery from overlapping spherical images. *Computational Visual Media 4*, 4 (2018), 367–383. 4, 6, 7, 10, 13

[PPG*18] PINTORE G., PINTUS R., GANOVELLI F., SCOPIGNO R., GOBBETTI E.: Recovering 3D existing-conditions of indoor structures from spherical images. *Computers & Graphics 77* (2018), 16–29. 4, 10, 12, 18, 20, 21, 22, 24

[Pri13] PRINCETON UNIVERSITY: SUN3D Database. https://sun3d.cs.princeton.edu/, 2013. [Accessed: 2019-09-25]. 4, 5

[Pri15] PRINCETON UNIVERSITY: SUNRGBD Database. http://3dvision.princeton.edu/projects/2015/SUNrgbd/, 2015. [Accessed: 2019-09-25]. 4

[Pri16] PRINCETON UNIVERSITY: SceneCG Dataset. https://sscnet.cs.princeton.edu/, 2016. [Accessed: 2019-09-25]. 4, 5

[Rec16] RECONSTRUCT INC: Reconstruct: A Visual Command Center. https://www.reconstructinc.com/, 2016. 2

[RF17] REDMON J., FARHADI A.: YOLO9000: Better, faster, stronger. In *Proc. CVPR* (2017), pp. 7263–7271. 20

[SCC12] SHIN H., CHON Y., CHA H.: Unsupervised construction of an indoor floor plan using a smartphone. *IEEE TPAMI 42*, 6 (2012), 889–898. 3

[SCD*06] SEITZ S. M., CURLESS B., DIEBEL J., SCHARSTEIN D., SZELISKI R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR* (2006), vol. 1, pp. 519–528. 3

[SD04] SCHINDLER G., DELLAERT F.: Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proc. CVPR* (2004), vol. 1, pp. I–I. 6, 12

[SFCH12] SHEN C.-H., FU H., CHEN K., HU S.-M.: Structure recovery by part assembly. *ACM TOG 31*, 6 (2012), 180:1–180:10. 18, 21, 28

[SFPU13] SCHWING A. G., FIDLER S., POLLEFEYS M., URTASUN R.: Box in the box: Joint 3D layout and object reasoning from single images. In *Proc. ICCV* (2013), pp. 353–360. 18, 19

[SGD*18] SCHUBERT D., GOLL T., DEMMEL N., USENKO V., STUECKLER J., CREMERS D.: The TUM VI benchmark for evaluating visual-inertial odometry. In *Proc. IROS* (2018). 4

[SHSC19] SUN C., HSIAO C.-W., SUN M., CHEN H.-T.: HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *Proc. CVPR* (June 2019). 5, 10, 12, 28

[SQLG15] SU H., QI C. R., LI Y., GUIBAS L. J.: Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proc. ICCV* (2015), pp. 2686–2694. 19

[SRLH15] SATKIN S., RASHID M., LIN J., HEBERT M.: 3DNN: 3D nearest neighbor. data-driven geometric scene understanding using 3D models. *International Journal of Computer Vision 111* (2015), 69–97. 18, 19, 26

[SS12] SANKAR A., SEITZ S.: Capturing indoor scenes with smartphones. In *Proc. UIST* (2012), pp. 403–412. 2, 27, 28

[SSG*17] SCHÖPS T., SCHÖNBERGER J. L., GALLIANI S., SATTLER T., SCHINDLER K., POLLEFEYS M., GEIGER A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR* (July 2017), pp. 2538–2547. 4

[SSS*08a] SINHA S. N., STEEDLY D., SZELISKI R., AGRAWALA M., POLLEFEYS M.: Interactive 3D architectural modeling from unordered photo collections. *ACM TOG 27*, 5 (Dec. 2008), 159:1–159:10. 25

[SSS08b] SNAVELY N., SEITZ S. M., SZELISKI R.: Modeling the world from internet photo collections. *International Journal of Computer Vision 80*, 2 (2008), 189–210. 17

[SSS09] SINHA S. N., STEEDLY D., SZELISKI R.: Piecewise planar stereo for image-based rendering. In *Proc. ICCV* (2009), pp. 1881–1888. 26, 27

[Sta16a] STANFORD UNIVERSITY: Bundle Fusion Dataset. https://graphics.stanford.edu/projects/bundlefusion, 2016. [Accessed: 2019-09-25]. 4, 5

[Sta16b] STANFORD UNIVERSITY: PiGraphs Dataset. https://graphics.stanford.edu/projects/pigraphs/, 2016. [Accessed: 2019-09-25]. 4

[Sta17] STANFORD UNIVERSITY: BuildingParser Dataset. http://buildingparser.stanford.edu/dataset.html, 2017. [Accessed: 2019-09-25]. 4, 5

[Str16] STRUCTIONSITE: VideoWalk. https://www.structionsite.com/products/videowalk/, 2016. 2

[SWK07] SCHNABEL R., WAHL R., KLEIN R.: Efficient RANSAC for point-cloud shape detection. *Computer Graphics Forum 26*, 2 (2007), 214–226. 13

[SWM*19] STRAUB J., WHELAN T., MA L., CHEN Y., WIJMANS E., GREEN S., ENGEL J. J., MUR-ARTAL R., REN C., VERMA S., CLARKSON A., YAN M., BUDGE B., YAN Y., PAN X., YON J., ZOU Y., LEON K., CARTER N., BRIALES J., GILLINGHAM T., MUEGGLER E., PESQUEIRA L., SAVVA M., BATRA D., STRASDAT H. M., NARDI R. D., GOESELE M., LOVEGROVE S., NEWCOMBE R.: The replica dataset: A digital replica of indoor spaces, 2019. arXiv:1906.05797. 4, 5

[SXZ*12] 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 TOG 31*, 6 (2012), 136:1–136:10. 18, 20, 21, 26

[SYMA12] STROILA M., YALCIN A., MAYS J., ALWAR N.: Route visualization in indoor panoramic imagery with open area maps. In *Proc. ICME Workshops* (2012), pp. 499–504. 27

[SYZ*17] SONG S., YU F., ZENG A., CHANG A. X., SAVVA M., FUNKHOUSER T.: Semantic scene completion from a single depth image. *Proc. CVPR* (2017). 4

[SZ12] SANCHEZ V., ZAKHOR A.: Planar 3D modeling of building interiors from point cloud data. In *Proc. ICIP* (2012), pp. 1777–1780. 24

[TCZ15] TURNER E., CHENG P., ZAKHOR A.: Fast, automated, scalable generation of textured 3D models of indoor environments. *IEEE Journal of Selected Topics in Signal Processing 9*, 3 (2015), 409–421. 1, 25, 27

[Tec15] TECHNICAL UNIVERSITY OF MUNICH: TUM LSI Dataset. https://hazirbas.com/datasets/tum-lsi/, 2015. [Accessed: 2019-09-25]. 4

[THA*10] TANG P., HUBER D., AKINCI B., LIPMAN R., LYTLE A.: Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in Construction 19*, 7 (2010), 829–843. 2, 3

[TZ12] TURNER E., ZAKHOR A.: Watertight as-built architectural floor plans generated from laser range data. In *Proc. 3DIMPVT* (2012), pp. 316–323. 3, 8, 10, 14, 18, 23

[TZ13] TURNER E., ZAKHOR A.: Watertight planar surface meshing of indoor point-clouds with voxel carving. In *Proc. 3DV* (2013), pp. 41–48. 3, 8, 10, 14, 17, 23

[TZ14] TURNER E., ZAKHOR A.: Floor plan generation and room labeling of indoor environments from laser range data. In *Proc. Int. Conf. on Computer Graphics Theory and Applications* (2014), pp. 22–33. 6, 7, 8, 10, 14, 18, 23

[Uni14] UNIVERSITY OF ZURICH: UZH 3D dataset. https://www.ifi.uzh.ch/en/vmml/research/datasets.html, 2014. [Accessed: 2019-09-25]. 4, 5

[Uni16] UNIVERSITY OF ZURICH: SceneNN Dataset. https://www.ifi.uzh.ch/en/vmml/research/datasets.html, 2016. [Accessed: 2019-09-25]. 4

[VSS14] VOLK R., STENGEL J., SCHULTMANN F.: Building information modeling (BIM) for existing buildings — literature review and future needs. *Automation in Construction 38* (2014), 109 – 127. 2

[VZ18] VEDALDI A., ZISSERMAN A.: Object instance recognition. http://www.robots.ox.ac.uk/~vgg/practicals/instance-recognition/index.html, 2018. [Accessed: 2018-10-24]. 21

[Was14] WASHINGTON UNIVERSITY: Washington RGBD dataset. https://rgbd-dataset.cs.washington.edu/dataset.html, 2014. [Accessed: 2019-09-25]. 4

[WGK10] WANG H., GOULD S., KOLLER D.: Discriminative learning with latent variables for cluttered indoor scene understanding. In *Proc. ECCV* (2010), Daniilidis K., Maragos P., Paragios N., (Eds.), pp. 435–449. 20

[WL09] WANG M. L., LIN H. Y.: Object recognition from omnidirectional visual sensing for mobile robot applications. In *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics* (2009), pp. 1941–1946. 21

[WLL*18] WU T., LIU J., LI M., CHEN R., HYYPPÄ J.: Automated large scale indoor reconstruction using vehicle survey data. In *Proc. UPINLBS* (2018), pp. 1–5. 3

[XAAH13] XIONG X., ADAN A., AKINCI B., HUBER D.: Automatic creation of semantically rich 3D building models from laser scanner data. *Automation in Construction 31* (2013), 325–337. 2, 22, 23

[XCF*13] XU K., CHEN K., FU H., SUN W.-L., HU S.-M.: Sketch2Scene: Sketch-based co-retrieval and co-placement of 3D models. *ACM TOG 32*, 4 (July 2013), 123:1–123:15. 26, 27

[XEOT12] XIAO J., EHINGER K. A., OLIVA A., TORRALBA A.: Recognizing scene viewpoint using panoramic place representation. In *Proc. CVPR* (June 2012), pp. 2695–2702. 4

[XF14] XIAO J., FURUKAWA Y.: Reconstructing the world's museums. *International Journal of Computer Vision 110*, 3 (Dec 2014), 243–258. 2, 25, 27

[XHS*15] XU K., HUANG H., SHI Y., LI H., LONG P., CAICHEN J., SUN W., CHEN B.: Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM TOG 34*, 6 (2015), 177:1–177:14. 20

[XMS14] XIANG Y., MOTTAGHI R., SAVARESE S.: Beyond PASCAL: A benchmark for 3D object detection in the wild. In *Proc. WACV* (2014), pp. 75–82. 19

[XOT13] XIAO J., OWENS A., TORRALBA A.: SUN3D: A database of big spaces reconstructed using SfM and object labels. In *2013 IEEE International Conference on Computer Vision* (Dec 2013), pp. 1625–1632. 4

[XSKT17] XU J., STENGER B., KEROLA T., TUNG T.: Pano2CAD: Room layout from a single panorama image. In *Proc. WACV* (2017), pp. 354–362. 10, 12, 18, 20

[YJL*18] YANG Y., JIN S., LIU R., , YU J.: Automatic 3D indoor scene modeling from single panorama. In *Proc. CVPR* (2018), pp. 3926–3934. 3, 12, 18, 20

[YLL*19] YAO Y., LUO Z., LI S., SHEN T., FANG T., QUAN L.: Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *Proc. CVPR* (June 2019). 4

[YRM*18] YANG B., ROSA S., MARKHAM A., TRIGONI N., WEN H.: Dense 3D object reconstruction from a single depth view. *IEEE TPAMI* (2018). 21

[YWP*19] YANG S.-T., WANG F.-E., PENG C.-H., WONKA P., SUN M., CHU H.-K.: DuLa-Net: A dual-projection network for estimating room layouts from a single RGB panorama. In *Proc. CVPR* (2019). 10, 12, 28

[YXL*19] YANG J., XU J., LI K., LAI Y., YUE H., LU J., WU H., LIU Y.: Learning to reconstruct and understand indoor scenes from sparse views. *CoRR* (2019). URL: http://arxiv.org/abs/1906.07892. 28

[YZ16] YANG H., ZHANG H.: Efficient 3D room shape recovery from a single panorama. In *Proc. CVPR* (2016), pp. 5422–5430. 3, 4, 10, 12, 13

[YZ18] YANG F., ZHOU Z.: Recovering 3D planes from a single image via convolutional neural networks. In *Proc. ECCV* (2018), pp. 85–100. 17

[YZS*19] 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 19*, 17 (Sep 2019), 3798. 10, 14, 28

[ZCC16] ZHANG E., COHEN M. F., CURLESS B.: Emptying, refurnishing, and relighting indoor spaces. *ACM TOG 35*, 6 (2016), 174:1–174:14. 25

[ZCSH18] ZOU C., COLBURN A., SHAN Q., HOIEM D.: LayoutNet: Reconstructing the 3D room layout from a single RGB image. In *Proc. CVPR* (2018), pp. 2051–2059. 3, 12

[ZGM18] ZHU J., GUO Y., MA H.: A data-driven approach for furniture and indoor scene colorization. *IEEE TVCG 24*, 9 (2018), 2473–2486. 25, 26

[ZSG*18] ZOLLHÖFER M., STOTKO P., GÖRLITZ A., THEOBALT C., NIESSNER M., KLEIN R., KOLB A.: State of the art on 3d reconstruction with rgb-d cameras. *Comput. Graph. Forum 37*, 2 (2018), 625–652. 2

[ZSL*15] ZHANG J., SCLAROFF S., LIN Z., SHEN X., PRICE B., MECH R.: Minimum barrier salient object detection at 80 FPS. In *Proc. ICCV* (2015), pp. 1404–1412. 20

[ZSP*19] ZOU C., SU J.-W., PENG C.-H., COLBURN A., SHAN Q., WONKA P., CHU H.-K., HOIEM D.: 3d manhattan room layout reconstruction from a single 360 image, 2019. arXiv:1910.04099. 12

[ZSTX14] ZHANG Y., SONG S., TAN P., XIAO J.: PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proc. ECCV* (2014), pp. 668–686. 3, 4, 10, 11, 12, 17, 18, 19, 20, 21

[ZXTZ15] ZHANG Y., XU W., TONG Y., ZHOU K.: Online structure analysis for real-time indoor scene reconstruction. *ACM TOG 34*, 5 (2015), 159:1–159:13. 5, 20

[ZYT17] ZHENG L., YANG Y., TIAN Q.: SIFT meets CNN: A decade survey of instance retrieval. *IEEE TPAMI 40*, 5 (2017), 1224–1244. 7, 21

[ZZL*19] ZHENG J., ZHANG J., LI J., TANG R., GAO S., ZHOU Z.: Structured3D: A large photo-realistic dataset for structured 3D modeling, 2019. arXiv:1908.00222. 5

[ZZZ*19] ZHENG L., ZHU C., ZHANG J., ZHAO H., HUANG H., NIESSNER M., XU K.: Active scene understanding via online semantic reconstruction. *Computer Graphics Forum 38*, 7 (2019), 103–114. 20, 28