Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Article Type (Research)

3D floor plan recovery from overlapping spherical images

Giovanni Pintore [1] (✉), **Fabio Ganovelli** [1], **Ruggero Pintus** [1], **Roberto Scopigno** [1], and **Enrico Gobbetti** [2]

Abstract We present a novel approach to automatically recover, from a small set of partially overlapping spherical images, an indoor structure representation in terms of a 3D floor plan registered with a set of 3D environment maps. We introduce several improvements over previous approaches based on color/spatial reasoning exploiting *Manhattan World* priors. In particular, we introduce a new method for geometric context extraction based on a 3D facets representation, which combines color distribution analysis of individual images with sparse multi-view clues. Moreover, we introduce an efficient method to combine the facets from different points of view in a single consistent model, considering the reliability of the facets contribution. The resulting capture and reconstruction pipeline automatically generates 3D multi-room environments where most of the other previous approaches fail, such as in presence of hidden corners and large clutter, even without involving additional dense 3D data or tools. We demonstrate the effectiveness and performance of our approach on different real-world indoor scenes. Our test data will be released to allow for further studies and comparisons.

Keywords Indoor reconstruction; spherical panoramic cameras; 360 degrees photography; multi-room environments.

1 CRS4, Visual Computing Group, Cagliari, Italy – `http://www.crs4.it/vic/` – E-mail: name.surname@crs4.it

2 CNR-ISTI, Visual Computing Group, Pisa, Italy – `http://vcg.isti.cnr.it` – E-mail: name.surname@crs4.it

Manuscript received: 2018-09-24; accepted: 201x-xx-xx.

1 Introduction

The attention of consumer-oriented industry towards spherical images has dramatically increased in recent years. Google and Facebook recently added support for 360° to their image and video sharing platforms and released reference camera designs for professional content producers [24]. Numerous consumer-level 360° cameras have just recently become available or will be released later this year, making fairly accessible for consumers to acquire and share panoramic images, or even to capture compelling imagery for stereo viewing in a head-mounted display [28]. While such spherical images could already be obtained by stitching conventional photographic shots, for instance with the help of special-purpose sensor fusion applications on mobile cameras and phones [4, 37], the emergence of these new 360° cameras is significantly reducing capturing efforts.

Large and complex environments can now be captured with very few single-shot 360° images, whose overlap can provide registration information. Such sparse, but visually rich, coverage is a very interesting and simple alternative to dense shape capture, as done with scanners or dense multi-view, especially in applications where location awareness and structure reconstruction is more important than fine geometric acquisition, such as guidance or security applications, which require structured models that support walkthroughs and are photorealistic enough to recognize real places by just looking at them [38]. In the indoor scenario, moreover, solutions based on low-cost devices play an even more important role also for privacy reasons, as they allow individual users to easily acquire and share their own environments using consumer-level tools, without forcing them to provide physical access to other persons for the scanning process [35].

Creating models of indoor environments just from

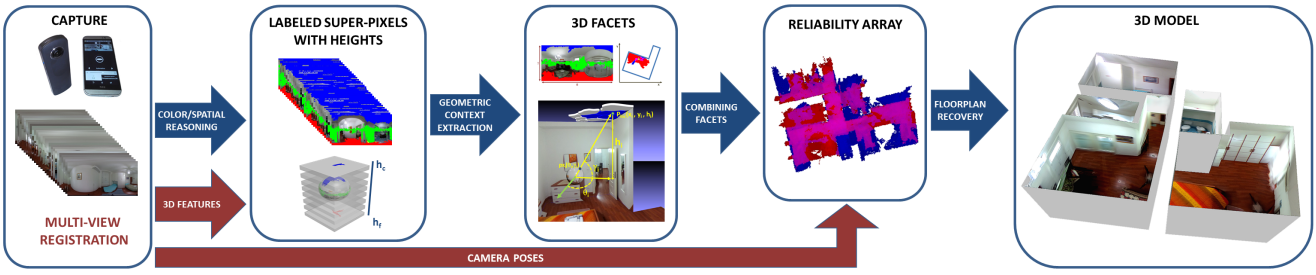


Fig. 1 Overview. For each spherical image, we perform an image-based classification based on *super-pixels*, labeling only those super-pixels that can be unambiguously assigned to floor, walls, and ceilings (Sec. 5.1), and in parallel we recover camera and features alignment through a multi-view registration of the images. We exploit these features to assign to each super-pixel the most likely *height* value. Once the heights are known we use a custom *3D mapping function* able to recover 3D world space points from image-space super-pixels (Sec. 5.2) to generate a 3D world space facets distribution and a 2D accumulation array. We exploit then facets distribution and accumulation array to recover the scene floor-plan and the relative 3D rooms shapes (Sec. 6).

visual data is, however, not an easy task. Major difficulties include poor texture detail, large occlusions, and complex floor-plan topology. Tackling these problems often lead to solutions that entails elaborate acquisition and stitching processes, and/or require complex reasoning to reconstruct invisible parts, often including manual intervention, especially in multi-room environments.

In recent years (see Sec. 2), research has focused on extending conventional image-based approaches for indoor reconstruction by exploiting panoramic imagery. However, these solutions still have many limitations. Solutions based on dense capture typically require long processing times and features to extract a dense point cloud. Faster solutions typically focus on one panoramic image per room, but are capable to infer 3D clues only under very limiting constrains (e.g., *Manhattan World*). Furthermore, all these methods are limited by the strict condition that all the corners of the room must be visible from a single point of view, which make them ineffective in many common indoor environments (e.g., L-shapes, multi-room scenes, corridors).

In order to address these issues, we propose a novel and light-weight approach, which efficiently improves over the analysis of individual images by exploiting multi-view clues (see Sec. 3).

Approach We acquire the scene through a small set of partially overlapping 360° images (Fig. 1) and we perform a multi-view registration on them. We generate, for each panoramic point-of-view, a simplified and compact representation of the viewed 3D space as labeled *3D facets*, obtained by augmenting a local color/spatial labeling of super-pixels with geometric information from multi-view 3D features (Sec. 5). The 3D facets from different point-of-views are, then, merged in order to find a consensus geometric context, from which to extract the overall indoor structure as a layout of rooms (Sec. 6). As a result, we obtain a 3D floor plan scaled to metric dimensions registered with a set of 3D environment maps.

Contribution Our main contribution to the state-of-the-art in indoor reconstruction is the following:

- we introduce a novel geometric context extraction approach based on the combination of color/spatial reasoning with sparse multi-view 3D features, dubbed *3D facets* (Sec. 5). This method improves over previous state-of-the-art approaches that try to infer 3D clues from *Manhattan World* vanishing lines priors [46] or from the image edgemap analysis [37];
- we introduce an efficient method to combine 3D facets from different images and evaluating their reliability (Sec. 6); this approach is more robust to clutter, occlusions and segmentation errors, compared to the single-view methods [37, 46] commonly adopted with panoramic images;
- we introduce a novel and practical image-based pipeline to automatically retrieve a multi-room indoor 3D layout from a small set of panoramic images. The indoor scene is quickly captured with commodity cameras, as well as the reconstruction is performed without the aid of externally calculated dense 3D data [6] or additional mobile tools [37]. While not all the individual components in this pipeline are novel by themselves, their elaboration and not-trivial combination significantly improve reconstruction capabilities.

This article is an invited extended version of our

PG 2018 contribution [36]. We here provide a more thorough exposition, but also significant new material, including the presentation of a refined pipeline and additional qualitative and quantitative results. Finally, we have attempted to further clarify the steps in our algorithms to facilitate their implementation and to make the transfer between abstract concepts and actual code as straightforward as possible.

Advantages Our approach allows for a consistent 3D structure extraction for complex multi-room environments. Only few overlapping images are required, and, thus, the method is much less time-consuming than dense multi-view approaches. Although sparse, the recovered 3D features, once integrated with the super-pixel segmentation and the multi-view reasoning, provide a more reliable spatial information than inferring 3D information only from a single image edge lines [37, 46], which, in contrast, are more prone to errors, manly due to the high distortion and low quality of indoor spherical images, and are limited by heavy constrains. Furthermore, with respect to previous approaches, the effective combination of contributions from different points of view allows the recovery of the rooms structure also in presence of large clutter, hidden corners, narrow corridors and multi-room structures, and, in general, even in presence of non-Manhattan World structures.

The effectiveness and performance of our approach is demonstrated on real-world scenes (see Sec. 7), including many cases of difficult texture-less walls and ceilings, where typically it is not possible to apply methods that require a denser and more regular feature coverage [6], as well as cluttered indoor environments. All data will be made publicly available for further studies.

2 Related Work

3D reconstruction of indoor architectural scenes is a very challenging problem. Compared to building exteriors, interiors are often dominated by clutter, barely lit surfaces, and texture-poor walls. Moreover visibility reasoning is more problematic due to the presence of interconnected rooms. The problem has thus attracted a lot of research in recent years.

Devices such as laser scanners, producing dense 3D point clouds, represent an effective solution for an accurate acquisition, but still require a lot of post-processing to extract structured models from raw data[32, 33, 45]. Moreover, the cost of the devices and the need of qualified personnel limits their use to specific application domains, such as Cultural Heritage or engineering. Modern mobile depth-sensing devices, such as RGB-D cameras, have become a promising alternative for widespread short-range 3D acquisition. However, rooms larger than a few meters, for example a hotel hall, are outside the depth range of most of these sensors and make the acquisition process more time consuming [14, 16, 20]. As for laser-scan data, heavy post-processing is also needed to transform the acquired high-density dataset into a structured model. A prominent example is the work of Ikehata et al. [18], which propose a 3D modeling framework that reconstructs an indoor scene as a structured model exploiting panoramic *RGB-D* images. Data-driven approaches with 3D model databases have also proved to be able to yield CAD-quality reconstructions [23, 34]. However the focus of these methods is so far on a clutter analysis in a small scale, such as a single room.

Purely image-based techniques are gaining popularity in several domains [2, 29] and, in certain situations, the accuracy of dense image-based methods is comparable to laser sensor systems at a fraction of the cost [43]. However, they typically require non-negligible acquisition and processing time, and most of the approaches fail in the presence of poor texture detail, typical of indoor environments. This has led to the emergence of methods that aid reconstruction by imposing domain-specific constraints. For example, several authors (e.g., [11, 12, 44]) exploit the heavily constraining *Manhattan World* [9] assumption to reconstruct the 3D structure of moderately cluttered interiors. Bao et al. [3], similarly to our work apply instead both multi-view geometry and single-view analysis, but focus on estimating a single room layout and the foreground objects rather than multi-room structures. In general, however, methods based on pin-hole image capture require a large number of shots. The recent emergence of consumer spherical cameras promises to improve visual capture of indoor environment, since each image covers the complete environment around the viewer, simplifying geometric reasoning, and very few images are required for a large coverage, simplifying the capture process and the features tracking.

Much of the work on omnidirectional images in the past years has been carried out in combination with specialized setups [17] or robotics solutions [8, 42]. In particular, omnidirectional cameras have been extensively used with special catadioptric systems [5, 30, 31] for SLAM and sparse reconstruction from *large motion* [48]. For dense depth map estimation, Li [26]

presented a fisheye stereo method, where the author reformulated a conventional stereo matching scheme for binocular spherical stereo system using the unified spherical model [13]. Kim and Hilton [22] also proposed a stereo matching method for a fisheye stereo camera, where a continuous depth map is obtained from a partial differential equation optimization, while Hane et al. [17] presented a real-time plane-sweeping algorithm which is suitable for images acquired with fisheye cameras. Going into the specifics of modern spherical panoramic cameras (SPC), Im et al. [19] propose a dense 3D reconstruction framework targeted for small motion of a SPC device. Their solution considers the SPC as two physical *fisheye* lenses on a rig, performing stereo calibration and bundle adjustment thanks to the overlapping field-of-view of the lens. What all these visual methods have in common, is that they rely on a sufficiently informative observed environment. In many practical cases, however, large parts of the camera image can become uninformative for SLAM, for instance in the presence of large many untextured walls or moving objects [7].

In recent years, efforts have focused on approaches for indoor reconstruction from panoramic images regardless of a special hardware (i.e., using the most common format of equirectangular image). Cabral et al. [6] adopted stitched equirectangular images to improve indoor reconstruction provided by a dense multi-view pipeline [12]. As clutter and homogeneous zones in indoor scenes tend to leave large reconstruction holes for image-based methods, their method exploits the labeling of the panoramas to complete the multi-view reconstruction obtained from pin-hole images. However, such approach required a considerable number of images and a dense point cloud, thus requiring considerable efforts in terms of user interaction and processing time.

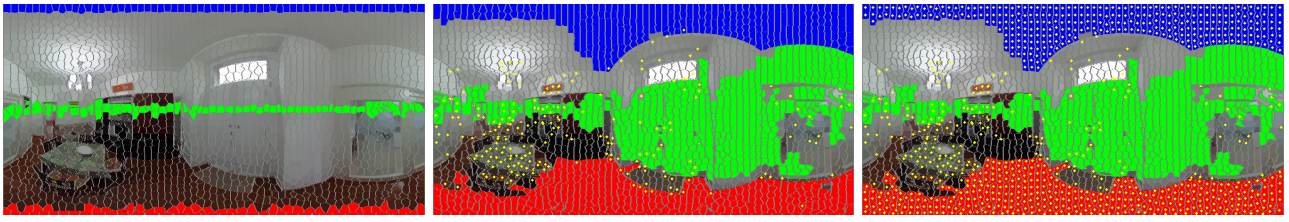
With the goal of minimizing user's burden and simplify geometric reasoning, recent state-of-the-art approaches [37, 46] focus on using only one panoramic image per room. Yang et al. [46] propose an efficient method to recover the 3D shape of a single room based on a constraint graph encoding the spatial configurations of Manhattan World line segments and super-pixels of a single panoramic image. Although effective in many indoor layouts, this approach is limited only to single room environment where all the corners are visible from the same point-of-view. Similarly to Yang et al. [46], Pintore et al. [37] integrate the super-pixel labeling through the analysis of the image's edgemap, extending the result for the single room to multi-room environments with the aid of motion sensors embedded in a mobile device. Although in a less restrictive way than Manhattan World, their approach works only by imposing fixed horizontal floor and ceiling plans, and with environments where all the structural features of the room can be captured with a single shot. This pipeline was recently extended with the purpose of recovering existing conditions [39]. The method uses multiple images, but only for aligning several rooms through the recovery of camera locations. No 3D features are used, and the method is still limited to same Manhattan World constraints.

In this work, we improve over previous solutions by presenting an approach that, starting from a small set of panoramic images, recovers the 3D floor plan of a multi-room environment, by exploiting at the same time multi-view 3D data and single-view image analysis. Such approach is more robust to errors and provides a consistent reconstruction even where previous methods fail.

3 Overview

Our pipeline, summarized in Fig. 1, starts from a set of partially overlapping equirectangular images. We assume that the input images are aligned to the gravity vector. This is easily obtained from mobile devices that have an IMU on board. If this is not the case, the alignment can be obtained by applying a 2D transformation so that vertical edges are aligned with the vertical direction in the image. We consider this alignment an orthogonal problem that can be solved prior to the pipeline, so as to work only with oriented images.

These images are analyzed in parallel (Sec. 5.1) to perform an image-based classification based on *super-pixels*, labeling only those super-pixels that can be unambiguously assigned to floor, walls, and ceiling. We also recover camera and 3D features alignment through a multi-view registration of the images. We exploit these features to assign to each super-pixel the most likely *height* value. Once the heights are known, we use a custom *3D mapping function* able to recover 3D world space points from image-space super-pixels (Sec. 5) to generate a 3D world space facets distribution and a 2D accumulation array. Finally, we exploit the facets distribution and accumulation array to recover the scene floor-plan and the relative 3D rooms shapes (Sec. 6). As a result, we recover a structured visually textured 3D model.



(a) *Initial labeling* (b) *Labels propagation* (c) *Heights propagation*

Fig. 2 Image labeling and features propagation. We show in Fig. 2(a) the initialization of the labeling process, which assigns regions of the image to *ceiling* (blue), *floor* (red), *wall* (green) zones, leaving undecided areas labeled as *unknown*. In Fig. 2(b) we show the labeling after the conservative propagation and we highlight super-pixels of which the height is known from the multi view registration (yellow centroids). In Fig. 2(c) we illustrate the final heights propagation.

4 Multi-view registration

As a first step in our pipeline, we run a SfM registration method [21] to extract each spherical camera orientation $[R]$ and pose $[T]$, as well as the 3D feature points. The 3D features obtained will in general be too sparse to serve a reconstruction (i.e., Fig. 13), in particular for an indoor scene, but their projection on the panoramic image tells us the spatial position of a subset of the pixels of the image. As we will see, we can use this very sparse information in conjunction with the segmentation obtained in Section 5.1 to guide the recovery of the room shape.

5 Geometric context extraction based on 3D facets

In order to infer a reliable geometric context for each point-of-view we define a simplified and compact representation of the indoor space, based on the combination of color/spatial reasoning on the images with multi-view 3D features. To this end we introduce a compact representation based on *3D facets*, generated by an appropriate transformation labeled super-pixels points.

5.1 Single-view conservative super-pixel labeling

As a first step to create 3D facets, we aim at conservatively finding small uniform regions of each image that can be assigned with high probability to the room boundaries. Compared to the segmentation and labeling approaches performed in single-view approaches [37, 46], which try to assign a geometric context to all the super-pixels, we only target to detect the most reliable attributions, thus avoiding the creation of wrong 3D facets in the following geometric context extraction step (Sec. 5.2), since our final goal is to integrate many partial, but reliable, image contributions.

Each image is segmented into super-pixels using a distance function D that combines color similarity and spatial proximity [1]. The 5D *Euclidean* distance is given by the distance function:

$$D = \sqrt{d_c^2 + \left(\frac{d_s}{N_s}\right)^2 m^2} \quad (1)$$

we define d_c and d_s respectively the Euclidean distance in *CIELAB* color space and image space, N_s the targeted spacing between super-pixels centers and m a constant value to weigh the relative importance between color similarity and spatial proximity. Choosing a large value for m (i.e. $m = 10$ in our experiments), produces an over-segmentation with respect to the real color distribution, with the goal of creating a fairly uniform spatial clustering and of preserving geometric coherence between centers.

We then perform a loose geometric context labeling, which assigns each super-pixel of the image to *ceiling*, *floor*, *wall* zones, leaving undecided areas labeled as *unknown*.

Since our images are known to be aligned to the gravity vector, we start by labeling as *ceiling* the most top row of super-pixels, *floor* the bottom ones and *wall* the ones lying on the image horizon - i.e., middle of the equirectangular image (Fig. 2(a)). Then, we iteratively propagate the labeling of each super-pixel to its neighbors. During labeling, we maintain in a global queue the distances from each *unknown* super-pixel to the closest of its labeled neighbors, in order to perform labeling in the order of increasing distance. Labeling (Fig. 2(b)) is performed by iteratively extracting the unlabeled pixel with the smallest distance to a neighbor, update its height, and update the queue after each assignment, recomputing the distances of all neighboring super-pixels. The process is made conservative by defining a threshold D_{max} for the distance functions (we experimented

with values ranging between [0.85-1.20]), stopping the labeling when the next propagation candidates have a distance larger than D_{max}.

5.2 Exploiting 3D features to create 3D facets

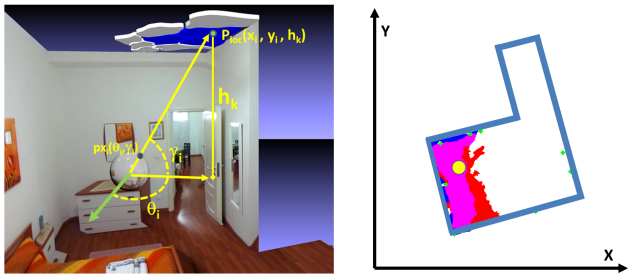


Fig. 3 3D facets from super-pixels. Left: from a pixel $px_i(\theta_i, \gamma_i)$ (θ and γ angles with respect to image's center direction - i.e. green arrow) we obtain a point $P_{loc}(x_i, y_i, h_k)$ in world space through the transform of eq. 2. As a result of this transform, super-pixels points generate horizontal facets in world space (i.e., blue ceiling facets). Right: labeled super-pixels of Fig. 2 transformed in *facets* and projected on the XY plane. Magenta zones are overlapping between ceiling and floor projections, yellow point is the camera position and the azure contour is the underlying shape of the room.

Given a superpixel SP_k labelled as *floor* or *ceiling*, we define as *facet* F_k the planar set of 3D points obtained by the super-pixel projection through the following transformation:

$$P_{loc}(\theta, \gamma, h_k) = \begin{cases} x_l = h_k / \tan\gamma * \cos\theta \\ y_l = h_k / \tan\gamma * \sin\theta \\ z_l = h_k \end{cases} \quad (2)$$

where $P_{loc} \in F_k$ is the 3D position of the pixel $(\theta, \gamma) \in SP_k$. The projection model is illustrated in Fig. 3. The origin of such *local* Cartesian coordinates is the position of the spherical camera, while the abscissa and ordinate of the *equirectangular image* respectively represent the azimuth θ and the tilt γ of the view's direction, that is a pixel (θ, γ) in the equirectangular image. Note that we do not assume that the whole vertical field is captured, but that we know how to map pixel coordinate to angles. Thus, we can cope with cameras that do not completely cover the vertical field (leeving uncaptured a small top and bottom area), assuming that the captured field is known.

In other words, a *floor* or *ceiling* facet F_k is a horizontal patch corresponding to a specific super-pixel SP_k, parametrized on its height h_k. Such representation has several advantages in order to identify the underlying structure. The footprint of *floor* and *ceiling* facets, in-fact, highlights the shape of the room (i.e., Fig. 3, right).

This model assumes that labeled super-pixels to be transformed must have an associated height. Initially only those on which 3D features fall have it (Fig. 2(b)). In order to propagate heights to all the labeled super-pixels we adopt a push-pull [27] height propagation algorithm, assuming that there is at least one height coming from SfM in a connected labeled region (Fig. 2(b)). This ensures that, through the described propagation process, height values will be assigned to all super-pixels in the floor and ceiling regions (Fig. 2(c)), which are two single connected regions by construction.

The facets recovered from a single image are not generally sufficient to define the shape of a room. In the next section we introduce an approach to efficiently combine these contributions to obtain a 3D floor plan.

6 Building 3D models by combining 3D facets from different images

Since we have the pose estimation for each camera, we can bring all the estimated facet points in a common reference frame (i.e., Fig. 4) by computing their global location as $P_{world} = \left[RT_i\right]^{-1} P_{loc}$, where RT_i is the transformation associated with the camera i. We exploit this mapping by first subdividing the model into separate rooms (Sec. 6.1), reconstruct the boundary of each room (Sec. 6.2 and Sec. 6.3), and finally produce a merged 3D model (Sec. 6.4).

6.1 Model partitioning

In order to subdivide the capture environment into separate rooms, we exploit a spatial reasoning approach, based on the occlusions between the poses track and the multi-view 3D points (Fig. 5). As discussed in the previous sections (i.e. Sec. 5), the recovered features are too sparse for a dense 3D point-based reconstruction, however they can provide enough information about strong occlusions along the path, such as a door or a narrowing.

To do this, we project on the same XY plane the feature points and the camera poses. We aggregate the feature points along LSD [15] segments (Fig. 5 left), evaluating when such lines intersect the camera trajectory. We exploit such breaks to divide the poses in groups, also discarding images eventually too close to the intersections (e.g., discarded images are not processed for shape recovery - see Tab. 1), since they most likely contain information that cannot be divided unambiguously between the two parts. Once the images are assigned to a defined space, each room shape is recovered just analyzing only the belonging spheremaps

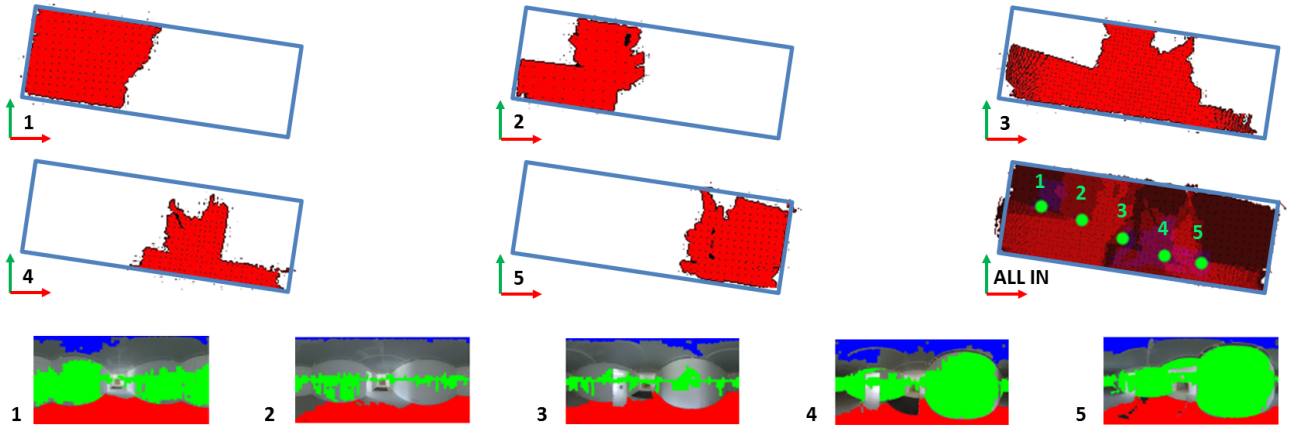


Fig. 4 Facets combination. Example of facets joining from five labeled images (only floor facets are showed to simplify the illustration). Their accumulation array is showed in the last screenshot. Red intensity represents floor labeling occurrences, blue the ceiling occurrences, magenta both ceiling and floor, while green dots show the camera poses.

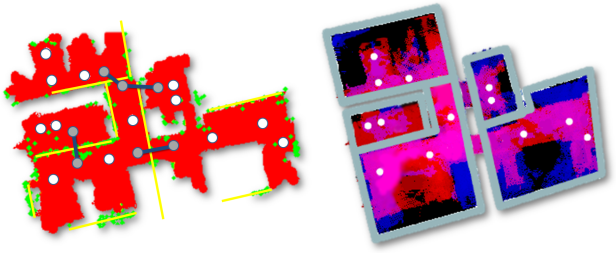


Fig. 5 Multi-room environment (D2 dataset). Left: we arrange the images (positions in white) in different rooms by grouping them (to improve illustration we show only floor facets as background). We exploit 3D features (green dots) to estimate strong occlusions between poses (yellow lines) and breaks among the camera trajectory (blue segments). Poses too close to the occlusions are discarded (grey poses). Right: once the images are grouped each room shape is recovered by projecting only the related images (3D reconstruction showed in Fig. 11).

(Fig. 5 right).

6.2 Room shape reconstruction

For each room, its 2D footprint can in principle be extracted by finding the bounding polygon in the XY plane of all the 3D facets. To do so, we first find the room's 2D bounding rectangle from the projection of 3D facet positions, and, then, project the facets coming from the room images on a regular grid, discretizing that bounding rectangle in order to obtain a footprint mask. The regular grid has a spacing of 4 cm in all the presented experiments. Finding a regularized contour of such a mask would provide us with the room boundary.

However, simply joining all the facets coming from the different cameras associated to the room works would be effective only if their generating super-pixels have been perfectly segmented and classified. However, mostly due to indoor imagery quality and spherical distortion, several errors could affect the labeling and the height assignment, and, therefore, the facet's 3D position.

Fig. 6 shows the effect of a noisy super-pixels segmentation. Super-pixels inside the boxes (Fig. 6 left) have been wrongly labeled as *floor* (actually they were part of the walls), and consequently assigned an incorrect label, height value, and 3D position.

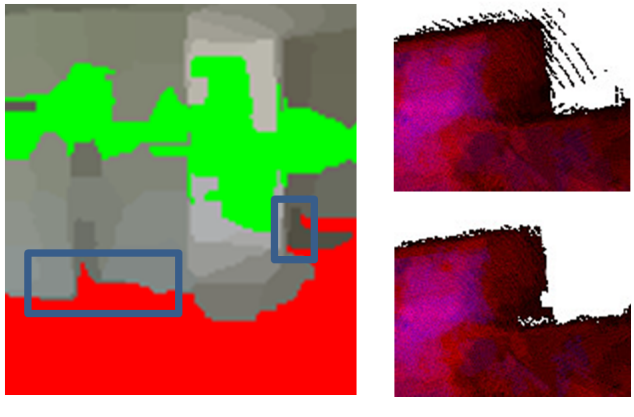


Fig. 6 Wrong classifications filtering. On the left the detail of some super-pixels misclassified. On the right: top, the effect of transforming super-pixels without evaluating their reliability; bottom, the overall results of merging the same part with our accumulation array (*D4* dataset 7, grid size 4 cm).

Although, in the proposed example, the error occurred only in one image, the result affects the entire shape (top-right detail of Fig. 6).

In our method, therefore, we propose a specific approach to join facets and make the reconstruction more robust, outperforming competing solutions [6, 37, 46], with respect to noisy segmentation and texture-less

regions. Since we have more than one labeled view for each part of the scene, we exploit this redundancy to assign a *reliability score* to each 3D point projected, and possibly to discard unreliable results.

In our approach, we project all the 3D points from the *ceiling* and *floor* facets on the XY grid, that we consider an *accumulation array* instead of a simple Boolean mask. Each cell contains the occurrences of a each labeled point, that is how many images cover that cell with the same label. Furthermore, each cell can be at the same time covered by ceiling and floor facets. Joining in the same cell both ceiling and floor contribution, and filtering them by considering the distribution of multi-view contributions, makes the room shape reconstruction more resilient against many clutter problems (i.e., furnitures covering the floor but not the ceiling). We evaluate the mean and the standard deviation σ of the occurrences in the array, then we choose a threshold of 2σ to remove less reliable cells (see for example the bottom right of Fig. 6). We experienced that about 96% of the values lie within the chosen threshold in all tests performed, and that cells with only one or two occurrences are usually discarded. Defining a small threshold on the minimum number of co-occurrences might therefore be an alternative viable alternative for filtering out spurious correspondences.

6.3 Room shape optimization

The final accumulation array provides a good approximation of the room boundary from which to extract the wall geometry. Since walls are vertical, their 2D footprint could be derived simply from the external boundary of the cells in the accumulation array that survive our filtering process. A side effect of such an approach is obviously the eventuality to filter also correct details, for example small peripheral parts of the structure that are barely seen and labeled only from a single image. To compensate for this effect, and to eventually complete parts that are not been labeled as ceiling or floor at all, we exploit in addition the data labeled as *wall* (Sec. 5.1).

We exploit such *wall* contribution as 2D *anchor points*, in combination with the 2D shape recovered from the ceiling and floor footprint. First, we apply an iterative end-point fit algorithm [10] (using as tolerance EPS a 2% of the arc length) to simplify the ceiling and floor footprint contour, obtaining a 2D polygon composed of $S_k(\bar{s_0}, \cdots, \bar{s_k})$ line segments, and we initialize $\bar{R}$ to this first polygonal approximation (Fig. 7, dotted yellow line). Then, we evaluate the initial distance of the *wall* anchor points to the S_k

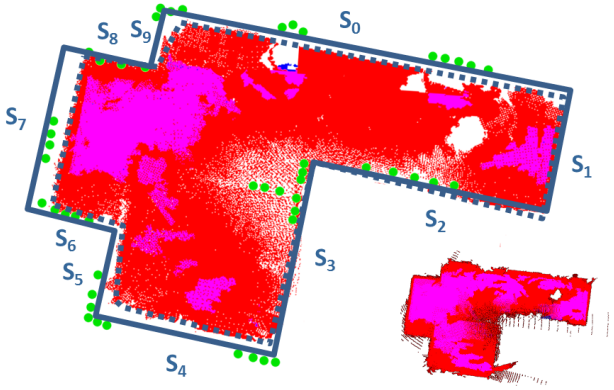


Fig. 7 Shape optimization. We exploit the data labeled as *wall* to perform an optimization on the room shape. For completeness we show in the bottom right thumbnail the footprint shape before filtering (see Sec. 6.2). Differences and points were emphasized to illustrate the method. (D4:Open space room example, see Results 7).

segments, in order to distribute a subset of them in to k $(W_0, \cdots, W_k)$ set of *constant* points, respectively the closest points sets to each $S_k(\bar{s_0}, \cdots, \bar{s_k})$ segment. Given the elements count $W_{i_{count}}$ of each subset points, with $i \in [0, \cdots, k]$, and representing the segments as a varying vector of $2k$ corners $\bar{R}(x_0, y_0, \cdots, x_k, y_k)$ (e.g., s_0 and s_k denotes the same corners in a closed polygon), we formalize the optimization problem as (Eq. 3):

$$R_{2k} \equiv \underset{\bar{R}}{argmin} \sum_{i=0}^{k} \sum_{j=0}^{W_{i_{count}}} dist(W_i(j), \bar{s_i})^2 \qquad (3)$$

which, once expressed in matrix form, can be solved as non-linear least squares problem with Levenberg-Marquardt iterations.

6.4 3D floor plan

Once the 2D shape of the indoor environment has been recovered we exploit the 3D information contained in the closest facets to define a 3D model for each room. An example of a room with sloped ceiling is illustrated in Fig. 8.

In order to generate the 3D room shape from the recovered 2D footprint, we identify the ceiling and floor facets candidates for providing heights. These candidates are found by extracting the ceiling and floor facets whose projection on the XY plane overlaps with a wall segment. Once we have identified the candidate facets, we exploit their heights to generate for each 2D corner a 3D edge formed by two 3D points (e.g., we select Fig. 8, h_0 from floor facets and h_1 from ceiling facets) . Then, in order to consider any intermediate variations in wall heights between original corners (e.g., Fig. 8 double sloped ceiling case), we

Fig. 8 3D room generation. We exploit the 3D information contained in facets closest to the recovered 2D shape to generate the 3D points. In the sloped ceiling case illustrated (*D3-Loft*) 3 height levels (i.e., h_0,h_1,h_2) have been recovered and associated to 3D vertices. Note that the windows and doors are actually sloped in this particular room.

(a) $D1 : Bath1$ (b) $D2 : living$

(c) $D3 : loft$ (d) $D5 : dining$

Fig. 9 Captured panoramic images. We captured many cases of textureless walls and ceilings, as well as moderately to heavily cluttered environments. It should be noted that such clutter is not evident in the reconstructed 3D models, which geometrically include only the boundary data and not the removed clutter.

check the height information of the candidate facets to each wall segment, inserting, in the case of significant variations new vertices in the ceiling 3D shape (e.g., Fig. 8 P_4 and P_7). We use a greedy method that iteratively inserts a new vertex when the difference in height from the current shape and the shape including the vertex has a maximum difference larger than 10cm. The vertices are scanned in order of decreasing error.

The approach proves capable to return a 3D reconstruction with non-trivial space arrangement, large occlusions, or presence of sloped ceilings, assuming that some height levels are detected, even sparsely, by the SfM pipeline. This is typically verified in practice, since the edge between ceiling and wall often lead to the presence of image features.

7 Results

To demonstrate our approach we developed a reconstruction pipeline that, starting from a collection of spherical images and their multi-view alignment, automatically produces a structured 3D floor plan in terms of interconnected rooms bounded by walls. This system has been implemented in C++ on top of *OpenCV*. To obtain camera registration we developed a tool based on the approach of Kangni and Laganiere [21]. Other available tools, such as *PhotoScan* (`http://www.agisoft.com/`), are equally valid for the same purpose.

7.1 Data collection

We evaluated our approach capturing real-world environments. We created ground truth data from on-site inspection aided by laser measures, comparing these reliefs to available blue prints.

We included in our results common indoor scenes, which typically have non-diffuse and homogeneous surfaces. In such typical indoor environments the lack of 3D information from Structure from Motion (SfM) and Multi-View Stereo (MVS) (i.e., Fig.13) makes the approaches based on the direct point-cloud analysis hardly practicable [6]. Furthermore, as the algorithm is explicitly designed to work with partial visibility on cluttered environments, we present results for cluttered scenes (examples of the captured panoramas in Fig. 9).

We captured equirectangular images, covering a full viewport of 360° longitude and 180° latitude, at the resolution of 5376x2688, by using a commodity *Ricoh Theta S* spherical camera (`https://theta360.com/en/about/theta/`). To maximize the bottom hemisphere coverage we mounted the camera on a tripod, also using a fixed distance of 170 cm from the floor, thus exploiting this information to obtain final models in real-world metric dimensions, thus allowing a direct comparison with ground truth. To recover the camera poses and multi-view features we acquired the images so that they always have a sufficient overlap, approximatively capturing at least two images for each room, with a maximum distance of 6 meters between them. We will make all dataset available to allow further studies and comparisons. The acquisition time has been within *20* minutes for each multi-room environment, whereas all reconstruction tests have been performed on an Intel i7 processor with 16GB RAM.

7.2 Room shape reconstruction performance

We present quantitative performances of our method in Tab. 1 detailing results for each room, contextually showing limitations of single-view approaches that also use similar geometric reasoning [37, 46]. To provide

Scene				Time					Error					
				Our				Y. [46]	Our		Y. [46]		P. [37]	
Name	Nc	Np	mq.	SP	Facets	Shape	Tot	Tot	l [%]	a [%]	l [%]	a [%]	l [%]	a [%]
D1:Living	3	2	13	10s	8s	2s	20s	18m09s	1	1	8	8	10	11
D1:Atrium	3	2	7	11s	8s	2s	21s	17m47s	6	7	NS	NS	NS	NS
D1:Corridor	5	4	8	24s	17s	4s	45s	17m43s	1	1	NS	NS	7	9
D1:Passage	3	1	2	6s	10s	1s	17s	19m25s	6	8	NS	NS	8	10
D1:Room1	2	2	12	12s	7s	2s	21s	22m29s	8	9	11	11	12	12
D1:Room2	2	2	8	10s	8s	2s	20s	17m31s	4	5	NS	NS	NS	NS
D1:Bath1	2	1	3	5s	6s	1s	12s	24m47s	4	8	NS	NS	NS	NS
D1:Bath2	2	1	5	5s	5s	1s	11s	22m53s	4	6	NS	NS	10	12
D1:Room3	2	2	10	12s	8s	2s	22s	17m27s	1	1	9	10	5	6
D1:Room4	3	2	12	10s	11s	2s	23s	19m55s	2	6	7	9	10	12
D1:Kitchen	3	2	9	11s	10s	2s	23s	21m41s	6	8	10	11	NS	NS
D2:Bedroom1	4	3	16	13s	12s	3s	28s	20m08s	2	3	NS	NS	12	14
D2:Living	6	3	17	15s	18s	3s	36s	19m42s	4	5	8	9	8	12
D2:Bedroom2	4	3	11	14s	12s	3s	29s	-	3	4	FP	FP	18	19
D2:Restroom	3	2	5	9s	9s	2s	20s	18m31s	4	6	9	9	8	10
D2:Kitchen	3	2	6	10s	9s	2s	21s	17m55s	2	2	NS	NS	12	14
D3:Attic	3	3	8	15s	12s	3s	30s	22m06s	9	10	NS	NS	NS	NS
D3:Loft	5	5	52	25s	20s	5s	50s	20m15s	9	10	NS	NS	NS	NS
D4:Reception	3	3	25	17s	10s	3s	30s	17m16s	8	8	10	12	16	18
D4:Office	5	4	52	20s	16s	3s	39s	17m44s	3	4	9	10	10	10
D4:Open space	11	11	200	1m05s	37s	10s	1m52s	-	8	9	NS	NS	NS	NS
D5:Dining	5	5	36	25s	18s	5s	48s	18m33s	7	8	NS	NS	NS	NS

Tab. 1 Room performance. We show reconstruction performance on real-world multi-room environment, detailing results per room to allow indicative comparison with single-view methods [37, 46]. N_c indicates the total number of images captured, including passages and connections, N_p, instead, is the number of processed images to obtain the shape, followed by the room area. *SP* column shows time effort to compute super-pixels, *Facets* the time to create the labeled facets, *Shape* the time to combine the facets and find the shape in world space. *Tot* columns show respectively the total time to compute the room model with our, Yang et al. [46] and Pintore et al. [37] methods. *NS* means *no structure*, that is when the reconstruction returns a model not comparable with the real ground truth structure. *FP* means failed processing.

such comparison we choose, among the captured poses of each room, the best captured view in terms of space coverage (e.g. maximum number of visible corners).

Scene field in Tab. 1 shows the number of captured poses for each room N_c, which also includes the poses exploited just to track the multi-view features (i.e. in the middle of a door) but not processed for the shape extraction, and the number N_p of poses actually employed for the reconstruction (see Sec.6.4). Beside we indicate the room area in square meters. The *SP* column shows the processing time needed to cluster, label and propagate the classification of about 2048 super-pixels (i.e. image scaled by 4 with respect to its original size) for N_p images. *Facets* column instead reports the time to create each room facets. This value includes the time to register N_c images (i.e. including a fraction of the global bundle adjustment time cost) and the time to create the labeled facets from N_p images. Indeed the number of captured and processed images increases with spatial dimensions and, above all, with the complexity of the environment (i.e., an *U-shape* room requires more images than a box-like room). *Shape* field shows the time to create the accumulation array from the N_p images and recover the 3D shape. *Tot* column summarizes the total time required to automatically generate a room with our method. Beside we indicate, for completeness, the time required to infer a 3D room layout with the single-view approach of Yang et al. [46] (CVPR2016 code snapshot: `https://github.com/YANG-H/Panoramix`). Reported time includes the pre-processing time (which was not indicated in the referenced paper) and the actual time to infer the room layout (about one minute for every room). Comparison highlights how the time required by our method to find the room structure from multiple images is significantly lower than the time required by this other approach to process a single image (e.g. seconds vs. minutes). In the *Error* field we indicate the maximum percent error for walls length e room area, compared with ground truth and the structures recovered with the comparable method of Pintore et al. [37], and with the method of Yang et al. [46], once their result (i.e., obj models) have been manually scaled to metric dimensions. It should be noted that, differently by our and the approach of Pintore et al. [37], the method of Yang et al. [46] is targeted to provide an up-to-scale and cluttered 3D

sketch of an indoor panorama, and not the underlying room structure as conceived in a floor plan. For the sake of clarity we compare numerical values of their method only when their reconstruction provides a comparable footprint of the room.

Numerical results confirm that, even for individual rooms, combining color analysis with multi-view clues is more effective than inferring the whole reconstruction from image segmentation and gradient analysis. As we expected, our approach returns a reliable reconstruction also when the compared approaches fail to find the room structure, such as in presence of hidden corners (i.e., D1:Atrium, D1:Room2), large clutter (i.e., D1:Kitchen, D2:Bedroom1, D3:Lounge), Sloped ceilings (i.e., D3:Attic, D3:Loft) or complex environment containing more than one of these issues at the same time (i.e., D4:Open space). The scenes with sloping

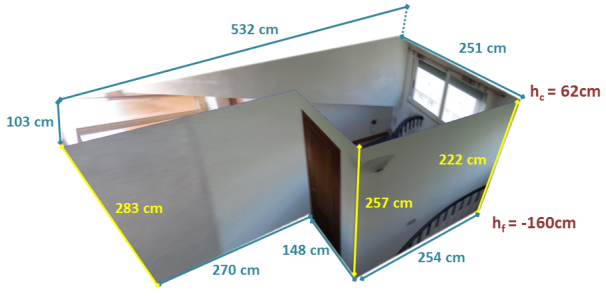


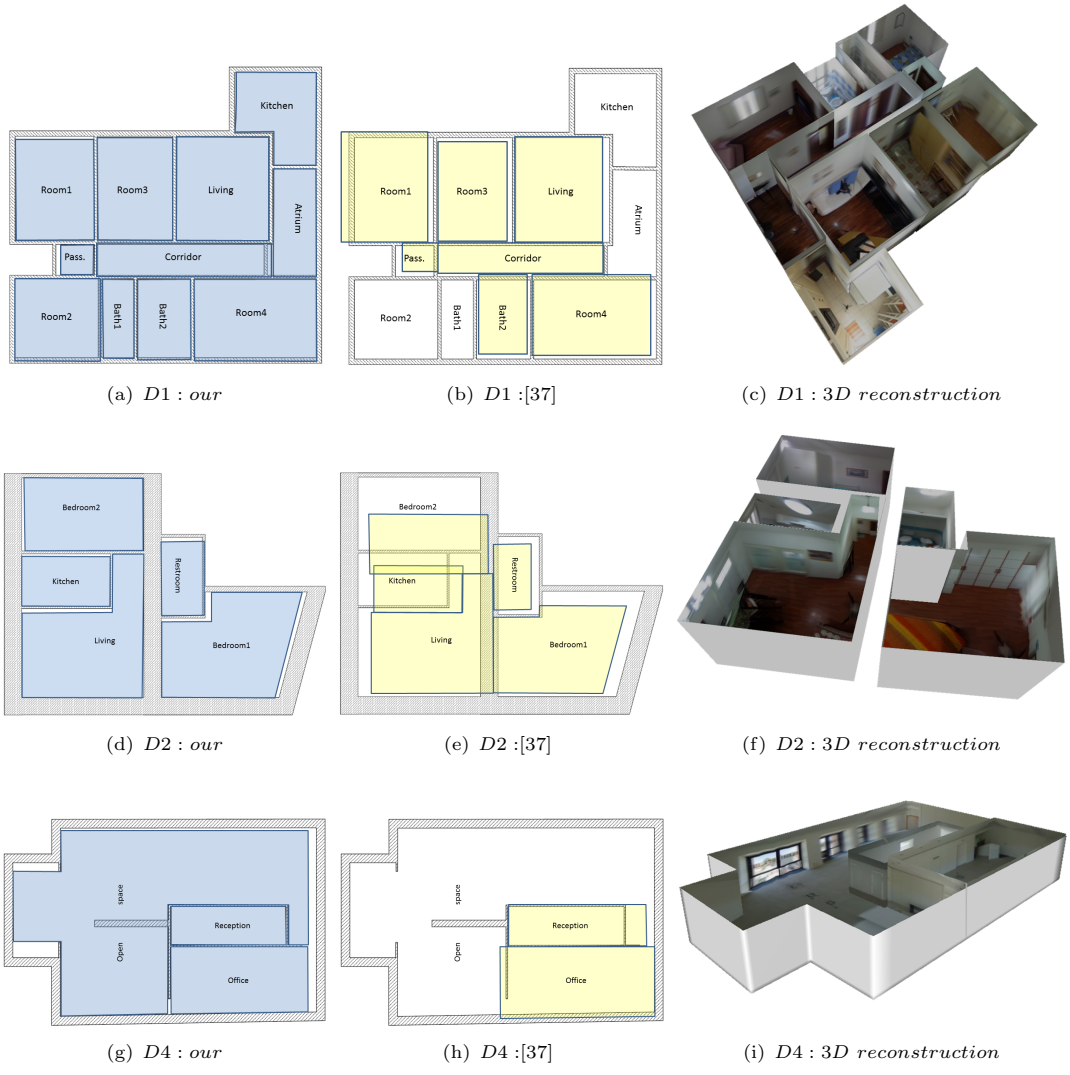
Fig. 10 Heights distribution. We show a sloped ceiling room case (*D3-Attic*).

ceilings highlight the ability of our system to handle scenes with different levels of height, differently by many other approaches [6, 37, 39, 46]. In our results we measured a height error between *6 cm* and *13 cm* in the case of Fig. 10 and Fig. 12(c), and a height error ranging from *8 cm* to *25 cm* in the case of double pitched roofs illustrated in Fig 8 and Fig. 12(e). In both cases we found the largest error on the wall with larger clutter.

7.3 Multi-room performance

Most of the benefits of our method are in its use in multiple and structured environments and, in general, where single view approaches are ineffective or less reliable. In terms of multi-room structure extraction our method is comparable with the method of Pintore et al. [37], which is the most close to ours, although limited by many more assumptions, among which having a single image per room. We exploit for the reconstruction the code provided by the authors [37], adapting their doors matching approach to the use of a spherical panoramic camera (i.e. their original approach was based on panoramic stitching). We show in Fig. 11 the comparison of the reconstructed floor-plans against a real and metrically scaled ground truth (background layer). We also show, besides, the 3D floor plans as textured models. Numerical performance are instead summarized in Tab. 1, detailing results for each room to provide an additional comparison with single-image state-of-the-art approaches (i.e., [46]). It should be noted that metrics such as *Pixel Classification Error* (percentage of pixels that disagree with ground-truth label) are not applicable to our method since our goal is to recover the underlying structure, exploiting parts of many images, which clearly cannot be remapped on the original images and their clutter.

In the first row we show the reconstruction of a typical apartment layout(*D1* dataset). As each room is a fairly regular structure, the main challenges are the splitting of spaces (eleven rooms) and the clutter. Our method 11(a) returns almost perfect spatial mapping and shape for each room, with an overall area error (calculated on real footprint *including* walls thickness), with respect to ground truth, of about *5%*. In the second column 11(b) we show the same environment reconstructed with the approach of Pintore et al. [37] where, mainly because of clutter, the reconstruction of some rooms failed (i.e., the method does not return a measurable reconstruction). Furthermore, due to the rooms joined through doors matching, considerable mapping errors are present. In the second row we present a different kind of structure, that is a residential environment situated within the walls of an old building, characterized by *Non-Manahttan World* corners and very thick walls (*D2* dataset). As our method performs even better than previous case, both in terms of mapping and area error (i.e., overall area error 4%), other approach again presents a higher area error (i.e. 17%) and inaccurate mapping, also due to presence of very thick walls (e.g. 65 *cm*). The third case is a larger and complex structure, where an office layout has been created into a former factory. Such layout is arranged in 3 functional spaces (reception, office, open space) along 290 square meters (*D4* dataset). In particular, the open space is distributed around the central reception and a septal wall, describing an U-shape impossible to be captured with only one view. Moreover, the presence of large areas of homogeneous color and large windows makes this structure hard to recover with a dense method (e.g., test showed at Fig. 13). Also in this case our method returns a reliable reconstruction 11(g) and a very low error, 8%, especially if considering the large size and the peculiar topology. On the other hand

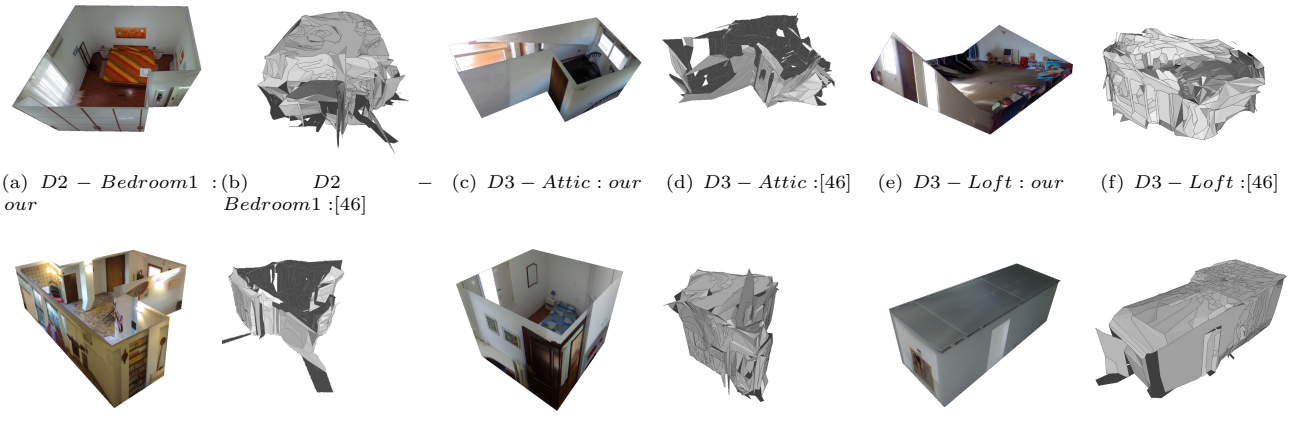


(a) $D1$: *our* (b) $D1$:[37] (c) $D1$: $3D$ *reconstruction*

(d) $D2$: *our* (e) $D2$:[37] (f) $D2$: $3D$ *reconstruction*

(g) $D4$: *our* (h) $D4$:[37] (i) $D4$: $3D$ *reconstruction*

Fig. 11 Recovered footprint and 3D models vs. ground truth floor plan. Comparisons against real, metrically scaled, ground truth (grey footprint), of our method (first column) and the multi-room approach of Pintore et al. [37] (second column). We show in the third column our final textured 3D floor plan. Ceilings and septal wall have been removed from 3D reconstruction to make the illustration more clear.



(a) $D2 - Bedroom1$: our (b) $D2 - Bedroom1$:[46] (c) $D3 - Attic$: our (d) $D3 - Attic$:[46] (e) $D3 - Loft$: our (f) $D3 - Loft$:[46]

(g) $D3 - Dining$: our (h) $D3 - Dining$:[46] (i) $D1 - Room2$: our (j) $D1 - Room2$:[46] (k) $D4 - reception$: our (l) $D4 - reception$:[46]

Fig. 12 Comparison with respect to a single-view approach We show some results of single room environment reconstructed with our approach, where single-view approaches tend to fail. We show, beside our reconstruction, the output of a single-view approach [46] for some of the cases marked as *NS* (no structure) in Tab. 1. We show, instead, in the last illustrations our 12(k) and [46] 12(l) reconstruction on a comparable case.

the compared approach, as it expect rooms where all corners are visible from a single point-of-view, definitely fails the reconstruction of the main room 11(h).

In should be noted that, in contrast to approaches that need an adequate number of 3D points to determine the room shapes [6], our method can effectively work on texture-poor environments, such as the presented cases. As we use SfM essentially for determining camera pose, after which we can work even if there is just a single 3D point per room in the case of horizontal ceiling and floor, and need more only in the presence of sloped ceilings.

7.4 Qualitative comparison

We show in Fig. 12 qualitative performances of our method on some cases where single-view approaches are not able to recover the underlying room structure (*NS* in tab. 1), and a visual comparison with the output of the Yang et al. pipeline [46] (e.g. a visual comparison with the method of Pintore et al. [37] was presented for the multi-room case (Fig. 11). Fig. 12(a) shows a room with non-Manhattan World corners, large clutter and a mirror on the wall. As our method correctly recover the room shape, the other approach misses a large fraction of the room walls, which are occluded by the wardrobe, as well as reconstructs wrong parts in the presence of a bed (e.g., *diagonal* lines motif between wall and floor leads to wrong vanishing lines estimation), a mirror and an open door. Fig. 12(c) and Fig. 12(e) show our reconstruction of single and double sloped ceiling environments. Incomplete reconstruction showed in Fig. 12(d) and Fig. 12(f), instead, highlight one of

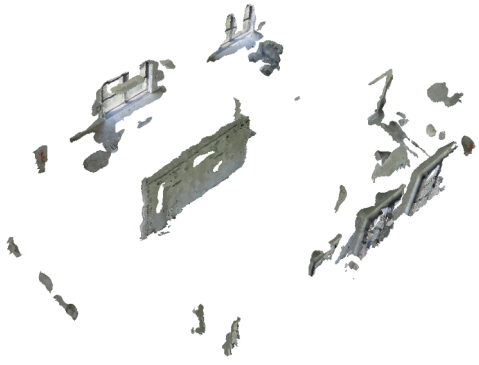


Fig. 13 Reconstruction from the point cloud only We show the output of a dense multi-view pipeline (i.e., *PhotoScan*) applied on the *D4* dataset. Mesh extraction and tiled model recovery have been also applied to enhance illustration.

the most critical points of the single view approaches applied to spherical images, namely the way in which the lines are extracted to provide a geometric context. In an equirectangular projection, in-fact, lines are not usually directly detectable (except for the vertical ones), but arbitrary perspective projections are generated to find them in an undistorted space [25], and then detected lines are transformed them back again in the original space. This approach works very efficiently for lines close to Manhattan World directions [47], but tends to fail for less conventional directions, as in the proposed examples. Fig. 12(h) and Fig. 12(j) show the reconstruction obtained from a single-point-view compared with ours (Fig. 12(g) and Fig. 12(i)), both due to clutter. In addition, Fig. 12(b) and Fig. 12(h)

examples highlight misclassification problems discussed at Sec. 6.2. We show instead in the last comparison (Fig. 12(k) vs. Fig. 12(l)) a case where the single-view approach returns a reliable structural reconstruction.

For completeness in Fig. 13 we show the output of a standard dense multi-view pipeline, applied on the same data reconstructed by our method in Fig. 11.

The reconstruction has been performed using *PhotoScan* on the original panoramic images, running respectively camera alignment and point cloud densification. As showed by the examples, such reconstructions contain several sparse details of interior clutter, but lack structural parts of the rooms (i.e., all ceilings and external walls), thus making methods that derive the structure from the point cloud unfeasible [6, 33].

Fig. 14 Failure case. We show a scene with open ceiling, stairs and a curved wall where our method failed to recover the structure. We show some of the original captured images, the resulting (wrong) transform in the 2D plane and the expected real shape (yellow).

In terms of limitations, our method targets the reconstruction of indoor environments in terms of rooms bounded by walls, ceilings and floors. We do not, thus, handle the rconstruction of furniture or additional architectural elements such as stairs. This is because the method explicitly looks only to reconstruct the bounding volume of rooms, together with multi-room connections.

Differently from *Manhattan World*, we do not require vertical planes to be orthogonal with respect to each other, and differently from *Manhattan World* [40], we can also handle sloped ceilings. We only assume that walls are vertical.

For the reconstruction to be successful, see Sec. 5.2, our method requires that there is at least one height coming from SfM in a connected labeled region, in order to automatically recover the ceiling height. In the case of sloped ceilings, we must have enough features to reconstruct the slope. While curved vertical walls can, in principle, be handled if enough features are present to define their 2D footprint, obtaining them is often a problem in practice, and often leads to failures (Fig. 14).

Since many indoor scenes, especially in office and apartment buildings, meet our method's assumptions [41] the above limitations can be considered acceptable.

8 Conclusions

We presented a novel and practical approach for recovering 3D indoor structures using low-cost 360° cameras. Our work has introduced several improvements over prior approaches aimed at extracting structural information without requiring a dense capture. In particular, our framework based on 3D facets combines a new approach for geometric context extraction, with a new technique for combining facets from different points of view in a single consistent 3D model, without strictly imposing Manhattan World constraints. As illustrated with our results, only few overlapping images are required to generate a 3D floor plan, even when other previous approaches fail, such as in presence of hidden corners, large clutter and more complex multi-room structures.

We envision, as a future work, to extend the mixing of single-view and multi-view labeling to extract other structural information from the data, such as the clutter in the rooms, in order to create a complete furnished 3D model.

Acknowledgements

This work was partially supported by projects VIGEC and 3DCLOUDPRO. The authors also acknowledge the contribution of Sardinian Regional Authorities.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 34(11):2274–2282, 2012.

[2] Autodesk. 123D Catch. `www.123dapp.com/catch`.

[3] S. Y. Bao, A. Furlan, L. Fei-Fei, and S. Savarese. Understanding the 3D layout of a cluttered room from

multiple images. In *Proc. IEEE WACV*, pages 690–697, 2014.

[4] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, Aug 2007.

[5] R. Bunschoten and B. Krose. Robust scene reconstruction from an omnidirectional vision system. *IEEE Transactions on Robotics and Automation*, 19(2):351–357, 2003.

[6] R. Cabral and Y. Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *Proc. CVPR*, pages 628–635, 2014.

[7] D. Caruso, J. Engel, and D. Cremers. Large-scale direct SLAM for omnidirectional cameras. In *Proc. IEEE IROS*, pages 141–148, 2015.

[8] P. Chang and M. Hebert. Omni-directional structure from motion. In *Proc. IEEE Workshop on Omnidirectional Vision*, pages 127–133, 2000.

[9] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proc. IEEE ICCV*, volume 2, pages 941–947, 1999.

[10] D. H. Douglas and T. K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973.

[11] A. Flint, D. Murray, and I. Reid. Manhattan scene understanding using monocular, stereo, and 3D features. In *Proc. IEEE ICCV*, pages 2228–2235, 2011.

[12] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *Proc. IEEE ICCV*, pages 80–87, 2009.

[13] C. Geyer and K. Daniilidis. A unifying theory for central panoramic systems and practical implications. In *Proc. ECCV*, pages 445–461, 2000.

[14] Google. Tango, 2014. `www.google.com/atap/projecttango/`.

[15] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: a line segment detector. *Image Processing On Line*, 2:35–55, 2012.

[16] R. Guo and D. Hoiem. Support surface prediction in indoor scenes. In *Proc. IEEE ICCV*, pages 2144–2151, 2013.

[17] C. Hane, L. Heng, G. H. Lee, A. Sizov, and M. Pollefeys. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *Proc. 3DV*, pages 57–64, 2014.

[18] S. Ikehata, H. Yang, and Y. Furukawa. Structured indoor modeling. In *Proc. IEEE ICCV*, pages 1323–1331, 2015.

[19] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. S. Kweon. *All-Around Depth from Small Motion with a Spherical Panoramic Camera*, pages 156–172. 2016.

[20] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3D-based reasoning with blocks, support, and stability. In *Proc. IEEE CVPR*, pages 1–8, 2013.

[21] F. Kangni and R. Laganiere. Orientation and pose recovery from spherical panoramas. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

[22] H. Kim and A. Hilton. 3D scene reconstruction from multiple spherical stereo pairs. *International Journal of Computer Vision*, 104(1):94–116, 2013.

[23] Y. M. Kim, N. J. Mitra, D.-M. Yan, and L. Guibas. Acquiring 3d indoor environments with variability and repetition. *ACM TOG*, 31(6):138:1–138:11, 2012.

[24] J. Kopf. 360° video stabilization. *ACM Trans. Graph.*, 35(6):195:1–195:9, Nov. 2016.

[25] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Proc. CVPR*, pages 2136–2143, 2009.

[26] S. Li. Binocular spherical stereo. *IEEE TITS*, 9(4):589–600, 2008.

[27] R. Marroquim, M. Kraus, and P. R. Cavalcanti. Special section: Point-based graphics: Efficient image reconstruction for point-based and line-based rendering. *Comput. Graph.*, 32(2):189–203, 2008.

[28] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski. Low-cost 360 stereo photography and video capture. *ACM Trans. Graph.*, 36(4):148:1–148:12, July 2017.

[29] Microsoft. Photosynth. `photosynth.net/`.

[30] B. Micusik and T. Pajdla. Autocalibration 3D reconstruction with non-central catadioptric cameras. In *Proc. IEEE CVPR*, pages I–58–I–65, 2004.

[31] B. Micusik and T. Pajdla. Structure from motion with wide circular field of view cameras. *IEEE TPAMI*, 28(7):1135–1149, 2006.

[32] C. Mura, O. Mattausch, A. Jaspe Villanueva, E. Gobbetti, and R. Pajarola. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers & Graphics*, 44:20–32, 2014.

[33] C. Mura, O. Mattausch, and R. Pajarola. Piecewise-planar reconstruction of multi-room interiors with arbitrary wall arrangements. *Computer Graphics Forum*, 35(7):179–188, 2016.

[34] L. Nan, K. Xie, and A. Sharf. A search-classify approach for cluttered indoor scene understanding. *ACM TOG*, 31(6):137:1–137:10, 2012.

[35] G. Pintore, F. Ganovelli, E. Gobbetti, and R. Scopigno. Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness. In *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 130–145. Springer, October 2016.

[36] G. Pintore, F. Ganovelli, R. Pintus, R. Scopigno, and E. Gobbetti. Recovering 3d indoor floor plans by exploiting low-cost spherical photography. In *PG2018 Short Papers Proceedings*, 2018. To appear.

[37] G. Pintore, V. Garro, F. Ganovelli, M. Agus, and E. Gobbetti. Omnidirectional image capture on mobile devices for fast automatic generation of 2.5D indoor maps. In *Proc. IEEE WACV*, pages 1–9, 2016.

[38] G. Pintore and E. Gobbetti. Effective mobile mapping

of multi-room indoor structures. *The Visual Computer*, 30, 2014.

[39] G. Pintore, R. Pintus, F. Ganovelli, R. Scopigno, and E. Gobbetti. Recovering 3d existing-conditions of indoor structures from spherical images. *Computers & Graphics*, 2018.

[40] G. Schindler and F. Dellaert. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–203–I–209 Vol.1, June 2004.

[41] A. G. Schwing and R. Urtasun. *Efficient Exact Inference for 3D Indoor Scene Understanding*, pages 299–313. 2012.

[42] M. Schnbein and A. Geiger. Omnidirectional 3D reconstruction in augmented Manhattan worlds. In *Proc. IEEE IROS*, pages 716–723, 2014.

[43] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, volume 1, pages 519–528, 2006.

[44] G. Tsai, C. Xu, J. Liu, and B. Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *Proc. IEEE ICCV*, pages 121–128, 2011.

[45] X. Xiong, A. Adan, B. Akinci, and D. Huber. Automatic creation of semantically rich 3D building models from laser scanner data. *Automation in Construction*, 31:325–337, 2013.

[46] H. Yang and H. Zhang. Efficient 3D room shape recovery from a single panorama. In *Proc. IEEE CVPR*, pages 5422–5430, 2016.

[47] Y. Zhang, S. Song, P. Tan, and J. Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 668–686, Cham, 2014. Springer International Publishing.

[48] S. Zingg, D. Scaramuzza, S. Weiss, and R. Siegwart. MAV navigation through indoor corridors using optical flow. In *Proc. IEEE IROS*, pages 3361–3368, 2010.

Giovanni Pintore is a researcher in the Visual Computing (ViC) group at the Center for Advanced Studies, Research, and Development in Sardinia (CRS4). He holds a Laurea (M. Sc.) degree (2002) in Electronics Engineering from the University of Cagliari. His research interests include multiresolution representations of large and complex 3D models, lightfield displays, reconstruction and rendering of architectural scenes exploiting mobile devices and the new generation mobile spherical cameras.

Fabio Ganovelli is a researcher in the Visual Computing Lab at ISTI-CNR in Pisa. He holds a Laurea (1995) and a Ph.D. degree (2001) in Computer Science from the University of Pisa. His research spans many areas of computer graphics and computer vision and is widely published in major journals and conferences.

Ruggero Pintus is a researcher in the Visual Computing (ViC) group at the Center for Advanced Studies, Research, and Development in Sardinia (CRS4). He holds a Laurea (M.Sc., 2003) and a Ph.D. degree (2007) in Electronics Engineering from the University of Cagliari, Italy. His research is currently focusing on acquisition, processing, and rendering of complex 3D models.

Roberto Scopigno graduated in Computer Science at Univ. of Pisa in 1984. He is a Research Director with CNR-ISTI and leads the Visual Computing Lab. He has been engaged in research projects concerned with scientific visualization, multi-resolution technologies, 3D range digitization, and CH applications. He has published more than 200 papers in international journals or conferences.

Enrico Gobbetti is the director of Visual Computing (ViC) at the Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Italy. He holds an Engineering degree (1989) and a Ph.D. degree (1993) in Computer Science from the Swiss Federal Institute of Technology in Lausanne (EPFL). Prior to joining CRS4, he held research and teaching positions at EPFL, UMBC, and NASA. Enrico's research spans many

areas of visual computing and is widely published in major journals and conferences.