

Joint Layout Estimation and Global Multi-View Registration for Indoor Reconstruction

Jeong-Kyun Lee^{*1}, Jaewon Yea^{*2}, Min-Gyu Park³, and Kuk-Jin Yoon^{†1}

¹Computer Vision Laboratory, GIST, Gwangju, South Korea

²LG Electronics, Incheon, South Korea

³Korea Electronics Technology Institute, Seongnam, South Korea

{leejk,kjyoon}@gist.ac.kr, jaewon.yea@lge.com, mpark@keti.re.kr

Abstract

In this paper, we propose a novel method to jointly solve scene layout estimation and global registration problems for accurate indoor 3D reconstruction. Given a sequence of range data, we first build a set of scene fragments using KinectFusion and register them through pose graph optimization. Afterwards, we alternate between layout estimation and layout-based global registration processes in iterative fashion to complement each other. We extract the scene layout through hierarchical agglomerative clustering and energy-based multi-model fitting in consideration of noisy measurements. Having the estimated scene layout in one hand, we register all the range data through the global iterative closest point algorithm where the positions of 3D points that belong to the layout such as walls and a ceiling are constrained to be close to the layout. We experimentally verify the proposed method with the publicly available synthetic and real-world datasets in both quantitative and qualitative ways.

1. Introduction

The popularization of low-cost consumer depth cameras has made a new perspective of solving various computer vision problems. Especially, with various depth sensors, simultaneous localization and mapping (SLAM) and 3D reconstruction [9, 27] have shown visually compelling results compared to conventional image-based approaches in an indoor environment. This is because the consumer depth camera robustly acquires depth measurements where the conventional image-based approaches frequently fail to esti-

^{*}Denote equal contribution.

[†]Corresponding author.

This work was done when Jaewon Yea and Min-Gyu Park were members of the Computer Vision Laboratory at GIST.

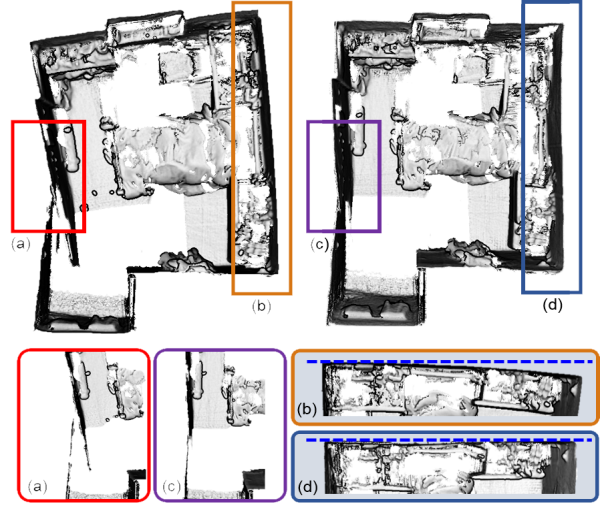


Figure 1. Comparison of the proposed method (right) with the state-of-the-art method [9] (left). With the aid of the scene layout, the proposed method preserves important structures of the scene such as walls and a floor.

mate accurate depth, *e.g.*, due to poorly textured regions. In this paper, we narrow our attention to the complete 3D reconstruction problem in an indoor environment using range measurements acquired from a consumer depth camera.

KinectFusion [27], one of the pioneering works, showed that a real-world object as well as an indoor scene can be reconstructed in real-time with GPU acceleration. It exploits the iterative closest point (ICP) algorithm [4] to track 6-DoF poses and the volumetric surface representation scheme with signed distance functions [12] to fuse 3D measurements. A number of following studies [9, 41, 43] have tackled the limitation of KinectFusion; as the scale of a scene increases, it is hard to completely reconstruct the scene due to the drift problem of the ICP algorithm as well as the large memory consumption of volumetric inte-

gration. To scale up the KinectFusion algorithm, Whelan *et al.* [41] presented a spatially extended KinectFusion, named as Kintinuous, by incrementally adding KinectFusion results as the form of triangular meshes. Moreover, they used a pose graph to alleviate the drift problem through graph optimization by identifying loop closures. Whelan *et al.* [43] also proposed ElasticFusion to overcome the problem by using the surface loop closure optimization and the surfel-based representation. On the other hand, this large-scale indoor reconstruction problem has been tackled from the view point of global registration [1, 9, 37]. Notably, Choi *et al.* [9] showed promising results. They utilized KinectFusion results as building blocks and developed a robust global registration scheme based on line-processes in the presence of sensor noise.

Furthermore, the large-scale indoor reconstruction problem has been tackled by considering the structural regularities of an indoor scene such as the axis-aligned geometry [45] and the planarity of the scene [25, 36, 48]. Xiao and Furukawa [45] showed that a museum-level indoor environment can be effectively reconstructed based on the Manhattan world assumption, *e.g.* walls, floors, and ceilings are parallel to one of the three orthogonal surfaces. However, this strong assumption about the scene generates oversimplified structures in practical situations. To avoid the potential oversimplification problem, Zhang *et al.* [48] analyzed planar and non-planar regions on the fly and integrated KinectFusion’s results seamlessly.

In this paper, we pose a new approach for accurate indoor reconstruction by jointly resolving the scene layout estimation problem and the global registration problem of range data. Given initially registered range data, we extract the envelope of an indoor scene, including walls, a floor, and a ceiling, through hierarchical agglomerative clustering and energy-based multi-model fitting to reduce redundant planes and to find dominant plane hypotheses. Then, we establish point-to-layout correspondences to constrain the position of these correspondences to be close to the layout. Finally, we register entire range data through the global ICP algorithm with pairwise and layout-based constraints. We repeat layout estimation and layout-based global registration procedures alternately until they converge. Note that we purely rely on range data and the layout is computed with a weak Manhattan world assumption, such that walls are not necessarily perpendicular to each other but perpendicular to the floor and the ceiling.

2. Previous Work

The literature review primarily focuses on indoor 3D reconstruction starting from KinectFusion [27] and its follow-up researches that aim at scaling-up KinectFusion.

After the breakthrough of Newcombe *et al.* [27], hundreds of papers have addressed the limitations of KinectFu-

sion. To overcome the scalability of the volumetric reconstruction approach, some works [41, 33, 21] adopted the concept of moving volume, which translates and rotates a reconstructed volume by using the estimated pose information. In the similar manner, Steinbrucker *et al.* [38] proposed a multi-scale octree data structure to modify the uniform volumetric structure into non-uniform volume. Henry *et al.* [20] proposed a multiple-volume representation to create a globally consistent indoor environment. Chen *et al.* [7] and Nießner *et al.* [28] proposed a memory-efficient hierarchical data structure for commodity graphics hardware to extend KinectFusion to large-scale scenes.

On the other hand, numerous researches focused on alleviating accumulation errors in reconstructing a large-scale environment. Conventional pairwise approaches [8, 32] as well as KinectFusion incrementally integrate a set of range data using the ICP algorithm; they suffer from accumulated errors in general. Therefore, the global registration approaches [3, 44, 29, 5, 37, 18, 24, 47, 1] have been developed to alleviate accumulation errors by optimizing the global poses simultaneously. Bergevin *et al.* [3] proposed a star-shaped network for global registration. Each range view can be interconnected with the world reference frame by sequential transformation multiplications. The transformations between the range views and the world frame are alternately optimized by the point-to-plane method [8]. Nishino and Ikeuchi [29] proposed a robust global registration method based on the M-estimator [22] to improve robustness against outlier correspondences. Arrigoni *et al.* [1] proposed the global registration method based on the low-rank and sparse (LRS) decomposition.

For the sake of robust global registration, a number of researches [5, 37] focused on identifying loop closures which must be acquired for global registration. As long as loop closures are properly identified, it is possible to reduce accumulation errors effectively. Several researches [41, 42, 38, 20] used visual features to identify loop closures, but they showed failure cases under pose variations as well as when revisiting poorly textured regions. Therefore, some researches [34, 14, 11, 9] focused on solving this problem with geometric features. Assuming loop closures are given from a set of accurate correspondences, the pose graph optimization scheme has been widely employed [41, 42, 38, 20, 9] owing to its real-time performance. It optimizes a pose graph constrained by pairwise transformations to balance the accumulated error. Tang and Feng [47] proposed the method to distribute the accumulated error by integrating loops incrementally. They minimized the bi-directional registration errors [24] of the virtual point pairs [32] using the global optimization technique [44].

In addition, a number of researches focused on the structural regularity of the indoor scene to elevate the quality

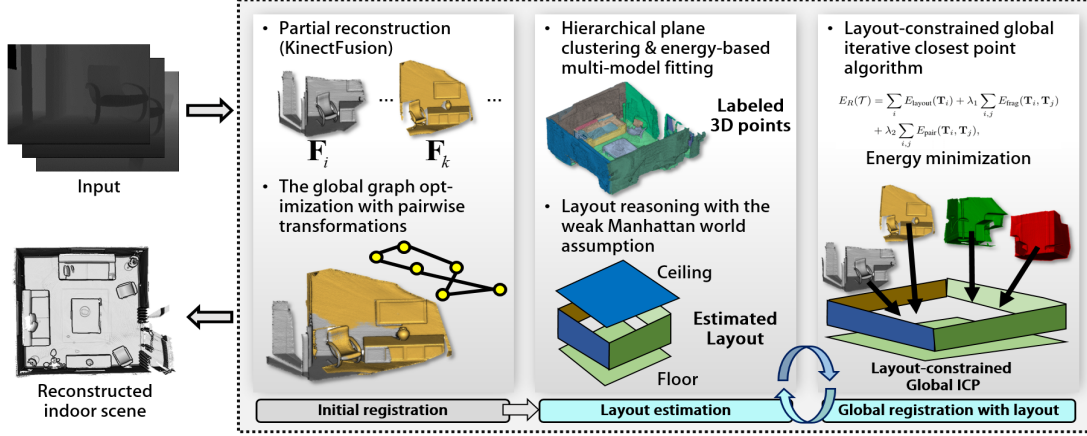


Figure 2. The overall procedure of the proposed method.

of reconstructed models. Basically, the planarity assumption [13, 2, 39, 48, 36] is the most commonly used one. Moreno *et al.* [36] proposed an incremental plane mapping scheme in which the relation between planes is identified by point features. Several studies [13, 2, 39] exploited planes and points to find frame-to-frame camera pose and to define an objective function for bundle adjustment. Ma *et al.* [25] estimated a global plane model and frame-to-frame pose in an alternative way in the EM framework. Zhang *et al.* [48] proposed an interactive reconstruction algorithm, in which the algorithm guides the person to capture designated spots.

The proposed method overcomes aforementioned problems through the layout-constrained global registration. The scene layout estimation problem has been tackled in the field of scene understanding [10, 15, 19] and object detection [17]. Some researches [30, 40, 50] proposed to enforce the global regularity (*e.g.* parallelism, orthogonality, and coplanarity) of the scene structures in an iterative fashion, assuming that *well-aligned* but noisy point clouds are given as input. However, we consider *inaccurately aligned* point clouds, *i.e.*, owing to drift errors as shown in the left of Fig. 1. Therefore, we perform the layout estimation and global registration jointly, and in particular, the proposed dominant plane estimation based on energy minimization provides locally optimal dominant planes without regard to the general global regularities.

3. Proposed Method

We propose a joint approach of scene layout estimation and global multi-view registration for 3D indoor reconstruction. As shown in Fig. 2, the overall procedure of the proposed method consists of three main steps: initial registration, layout estimation, and global registration. In the initial registration step, we sequentially construct scene fragments from range data and then align them in the world coordinate system. Afterward, we alternate between layout estimation and global registration procedures in iterative fashion using

scene fragments. The model reconstructed by the global registration is refined using [49].

3.1. Initial Registration

For initial registration, we partially reconstruct the captured indoor scene to produce a set of scene fragments and then register them in the world coordinate system, which is similar to the previous study of Choi *et al.* [9]. Here, the underlying assumption is that each scene fragment contains a negligible amount of accumulation errors so that the large-scale 3D reconstruction problem turns into the problem of aligning all the scene fragments. To construct a scene fragment $\mathbf{F}_i \in \mathcal{F}$, we simply use KinectFusion [27] for every N frames, *e.g.* 50, which is a volumetric approach to reconstruct a scene with truncated signed distance functions [12]. Afterwards, we find pairwise transformations $\mathbf{T}_{i,i+1}$ for all pairs of the consecutive fragments and align all the fragments in the world coordinate system based on sequential multiplication of the pairwise transformations.

Loop closure detection: The set of the registered fragments via the sequential multiplication of the pairwise transformations usually has a large amount of accumulated pose errors as well as misaligned range data. Therefore, it is necessary to identify loop closures to diffuse drift errors across all the fragments. To detect loop closures, we align all pairs of the inconsecutive fragments using the FPFH descriptor [34] and check the overlap ratio of the aligned fragments. If the overlapping ratio between the fragments \mathbf{F}_i and \mathbf{F}_j exceeds a predefined percentile, *e.g.* 30%, we determine the fragment pair as a loop closure and define its pairwise transformation as $\mathbf{T}_{i,j}$.

Pose graph optimization: Given a set of loop closures, we minimize the drift errors through the pose graph optimization. Here, we adopt the line process as in [9] to handle spurious loop closures obtained by low distinctiveness of 3D local descriptors. For a set of fragments $\mathcal{F} = \{\mathbf{F}_0, \dots, \mathbf{F}_n\}$, we define a set of transformations $\mathcal{T} = \{\mathbf{T}_0, \dots, \mathbf{T}_n\}$ where

\mathbf{T}_i is a transformation from a fragment \mathbf{F}_i to world reference coordinates and a pairwise transformation from \mathbf{F}_j to \mathbf{F}_i is expressed as $\mathbf{T}_{i,j} = \mathbf{T}_i^{-1} \circ \mathbf{T}_j$. Then, given pairwise transformations $\mathbf{T}_{i,j}$ between the fragments \mathbf{F}_i and \mathbf{F}_j , we estimate the transformations \mathcal{T} of the fragments \mathcal{F} and a line process \mathcal{L} by minimizing the following function,

$$E_L(\mathcal{T}, \mathcal{L}) = \sum_i f(\mathbf{T}_i, \mathbf{T}_{i+1}, \hat{\mathbf{T}}_{i,i+1}) + \sum_{i,j} l_{ij} f(\mathbf{T}_i, \mathbf{T}_j, \hat{\mathbf{T}}_{i,j}) + \sum_{i,j} \Psi(l_{ij}), \quad (1)$$

where $f(\mathbf{T}_a, \mathbf{T}_b, \hat{\mathbf{T}}_{a,b})$ measures the difference between the pre-computed pairwise transformation $\hat{\mathbf{T}}_{a,b}$ and the pairwise transformation computed from \mathbf{T}_a and \mathbf{T}_b . $l_{ij} \in \mathcal{L}$ is a parameter of a line process. $\Psi(l_{ij}) = \sqrt{1 - l_{ij}^2}$ is a constraint to maximize the number of inlier loop closures. If an estimated parameter l_{ij} is larger than a threshold, the loop closure between the fragments i and j is determined as a correct loop closure.

3.2. Layout Estimation

To estimate the scene layout, which consists of a set of planes such as a ceiling, a floor, and walls, we find the dominant planes $\mathcal{P}_{\text{dominant}}$ in the scene and then determine layout planes $\mathcal{P}_{\text{layout}}$ from $\mathcal{P}_{\text{dominant}}$. To extract a set of dominant planes, we compute and cluster plane parameters from supervoxels [31] of each fragment and subsequently merge similar plane parameters in the world coordinate system.

Dominant plane extraction: Initially, we divide a fragment \mathbf{F}_i into a set of supervoxels [31], $\mathcal{S} = \{\mathcal{S}_1 \cup \mathcal{S}_2 \cup \dots \cup \mathcal{S}_K\}$, and generate plane hypotheses using the supervoxels. To generate a plane hypothesis π_l in the fragment \mathbf{F}_i , we compute a plane parameter from the center points of three adjacent supervoxels because it improves computational efficiency in comparison with the way that a plane parameter is computed using all the points in a sampled supervoxel. Here, the number of initial plane hypotheses proportionally increases as the scale of a scene increases, and there might be a lot of similar planes owing to largely planar regions such as walls. Therefore, we cluster initial plane hypotheses through two plane clustering steps. First, we merge the plane hypotheses using the hierarchical agglomerative clustering [26]. The distance between a supervoxel \mathcal{S}_k and a plane hypothesis π_l is computed as

$$C(\pi_l, \mathcal{S}_k) = \frac{1}{|\mathcal{S}_k|} \sum_{\mathbf{p} \in \mathcal{S}_k} d(\pi_l, \mathbf{p}). \quad (2)$$

The distance function $d(\cdot, \cdot)$ is defined as

$$d(\pi_l, \mathbf{p}) = \frac{|\pi_l^\top \bar{\mathbf{p}}|}{\sqrt{a_l^2 + b_l^2 + c_l^2}}, \quad (3)$$

where a plane parameter π_l is denoted as $\pi_l = [a_l, b_l, c_l, 1]^\top$ and $\bar{\mathbf{p}}$ is a homogeneous representation of a 3D point \mathbf{p} . Some supervoxels with similar plane hypotheses are grouped together via the clustering method and used to recalculate plane parameters. However, there still exist some plane hypotheses that are on an identical wall but not grouped together because of local distortion in the vicinity of the fragment's border. Thus, as the second step, we assign the recomputed plane hypotheses to each 3D point by minimizing an energy function via graph cuts [6]. Given a set of 3D points, *i.e.*, \mathbf{F}_i , and a set of plane parameters, denoted by \mathcal{P}_i , the problem is defined as finding a mapping function h from a point $\mathbf{p} \in \mathbf{F}_i$ onto a plane parameter $\pi_l \in \mathcal{P}_i$ (*i.e.*, $h: \mathbf{F}_i \mapsto \mathcal{P}_i$). An energy function E_P is defined as

$$E_P(h) = \sum_{\mathbf{p} \in \mathbf{F}_i} D_{\mathbf{p}}(h_{\mathbf{p}}) + \sum_{\mathbf{p} \in \mathbf{F}_i, \mathbf{q} \in \mathcal{N}_{\mathbf{p}}} V_{\mathbf{p},\mathbf{q}}(h_{\mathbf{p}}, h_{\mathbf{q}}). \quad (4)$$

The data term $D_{\mathbf{p}}$ is defined in the same manner as Eq. (3) to measure the distance between a point \mathbf{p} and a plane parameter π_l . We employ the Potts model [6] as the smoothness term $V_{\mathbf{p},\mathbf{q}}$ to preserve continuity of a plane parameter between neighboring points. The Potts model is defined as $V_{\mathbf{p},\mathbf{q}}(h_{\mathbf{p}}, h_{\mathbf{q}}) = \alpha_{\mathbf{p},\mathbf{q}} T(h_{\mathbf{p}} \neq h_{\mathbf{q}})$ where $\alpha_{\mathbf{p},\mathbf{q}}$ is a penalty weight and T is 1 if the argument is true and otherwise 0. $\mathcal{N}_{\mathbf{p}}$ represents neighboring points of \mathbf{p} . The neighboring points are determined as points within a predefined distance among points obtained by the k -nearest neighbor (k -NN) search algorithm [35]. In addition, we employ a null-plane hypothesis π_\emptyset to avoid assigning plane hypotheses to a point that has a large displacement from the plane. Therefore, the data term is redefined as

$$D_{\mathbf{p}}(h_{\mathbf{p}}) = \begin{cases} d(h_{\mathbf{p}}, \mathbf{p}), & \text{if } h_{\mathbf{p}} \neq \pi_\emptyset \\ \gamma, & \text{otherwise} \end{cases}, \quad (5)$$

where γ is a constant. Here, the role of the null hypothesis is to ignore noisy measurements or points on non-planar surfaces. As a result, we obtain a smaller number of merged plane hypotheses. The plane hypotheses in each fragment \mathbf{F}_i are transformed from the fragment coordinate system to the world reference coordinate system. Example of clustered planes are shown in Fig. 3(a).

With the clustered and transformed plane hypotheses, we find a set of dominant planes $\mathcal{P}_{\text{dominant}}$, which enables to represent the scene with a small number of plane hypotheses. To find $\mathcal{P}_{\text{dominant}}$, we employ the hierarchical agglomerative clustering again. Here, instead of comparing supervoxels, we compute the distance function of Eq. (2) using groups of 3D points with the same plane hypothesis via Eq. (4). Consequently, it is possible to acquire the set of planes that best describe the scene as shown in Fig. 3(b), where different colors indicate that different plane hypotheses are assigned.

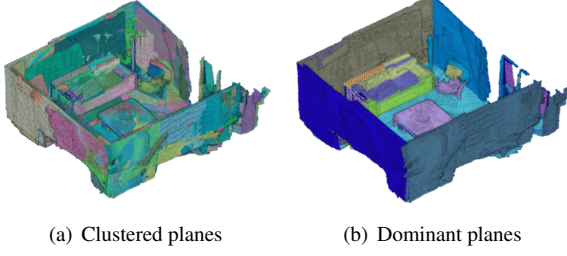


Figure 3. The result of the hierarchical agglomerative clustering and energy-based multi-model fitting. This procedure approximates the scene with the a small number of planes. Therefore, it is easy to find the scene layout from these planes.

Layout plane estimation: Given dominant planes $\mathcal{P}_{\text{dominant}}$ and clustered point clouds, we estimate the scene layout which can be understood as an envelope of an arbitrary indoor space that includes the ceiling, floor, and walls. Here, we assume a weak Manhattan world in which all the walls are orthogonal to the ceiling and the ground floor, but the walls are not necessarily orthogonal to each other. However, in practice, captured planes are hard to be perfectly planar due to the measurement noise, and therefore, we make planes orthogonal to each other if they are quite close to be orthogonal.

We find the scene layout planes $\mathcal{P}_{\text{layout}}$ through two steps. In the first step, we find the ceiling or ground floor, called a base plane, assuming that one of them is the largest planar region among all the plane hypotheses. The base plane is determined by computing the areas of dominant planes instead of simply counting the number of clustered 3D points because the density of 3D points significantly differs depending on the amount of acquired range data. To estimate the area of a dominant plane, we generate a 2D occupancy grid map on the dominant plane and project labeled 3D points on to the dominant plane. Then, we count the number of occupied grids.

In the second step, we find a set of planes that are orthogonal to the base plane determined in the first step as follows. We generate a 2D occupancy grid map on the base plane as shown in Fig. 4. Then, we project all the 3D points onto the base plane and fill each cell of the grid map if the density of points is larger than a predefined value. Here, an empty cell indicates that it is either outside the room or inside the room but not measured. From the occupancy grid, we determine the boundary of occupied grids, denoted by ∂O , via the morphological boundary detection [23] that can handle an arbitrary shape. Finally, we select the set of planes by following criteria:

$$L(\pi_i) = \begin{cases} 1, & \text{if } (\vec{n}_i \cdot \vec{n}_{\text{base}}) < \tau_1 \text{ and } g(\partial O, \pi_i) < \tau_2 \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

where \vec{n}_i and \vec{n}_{base} are normal vectors of a selected dominant plane and base plane, respectively. Therefore, the first

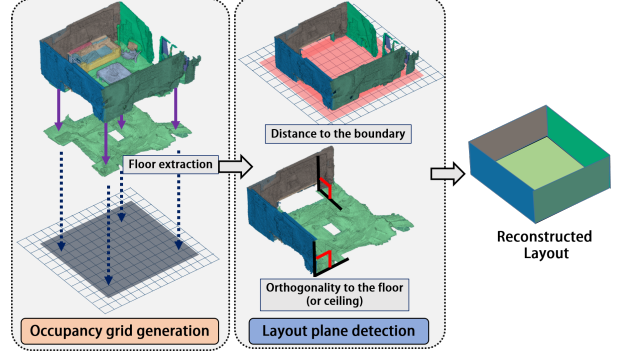


Figure 4. Layout estimation procedure. We extract a floor (or a ceiling) and generate a 2D occupancy grid. Afterwards, we find a set of layout planes by checking two criteria, boundary distance and orthogonality.

criterion checks the perpendicularity between two planes. The second criterion checks the distance between the plane π_i and the boundary ∂O because the layout planes, especially walls, surround the space. τ_1 and τ_2 are two user-defined parameters. The distance function g is defined as

$$g(\partial O, \pi_i) = \frac{1}{|\mathcal{S}_{\pi_i}|} \sum_{\mathbf{p} \in \mathcal{S}_{\pi_i}} |\partial O - \mathbf{p}_{\text{proj}}|, \quad (7)$$

where \mathbf{p}_{proj} is the projected point of \mathbf{p} on to the base plane and \mathcal{S}_{π_i} denotes a set of 3D points that belong to the plane π_i . $|\mathcal{S}_{\pi_i}|$ is the cardinality of \mathcal{S}_{π_i} . If two criteria are satisfied, we regard the corresponding dominant plane as a layout plane. Figure 4 shows the result of layout plane estimation.

3.3. Global Registration with Scene Layout

As the last step, we reconstruct the entire scene by registering all the fragments \mathcal{F} with the aid of the estimated scene layout in the world coordinate system. We pose a global optimization problem for the layout-constrained global registration. To resolve the problem, we introduce a joint approach of the layout estimation and the global registration because they depend on each other. A detailed description is given in the following subsections.

Terminologies: Let \mathcal{I} denote a set of fragment pairs. If a pair of fragments, $(\mathbf{F}_i, \mathbf{F}_j) \in \mathcal{I}$, has an overlapping region, we define a set of correspondences, $\mathcal{C}_{i,j}$, between their points in the overlapping region. Similarly, we define a set of correspondences, \mathcal{C}_i , between each fragment \mathbf{F}_i and the layout where the corresponding point of the layout is a virtual point on the layout plane. For example, we project a point \mathbf{p} of a fragment \mathbf{F}_i onto the nearest layout plane and establish a correspondence if the distance between the point and the virtual (projected) point is small. For a point $\mathbf{p} \in \mathbb{R}^3$ and a transformation \mathbf{T} , a transformed point is represented by $\mathbf{T}(\mathbf{p}) = \mathbf{R}\mathbf{p} + \mathbf{t}$ where $\mathbf{R} \in SO(3)$ is a rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is a translation vector.

Objective function: For the global registration of all fragments, we define the following energy function,

$$E_R(\mathcal{T}) = \sum_i E_{\text{layout}}(\mathbf{T}_i) + \lambda_1 \sum_{i,j} E_{\text{frag}}(\mathbf{T}_i, \mathbf{T}_j) + \lambda_2 \sum_{i,j} E_{\text{pair}}(\mathbf{T}_i, \mathbf{T}_j), \quad (8)$$

where λ_1 and λ_2 are weighting parameters and are determined depending on the numbers of points and fragment pairs. The first term is to minimize the distance between correspondence points of the layout and each fragment. Among many metrics [4, 8, 24], we use the point-to-plane metric [8]. By the metric, the first term is defined as

$$E_{\text{layout}}(\mathbf{T}_i) = \sum_{(\mathbf{p}, \mathbf{q}) \in \mathcal{C}_i} \|(\mathbf{T}_i(\mathbf{p}) - \mathbf{q})^\top \mathbf{R}_i \mathbf{n}_{\mathbf{p}}\|^2, \quad (9)$$

where $\mathbf{n}_{\mathbf{p}}$ is a normal vector of \mathbf{p} and \mathbf{q} is a virtual (projected) point on the layout. Since the layout is estimated under the weak Manhattan world assumption in Sec. 3.2, the aligned fragments have axis-aligned geometry, *e.g.*, orthogonality between a wall and the ceiling. It is worthy of note that, since we only constrain the positions of points along the envelope of the scene, objects inside the space are not necessarily planar. The second term is to minimize the distance between correspondence points of each pair of fragments. In the same way as the point-to-plane metric [8], the second term is defined as

$$E_{\text{frag}}(\mathbf{T}_i, \mathbf{T}_j) = \sum_{(\mathbf{p}, \mathbf{q}) \in \mathcal{C}_{i,j}} \|(\mathbf{T}_i(\mathbf{p}) - \mathbf{T}_j(\mathbf{q}))^\top \mathbf{R}_i \mathbf{n}_{\mathbf{p}}\|^2. \quad (10)$$

For the last term, we incorporate the pairwise transformation constraint as

$$E_{\text{pair}}(\mathbf{T}_i, \mathbf{T}_j) = \delta \left(\mathbf{T}_i \circ \tilde{\mathbf{T}}_{i,j} - \mathbf{T}_j \right), \quad (11)$$

where $\tilde{\mathbf{T}}_{i,j}$ is a pairwise transformation estimated by the iterative closest point (ICP) algorithm [8] and δ is the sum of the norms of elements. The pairwise transformation constrains the feasible solution space to avoid a degenerate situation, *e.g.*, a fragment moves too much or the scene structures are collapsed. To optimize Eq. (8), we use the widely known Gauss-Newton method.

Joint optimization: Since the layout estimation and the global registration problems are closely related to each other, we alternately estimate the scene layout and the optimal transformations instead of solving the complex joint estimation problem. Algorithm 1 shows the entire procedure. Initially, we regard that the initial transformations \mathcal{T}_0 and fragments \mathcal{F} are given. Here, we set the coordinates of the first fragment to the world reference coordinates so that \mathbf{T}_0 is fixed to an identity matrix. Afterwards, we repeatedly estimate the scene layout and minimize the objective function

Algorithm 1 Joint layout estimation and global registration

Require: $\mathcal{F}, \mathcal{T}_0$

Ensure: \mathcal{T}

- 1: establish $\mathcal{C}_{i,j} \forall (\mathbf{F}_i, \mathbf{F}_j) \in \mathcal{I}$
 - 2: $\mathcal{T} \leftarrow \mathcal{T}_0$
 - 3: **repeat**
 - 4: estimate $\mathcal{P}_{\text{layout}}$ using the method of Sec. 3.2
 - 5: establish $\mathcal{C}_i \forall \mathbf{F}_i \in \mathcal{F}$
 - 6: **repeat**
 - 7: compute $\Delta \mathcal{T}$ using Eq. (8)
 - 8: $\mathcal{T} \leftarrow \mathcal{T} + \Delta \mathcal{T}$
 - 9: **until** N times
 - 10: transform \mathcal{F} using \mathcal{T}
 - 11: **until** M times
-



Figure 5. Joint optimization procedure. As we iterate layout estimation and global registration procedures, the fragmented structures merge into a wall region.

in Eq. (8). We experimentally confirmed that the inner and outer loops in Algorithm 1 generally converge commonly within 10 and 20 iterations, respectively.

The optimized process is shown in Fig. 5. As the number of the iterations increases, the curved walls are straightened more. The reconstructed room has a cuboid shape at the end of the iterations. Consequently, the joint approach improves the global registration and the layout estimation.

4. Experimental Results

We experimentally verified the proposed method in quantitative and qualitative ways by using publicly available datasets: the augmented ICL-NUIM dataset [9] and the SUN3D dataset [46]. The former is a synthetic dataset generated in consideration of a noise model of a consumer depth camera. Since this dataset provides the ground truth trajectories and 3D structures, we performed the quantitative evaluation using this dataset. In contrary, the SUN3D dataset was captured in the real-world environment using a hand-held camera and did not provide the ground truth information. Thus, we use this dataset to confirm the feasibility of the proposed method in practical situations. For evaluation, we compare the proposed method with the state-of-the-art methods [41, 43, 46, 9]. Here, Kintuous [41] and ElasticFusion [43] are online methods, and SUN3D structure from motion (SFM) [46], the Choi *et al.* method [9], and the proposed method are offline methods. *Please note*

Table 1. Reconstruction performance evaluation in terms of average and median errors by using four synthetic datasets. The unit of error is centimeter. The best performance in each row is represented in bold.

		Kint. [41]	Elas. [43]	SUN3D [46]	Choi [9]	Ours
<i>Liv.1</i>	Avg.	13.19	9.31	12.69	5.41	2.72
	MED	7.47	4.96	5.85	4.39	1.56
<i>Liv.2</i>	Avg.	11.60	12.11	10.53	7.12	5.43
	MED	7.45	6.41	5.79	3.65	3.25
<i>Off.1</i>	Avg.	9.01	4.89	34.41	3.51	4.02
	MED	5.75	2.67	28.04	2.64	2.72
<i>Off.2</i>	Avg.	9.48	5.36	33.09	3.52	3.14
	MED	4.33	2.30	29.61	1.92	1.79

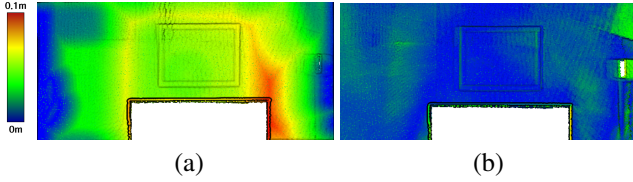


Figure 6. Visualization of reconstruction errors of the Choi *et al.* method [9] (a) and the proposed method (b) in a wall region of the *Livingroom1* dataset. The proposed method shows consistently lower errors than (a), with the aid of layout information.

that detailed parameters used for our experiments and more results can be found in the supplementary material.

Reconstruction quality: To measure the quality of estimated structures, we compute the average and the median of errors. The errors are defined as the distance between an estimated 3D point and its closest ground truth point. Table 1 shows the reconstruction errors of the proposed method and those of the state-of-the-art methods. Overall, the proposed method shows superior results compared to the state-of-the-art methods, except the *Office1* dataset. Occasionally, the reconstruction error of the proposed method is slightly higher than that of the Choi *et al.* method because of over-fitting noisily reconstructed fragments to the scene layout. However, in most cases, the proposed method shows more accurate results than the Choi *et al.* method because planar structures are preserved better than the results of the Choi *et al.* method.

It is worthy of note that the global registration without the layout information frequently shows bended walls and floors owing to noisy measurements as shown in Fig. 7. In contrast, the proposed method preserves largely planar structures with the aid of scene layout information. For example, Fig. 6(a) shows non-uniform reconstruction errors in the wall region reconstructed by the Choi *et al.* method. The proposed method shows consistently small errors over the wall in Fig. 6(b). This advantage primarily comes from the layout information. Thus, we claim that the layout information is an important cue for accurate indoor scene reconstruction.

Trajectory accuracy: The trajectory error is measured in

Table 2. Comparison of trajectory errors in terms of root mean squared errors (RMSE) and median errors for synthetic datasets. The errors are measured in centimeter. The best performance in each row is represented in bold.

		Kint. [41]	Elas. [43]	SUN3D [46]	Choi [9]	Ours
<i>Liv.1</i>	RMSE	57.36	59.02	32.22	9.87	9.49
	MED	45.16	43.83	27.28	7.88	8.18
<i>Liv.2</i>	RMSE	29.32	37.09	29.13	13.63	12.18
	MED	27.16	24.67	24.15	11.81	10.50
<i>Off.1</i>	RMSE	18.29	13.10	50.84	6.22	9.95
	MED	12.11	9.69	42.68	5.39	9.31
<i>Off.2</i>	RMSE	27.18	13.26	29.75	8.89	6.93
	MED	25.25	11.89	28.40	9.02	5.86

terms of the root mean squared error (RMSE) and median error between the ground-truth trajectory and an estimated camera trajectory. Since an accurate camera trajectory implies the accurate registration of scene fragments, we use this metric to evaluate various indoor reconstruction methods. Table 2 shows the trajectory errors of the proposed method and the state of the art methods. The proposed method outperforms other methods except the *Office1* dataset similarly as in Table 1. This quantitative comparison also confirms that the proposed method is promising, especially when previous approaches cannot preserve the global scene structures well.

Qualitative evaluation: We compare reconstructed results of some selected methods [41, 46, 9] in challenging real-world datasets provided by Xiao *et al.* [46]. As shown in Fig. 7, Kintinuous, SUN3D SFM, and the Choi *et al.* method cannot preserve the genuine structure of walls in the real-world datasets. These results of the methods except our method show curved structures as well as largely distorted walls. Moreover, the SUN3D SFM shows noisy 3D points along the wall region. However, in the presence of a large amount of errors, the proposed method shows significantly improved results as shown in Fig. 7(e) and 7(j). The effect of the layout-constrained registration can be found clearly in real-world datasets.

For the qualitative evaluation using a weak Manhattan world scene, the results of *hotel_stb_scan3* dataset are used for comparison with other methods. As shown in Fig. 8(a) and 8(b), the Choi *et al.* method and ours yield good reconstruction results in the weak Manhattan world scene. In addition, we performed the experiments only using the first 1,000 frames out of 3,756 frames of the same dataset so that loop closures were not detected. Here, it is observed that the result reconstructed by the Choi *et al.* method (Fig. 8(c)) is slightly bent, whereas the result by our method (Fig. 8(d)) is similar to the result by applying loop closing (Fig. 8(b)), with the aid of layout information. Similarly, loop closures are not detected properly in Fig. 7(d) and 7(i), and therefore, reconstruction results are poor. Nevertheless, the proposed method recovers rectangular shapes of the scene in Fig. 7(e) and 7(j). Slight quality degradation can occasion-

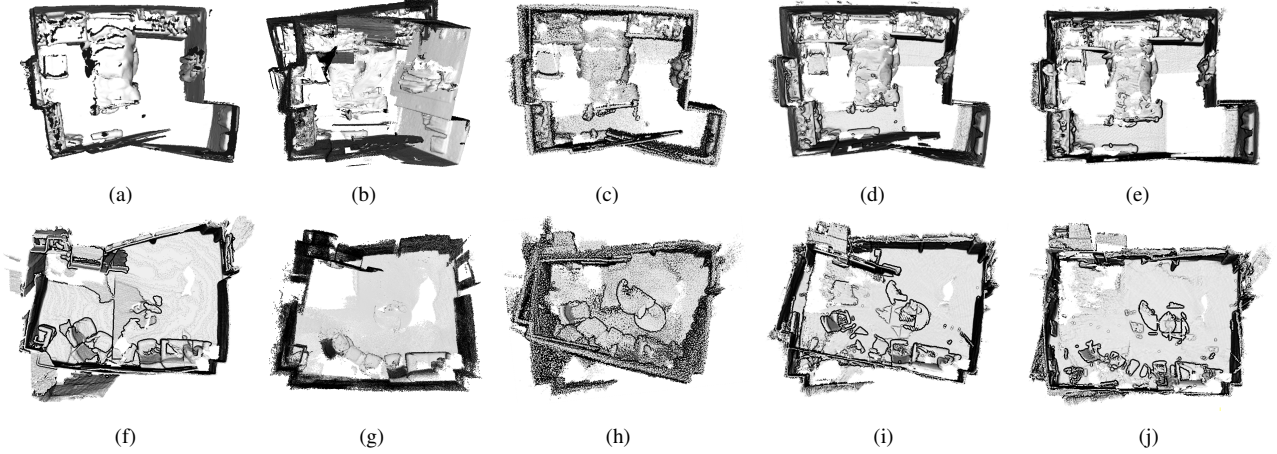


Figure 7. Comparison of reconstructed results for real-world datasets [46]. From the left to the right; Kintinuous [41], ElasticFusion [43], SUN3D SFM [46], Choi *et al.* [9], and the proposed method, respectively. From the top to the bottom; *mit_dorm_next_sj* and *mit_lab_hj* datasets, respectively. Reconstructed results are compared using the top-view to clearly show registration errors in wall regions.

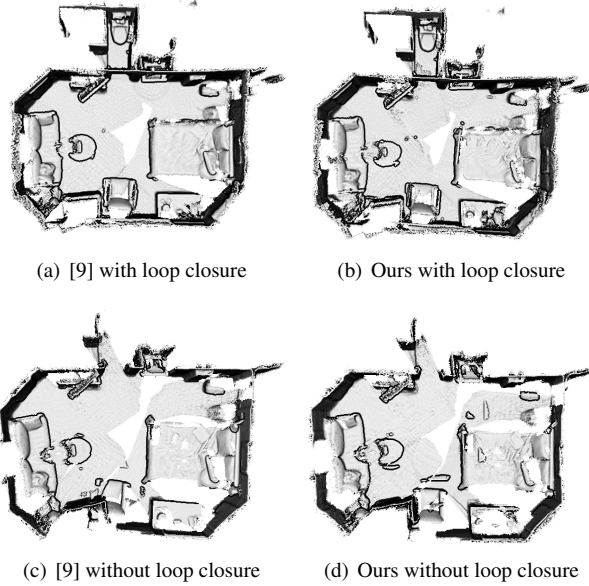


Figure 8. Reconstruction results in the weak Manhattan world scene.

ally happen if the reconstructed model is overfitted, *e.g.*, as in *Office1*. However, the proposed method improves the reliability of reconstruction while retaining accurate reconstruction results.

Computational complexity: The proposed method was implemented mixedly in MATLAB and C++ and ran on 2.6 GHz CPU with single core. For the *Livingroom1* dataset, the total computational time of our method was about 3 hours, which was two times of that of the Choi *et al.* method implemented in C++. The largest burden in the process is the dominant plane extraction step. The process of the dominant plane extraction has a complexity of $O(nml)$

since it computes the distance between n plane hypotheses and m points in l fragments. Although the proposed method is an offline method as [46, 9] and is slower than the state-of-the-art methods, we explicitly assure that our method offers more reliable, robust, and accurate indoor 3D reconstruction results in comparison with the real-time methods [41, 43] and other state-of-the-art offline methods [46, 9].

5. Conclusion

We have presented an indoor 3D reconstruction algorithm that alternately resolves two complementary problems, scene layout estimation and global registration, in iterative fashion. Given initially registered scene fragments, we estimate the envelope of a scene through hierarchical clustering and energy-based multi-model fitting and find a minimum set of planes that best describe the entire scene. From these plane hypotheses, we extract the scene layout that surrounds the entire point cloud, assuming that they coincide with walls, a floor, and a ceiling. We exploit the scene layout information to obtain globally consistent reconstruction results by constraining the global registration problem with scene layout information. We verified the superiority of the proposed method by using various datasets, including a challenging real-world dataset.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant (No. NRF-2015R1A2A1A01005455) and ‘The Cross-Ministry Giga KOREA Project’ grant (GK17P0300, Real-time 4D reconstruction of dynamic objects for ultra-realistic service) funded by the Korea government(MSIT).

References

- [1] F. Arrigoni, B. Rossi, and A. Fusiello. Global registration of 3d point sets via lrs decomposition. In *ECCV*, 2016.
- [2] E. Ataer-Cansizoglu, Y. Taguchi, S. Ramalingam, and T. Garaas. Tracking an rgb-d camera using points and planes. In *ICCV Workshops*, 2013.
- [3] R. Bergevin, M. Soucy, H. Gagnon, and D. Laurendeau. Towards a general multi-view registration technique. *TPAMI*, 18(5):540–547, 1996.
- [4] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *TPAMI*, 14(2):239–256, Feb. 1992.
- [5] D. Borrmann, J. Elseberg, K. Lingemann, A. Nchter, and J. Hertzberg. Globally consistent 3d mapping with scan matching. *Robotics and Autonomous Systems*, 56(2):130–142, 2008.
- [6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 23(11):1222–1239, 2001.
- [7] J. Chen, D. Bautembach, and S. Izadi. Scalable real-time volumetric surface reconstruction. *ACM Trans. Graph. (TOG)*, 32(4), July 2013.
- [8] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.
- [9] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *CVPR*, 2015.
- [10] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *CVPR*, 2013.
- [11] R. Cupec, E. K. Nyarko, D. Filko, A. Kitanov, and I. Petrović. Place recognition based on matching of planar surfaces and line segments. *IJRR*, 34(4-5):674–704, 2015.
- [12] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996.
- [13] M. Dou, L. Guan, J.-M. Frahm, and H. Fuchs. Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held rgb-d camera. In *ACCV*, 2012.
- [14] E. Fernández-Moral, W. Mayol-Cuevas, V. Arévalo, and J. González-Jiménez. Fast place recognition with plane-based maps. In *ICRA*, 2013.
- [15] A. Furlan, D. Miller, D. G. Sorrenti, L. Fei-Fei, and S. Savarese. Free your camera: 3d indoor scene understanding from arbitrary camera motion. In *BMVC*, 2013.
- [16] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, 2009.
- [17] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011.
- [18] N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann. Robust global registration. In *Symposium on geometry processing*, volume 2, 2005.
- [19] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010.
- [20] P. Henry, D. Fox, A. Bhowmik, and R. Mongia. Patch volumes: Segmentation-based consistent mapping with rgb-d cameras. In *International Conference on 3D Vision (3DV)*, 2013.
- [21] F. Heredia and R. Favier. Kinectfusion extensions to large scale environments. <http://www.pointclouds.org/blog/srcs/fheredia/index.php>, 2012. Online; accessed 11-May-2012.
- [22] P. J. Huber. *Robust statistics*. Springer Berlin Heidelberg, 2011.
- [23] M. Iwanowski. Morphological boundary pixel classification. In *International Conference on "Computer as a Tool"*, 2007.
- [24] Y. Liu, W. Zhou, Z. Yang, J. Deng, and L. Liu. Globally consistent rigid registration. *Graphical Models*, 76(5):542–553, 2014.
- [25] L. Ma, C. Kerl, J. Stueckler, and D. Cremers. Cpa-slam: Consistent plane-model alignment for direct rgb-d slam. In *ICRA*, 2016.
- [26] L. Magri and A. Fusiello. T-linkage: A continuous relaxation of j-linkage for multi-model fitting. In *CVPR*, June 2014.
- [27] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.
- [28] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Trans. Graph. (TOG)*, 32(6):169, 2013.
- [29] K. Nishino and K. Ikeuchi. Robust simultaneous registration of multiple range images. In *ACCV*, 2002.
- [30] S. Oesau, F. Lafarge, and P. Alliez. Planar shape detection and regularization in tandem. In *Computer Graphics Forum*, volume 35, pages 203–215, 2016.
- [31] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *CVPR*, 2013.
- [32] K. Pulli. Multiview registration for large data sets. In *Second International Conference on 3-D Digital Imaging and Modeling*, 1999.
- [33] H. Roth and M. Vona. Moving volume kinectfusion. In *BMVC*, 2012.
- [34] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *ICRA*, 2009.
- [35] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *ICRA*, 2011.
- [36] R. F. Salas-Moreno, B. Glocker, P. H. J. Kelly, and A. J. Davison. Dense planar slam. In *ISMAR*, 2014.
- [37] T. Shiratori, J. Berclaz, M. Harville, C. Shah, T. Li, Y. Matsushita, and S. Shiller. Efficient large-scale point cloud registration using loop closures. In *International Conference on 3D Vision (3DV)*, 2015.
- [38] F. Steinbrucker, C. Kerl, and D. Cremers. Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *ICCV*, 2013.
- [39] Y. Taguchi, Y.-D. Jian, S. Ramalingam, and C. Feng. Point-plane slam for hand-held 3d sensors. In *ICRA*, 2013.
- [40] Y. Verdie, F. Lafarge, and P. Alliez. Lod generation for urban scenes. *ACM TOG*, 34(3):30, 2015.
- [41] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended Kinect-Fusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

- [42] T. Whelan, M. Kaess, J. J. Leonard, and J. McDonald. Deformation-based loop closure for large scale dense rgb-d slam. In *IROS*, 2013.
- [43] T. Whelan, S. Leutenegger, R. S. Moreno, B. Glocker, and A. Davison. Elasticfusion: Dense slam without a pose graph. In *RSS*, 2015.
- [44] J. A. Williams and M. Bennamoun. Simultaneous registration of multiple point sets using orthonormal matrices. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000.
- [45] J. Xiao and Y. Furukawa. Reconstructing the world’s museums. In *ECCV*, 2012.
- [46] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013.
- [47] J. F. Y. Tang. Hierarchical multiview rigid registration. *Computer Graphics Forum*, 34(5):77–87, 2015.
- [48] Y. Zhang, W. Xu, Y. Tong, and K. Zhou. Online structure analysis for real-time indoor scene reconstruction. *ACM Trans. Graph. (TOG)*, 34(5):159:1–159:13, Nov. 2015.
- [49] Q.-Y. Zhou and V. Koltun. Simultaneous localization and calibration: Self-calibration of consumer depth cameras. In *CVPR*, 2014.
- [50] Q.-Y. Zhou and U. Neumann. 2.5d building modeling by discovering global regularities. In *CVPR*, pages 326–333, 2012.