3D Indoor Scene Completion via Room Layout Estimation

Wei Xia*, Rongfeng Lu*, Yaoqi Sun, Chenghao Xu, Kun Lv, Yanwei Jia, Zunjie Zhu and Bolun Zheng

School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang, China
Email: wx960519@gamil.com. “*” shares equal contribution to this work.

Abstract. Recent advances in 3D reconstructions have shown impressive progress in 3D indoor scene reconstruction, enabling automatic scene modeling; however, holes in the 3D scans hinder the further usage of the reconstructed models. Thus, we propose the task of layout-based hole filling for the incomplete indoor scene scans: from the mesh of a scene model, we estimate the scene layout by detecting the principal planes of a scene and leverage the layout as the prior for the accurate completion of planar regions. Experiments show that guiding scene model completion through the scene layout prior significantly outperforms the alternative approach to the task of scene model completion.

Keywords. Shape completion; layout estimation; indoor scene.

1. Introduction
Recently, we have seen remarkable advances in real-time 3D indoor scene reconstruction, driven by the breakthrough of hardware (such as range sensors and GPUs) as well as the progress of technical research [1, 2]. However, due to the incomplete scan from users, the reconstructed scene models are always with many holes, which hinders the understanding for the reconstructed models and reduces the interactivity of the human-computer interaction (HCI). For instance, methods such as Scene CAD [3] use wall corners for the scene layout estimate, however, they cannot detect the corners when the corner regions are not scanned, leading to the failure of those methods.

The floor, wall and ceiling are three principal components to construct an indoor scene model. Hence, this paper defines the indoor layout estimation of the scene model as the extraction of the main plane in different scenes. Nowadays, the plane detection problem can be roughly divided into three categories: the method based on RANSAC [4], the method based on Hough transform and the method based on region growth.

Lots of works have been developed to infer indoor structure layout information from 2D/3D input. Scan2BIM [5] takes 3D point cloud model as input to generate building information models under Manhattan-hypothesis constraints. They detect planes and figure out the intersection relationship between different planes to generate loops to construct a room layout.

The existing completion methods are divided into two kinds, one is based on a 3D point cloud, such as PCN algorithm [6] can complete the missing point cloud. The other is mesh-based completion algorithm, which can scan the 3D model with holes and fill the original surface gap area with new triangular patches. However, most of the existing completion algorithms work on smaller objects [7, 8], such as the Stanford-
bunny model. At the same time, the completed region is required a full envelope closed-loop region, for more complex models they may be impossible to repair objects with excellent result.

Considering the difficulty of the end-to-end scene model completion, in this work, we mainly focus on the completion of principal planes of a scene. To recover the realistic scene model, we take the full usage of the scene layout as the prior for the completion. As for the layout estimation of a scene model, we detect principal planes from the mesh model by implementing the triangle mesh clustering via planar properties at first. Then we estimate all corner points of a scene according to the intersection between different planes to obtain the final layout. A hole among a wall corner in the layout will be redefined as three holes belong to two walls and a floor. Then, similar to [8], we develop a biharmonic field to calculate the isosurface of each hole for the completion. Besides, we also recover holes with a sharp feature on the basis of planar information, which cannot be done by the method of [8], since there would be an interruption among neighboring objects so that they could not figure out the sharp area. Our proposed approach enabling the separation of objects and planes for the hole filling. The pipeline of this paper is shown in figure 1.

![Figure 1](image)

**Figure 1.** The red part is layout estimation process that take an incomplete mesh as input (left) to do supervoxel segmentation and cluster at first to reduce the cost of calculation. After that, all planes are detected on the basis of segmented supervoxel. Meanwhile, principle planes are filtered used to estimate scene layout. In the blue section, layout prior is offered as supplements to support hole detection process. Later, isosurface of each hole is calculated and holes are repaired based on that.

2. Indoor Scene Layout Estimation

2.1. Principal Plane Detection

Given an indoor scene model contains millions of triangular patches generally that immensely increase the computation cost. Hence we adopt an over-segmentation method [9] to shrink the volume of mesh number cost, meanwhile it could guarantee local feature constantly. Firstly, a group of clustering hyperpatches \( S = \{S_i\}_{i=1}^{N} \) are generated on the mesh model. Each cluster hyperpatch has centroid \( c_i \), normal vector \( n_i \) and curvature \( u_i \). We use RANSAC sequentially to extract pre-selected planes on clustered-mesh as set: \( Y = \{p_1, p_2, \ldots, p_l\} \). The process of selecting building structures from pre-selected planes is defined as a task of minimizing the energy function with the constraint of geometry:

\[
P^* = \arg\min P E(P)
\]  

(1)

The \( E(P) \) is the energy function extracted from the main structure of indoor buildings. For this method, energy function is defined as:
\[ E(P) = \sum_{p \in P} Z(p) + \sum_{p_m, p_n \in P} H(p_m, p_n) + \sum_{p_m, p_n \in P} O(p_m, p_n) \]  \hspace{1cm} (2)

where \( Z(p_m) \) constrains the authenticity of planes. There will be many hyperfaces on a plane if it is a real plane in the dense model. If the number of hyperfaces on the plane is less than a threshold, the plane will not be considered as a building structure. On the contrary, we consider a plane as an indoor building structure if it contains a great number of hyperfaces. \( H(p_m, p_n) \) illustrates the geometric constraints of the main structure inside buildings. The building structures are parallel or perpendicular to each other in majority circumstances, for instance, walls are perpendicular to floor and the ceiling, and the floor is parallel to the ceiling. Therefore, we use \( H(p_m, p_n) \) to identify whether two planes meet the parallel or vertical relationship, and then filter the main structure of the building. In addition, in order to avoid repeated extraction of overlapping planes, an overlapping constraint \( O(p_m, p_n) \) is set. In equation (2), the \( Z(p_m) \) is defined as follow:

\[
Z(p_m) = \begin{cases} 
1 & |\xi(p_m)| < \lambda \\
\exp\left(\frac{\lambda - |\xi(p_m)|}{\lambda}\right) & |\xi(p_m)| \geq \lambda 
\end{cases} \hspace{1cm} (3)
\]

The \( \xi(p_m) \) represents a set of points in which the distance between point \( c_i \) and plane \( p_i \) is less than a threshold. And it could be described as \( \xi(p_m) = \{x \in \mathbb{R} | r(c_i, p_m) < \theta \} \), \( r(c_i, p_m) \) is the distance between \( c_i \) and \( p_m, \theta \) is the noise threshold of a plane. In equation (3), \( \lambda \) represents the minimum number of hyperfaces that are contained in a building structure. The building structures are parallel or perpendicular to each other in majority circumstances, for instance, walls are perpendicular to floor and the ceiling, and the floor is parallel to the ceiling. Therefore, we use \( H(p_m, p_n) \) to identify whether two planes meet the parallel or vertical relationship, and then filter the main structure of the building. In addition, in order to avoid repeated extraction of overlapping planes, an overlapping constraint \( O(p_m, p_n) \) is set. In equation (2), \( Z(p_m) \) is defined as follow:

\[
H(p_m, p_n) = \frac{\exp(A(p_m, p_n) - A(p_m, p_n))}{\exp(A(p_m, p_n)) + \exp(-A(p_m, p_n))} \hspace{1cm} (4)
\]

where \( A(p_m, p_n) = |90 - A(p_m, p_n), A(p_m, p_n)| \), \( A(p_m, p_n) \) indicates the angle between plane \( p_m \) and plane \( p_n \). The design of energy function in equation (4) is motivated by an activation function \( \tanh \). In equation (2), \( O(p_m, p_n) \) is defined as following:

\[
O(p_m, p_n) = \frac{|\xi(p_m) \cap \xi(p_n)|}{\min(|\xi(p_m), \xi(p_n)|)} \hspace{1cm} (5)
\]

Based on equation (2)(3)(4)(5), we construct an energy function to extract building structure layout. Solving minimum energy function (1) can be seen as a NP-hard problem. We convert minimizing energy function (2) to minimizing binary function (6):

\[
E(X) = \sum_{m=1}^{M} Z(p_m)x_m + \sum_{p_m, p_n \in P} H(p_m, p_n)x_m x_n + \sum_{p_m, p_n \in P} O(p_m, p_n)x_m x_n \hspace{1cm} (6)
\]

where \( X = [x_1, x_2, \ldots, x_M] \) is the vector if binary variable, when \( x_m = 1 \) it represents selected plane \( p_m \). Energy function is non-submodular [10] so that the minimizing task could still be treated as a NP-hard problem. We adopt a local sub-modularization method [11] to optimize the equation (6). Then perform a point-to-plane assignment, and mark the points that belong to the same plane into the same label. At this time, the main structure of the building in the scene is successfully proposed. In this over-segmentation method, large amount of data will be compressed into limited number of supervoxels which remarkably decrease the size of input, meanwhile it also can keep local features trackable.

2.2. Layout Detail Estimation

We calculate intersection points between floor and walls since plane parameters are extracted in section 3.1. We assume floor is orthogonal to all walls so that an intersection point between two adjacent walls and floor can be seen as corner point in a scene model. We state the problem into the calculation of intersection point between three orthogonal planes [12]. Gauss elimination method is used in this issue.
3. Hole Filling

3.1. Completion Strategy

Bi-harmonic filed is proposed in [8] to tackle with hole completion task at first, which can recover signal mesh object properly. However, this method cannot perform efficiently in a complex indoor scene because neighboring holes from different objects will intersect each other during repairing. Meanwhile, this method could not resume sharp geometry feature accurately like corner or wall edge.

We propose a novel approach to remedy defects of the bi-harmonic field algorithm in indoor scene completion task. We extract the principle plane parameter at first from inputting indoor model via plane detection algorithm (see section 3) which is constrained with building structure. Then, we predict the scene layout on the basis of model architecture as prior so that we can separate clutters (chairs, tables, or other furniture) and principle planes which minimizes the interfection dramatically between different objects during repairing. Separating them allows us to fill holes in layout more effectively (walls and floor) and clutters. Experiment results in section 5.3 demonstrate our method fixes hole more efficiently compare to bi-harmonic field approach. Our method takes scene layout as prior to literally specify the target area and remarkably optimize the completion effect about areas with sharp geometry feature.

3.2. Hole Analysis

In figure 2, we categorize the holes among planes of a 3D scan into four different types.

![Figure 2](image)

**Figure 2.** Different types of holes: (a) All points at the hole boundary belong to the same plane; (b) All points at the hole boundary belong to two different places; (c) All points at the hole boundary belong to three different planes; (d) All points at the hole boundary belong to four or more different planes.

3.3. Boundary and Biharmonic Field

We generate a 3D bounding box around each hole after the detection process that is created via expanding 20% outward in each coordinate direction. Finally, we use an approach on the basis of voxel to recover holes inside all bounding boxes.

First, inside each 3D bounding box $B_h$ we generate a grid $G_h$ with the same resolution that could explicitly guarantee each triangle mesh would not exceed the size of $B_h$. The resolution setting will keep updating depends on the resolution of the voxel that is mentioned in section 3.1. Secondly, we calculate the signed distance from $B_h$ to mesh and the unsigned distance between $B_h$ and the hole boundary. We divide $G_h$ into two different areas $R_d$ and $R_b$ in $B_h$ derived from distance values, $R_d$ area could calculate unsigned distance robustly due to close to mesh. $R_b$ area is neighboring to hole area, we extend smoothly the value from $R_b$ to $R_d$ with the assistance of $R_d$ area. Then we construct a truncated signed distance function (TSDF) on the basis of the signed distance to mesh in $R_d$ area, then we obtain a discrete bi-harmonic filed via figuring out constrained optimization problem in $R_b$ area, furthermore, we get zero isosurface in $R_b$ area. The computation process could be accelerated by a multi-grid solver, and we use unsigned distance filed $R_d$ as boundary constrained condition during bi-harmonic calculation in the
solving process so that the zero isosurface can be evaluated more robust in $R_b$. Once bi-harmonic field is calculated in $R_b$ area, we complete the filling task by stitching zero isosurface and mesh around hole boundary. Further detail about the biharmonic calculation and unsigned or signed distance field evaluation could check [8].

Inside an indoor scene, building architecture and spatial distance are used to segregate objects via the above algorithm. Secondly, holes can be detected separately on different objects. Thus, we determine to take a 2-ring triangle mesh as the beginning edge and use the proposed algorithm to do hole detection and hole filling tasks.

4. Experiment And Result

4.1. Dataset
The experimental dataset we adopt from the indoor scenes of ScanNet [13]. However, we found that the dataset from ScanNet always accompany with holes that is not proper enough to be referred as ground truth so that we took SceneNet as well. SceneNet [14] could offer a full envelope 3D mesh scene so that we could evaluate our algorithm effect intuitively.

4.2. Layout estimation
We take an indoor mesh with holes as input to reproduce the plane detection result on the basis of [9] shows in figure 3. It visualizes point cloud plane detection result in figure 3b. Moreover, we only extract principle components from original input as shown in figure 3c, after that layout can be estimated according to the method is mentioned in section 3.2.

![Figure 3](image)

(a) Input model       (b) Planes detection     (c) Principle planes detection     (d) Layout estimation

**Figure 3.** The plane detection approach from [9] takes (a) as input and output point cloud in (b). (c) the detection result from ours in mesh representation; (d) the final layout estimation result.

4.3. Filling Result
We first reproduce the experiment result of [7] in ScanNet dataset. Then we compare it with our result against to separate clutter and planes in figure 4 and figure 5. Only a few models in the sceneNet is full envelope so that we choose synthetic data from sceneNet and compare two different methodologies in dealing with different type of holes. It can be clearly found that the prior information prompts the completion effect significantly and scale the extendable area for the ceiling hole narrowly. The following experiments take incomplete model as input shows the repairing effect between our approach and Oscars algorithm.
Figure 4. (a) is input model, (b) is the result from [7] that could be easily found that the repairing effect of plane and chairs are affected by each other, the ground floor is connected to chairs after repairing and the ground does not extend horizontally or bend severely. On the other hand, clutters and planes are separated into two parts, then repairing them independently, the ground extends in a limited square region, the chair would not be affected by the hole lies on the ground.

Figure 5. An experiment takes incomplete model as input shows the repairing effect between our approach and Oscars algorithm. In order to verify our algorithm could be applied in border-less model that we do an experiment in a border-less scene. The experiment demonstrates that our approach could manipulate principle plane extension in a limited region (figure 6). On the contrary, Oscar’s method performs terrible on repairing due to infinite boundary. Meanwhile, some clutters cross through the ground floor which destroys the hole detection and isosurface calculation process so that the ground floor and clutters are erased after the completion process.

Figure 6. An experiment demonstrates the completion result from ours and Oscars.
5. Limitations
One of the limitations of our approach is that we only focus on the holes in planes, which means that we cannot accurately fill holes among objects in indoor scenes. Besides, the proposed layout estimation only works on single room due to our heuristic strategy of the principal plane detection. On the other hand, after the repairing process is done the whole topology structure will be amended into a brand-new form so that it will be challenging to do a quantitative evaluation with other approaches.

6. Conclusion
In this paper, we have presented an approach for the hole filling task, which generates the complete model for the reconstructed indoor scene with the layout auxiliary. The completion process is divided into two sub-tasks: the layout estimation as well as the hole identification and repairing. For a 3D scan of an indoor scene, we leverage the supervoxel segmentation to guide the detection of principal planes, which is then used to construct the layout of the scene. In the second sub-task, we identify the types of different holes based on the estimated layout, and simplify complicated ones into two or easier holes to keep the planar assumption for each hole. Finally, the isosurfaces are derived for the filling of all holes, generating the complete mesh model for the indoor scene. We show that our layout guided hole filling outperforms the alternative approach on the task of indoor scene model completion. In the future, we would like to explore 3D geometry learning methods for scene completion tasks.

Acknowledgements
This work was supported by the National Key Research and Development Program of China under Grant (2020YFB1406604), National Nature Science Foundation of China (61931008, 61671196, 62071415, 62001146, 61701149, 61801157, 61971268, 61901145, 61901150, 61972123), Zhejiang Province Nature Science Foundation of China (LR17F030006, Q19F010030), 111 Project, No. D17019.

References
[1] Dai A, Nießner M, Zollhöfer M, Izadi S and Theobalt C 2017 Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration ACM Transactions on Graphics (ToG) 36 (4) 1.
[2] Whelan T, Salas-Moreno R F, Glocker B, Davison A J and Leutenegger S 2016 ElasticFusion: Real-time dense SLAM and light source estimation The International Journal of Robotics Research 35 (14) 1697-1716.
[3] Avetisyan A, Khanova T, Choy C, Dash D, Dai A and Nießner M 2020 Scenecad: Predicting object alignments and layouts in RGB-D scans Computer Vision–ECCV 2020: 16th European Conference (Glasgow, UK, August 23-28, 2020) Part XXII 16 pp 596-612.
[4] Schnabel R, Wahl R and Klein R 2007 Efficient RANSAC for point-cloud shape detection Computer Graphics Forum (Oxford, UK: Blackwell Publishing Ltd.) pp 214-226.
[5] Murali S, Speciale P, Oswald M R and Pollefeys M 2017 Indoor Scan2BIM: Building information models of house interiors 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp 6126-6133.
[6] Yuan W, Khot T, Held D, Mertz C and Hebert M 2018 PCN: Point completion network 2018 International Conference on 3D Vision (3DV) (IEEE) pp 728-737.
[7] Guo X, Xiao J and Wang Y 2018 A survey on algorithms of hole filling in 3D surface reconstruction The Visual Computer 34 (1) 93-103.
[8] Argudo O, Brunet P, Chica A and Vinacua Â 2015 Biharmonic fields and mesh completion Graphical Models 82 137-148.
[9] Papon J, Abramov A, Schoeler M and Worgotter F 2013 Voxel cloud connectivity segmentation-supervoxels for point clouds Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 2027-2034.

[10] Kolmogorov V and Rother C 2007 Minimizing nonsubmodular functions with graph cuts-A review IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (7) 1274-1279.

[11] Gorelick L, Boykov Y, Veksler O, Ben Ayed I and Delong A 2014 Submodularization for binary pairwise energies Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 1154-1161.

[12] Goldman R 1990 Intersection of three planes Graphics Gems p 305.

[13] Dai A, Chang A X, Savva M, Halber M, Funkhouser T and Nießner M 2017 ScanNet: Richly-annotated 3D reconstructions of indoor scenes Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5828-5839.

[14] Handa A, Patraucean V, Badrinarayanan V, Stent S and Cipolla R 2015 SceneNet: understanding real world indoor scenes with synthetic data arXiv preprint arXiv:1511.07041.