Floorplan-Jigsaw: Jointly Estimating Scene Layout and Aligning Partial Scans

Cheng Lin   Changjian Li   Wenping Wang

The University of Hong Kong

{clin, cjli, wenping}@cs.hku.hk

Abstract

We present a novel approach to align partial 3D reconstructions which may not have substantial overlap. Using floorplan priors, our method jointly predicts a room layout and estimates the transformations from a set of partial 3D data. Unlike the existing methods relying on feature descriptors to establish correspondences, we exploit the 3D "box" structure of a typical room layout that meets the Manhattan World property. We first estimate a local layout for each partial scan separately and then combine these local layouts to form a globally aligned layout with loop closure. Without the requirement of feature matching, the proposed method enables some novel applications ranging from large or featureless scene reconstruction and modeling from sparse input. We validate our method quantitatively and qualitatively on real and synthetic scenes of various sizes and complexities. The evaluations and comparisons show superior effectiveness and accuracy of our method.

1. Introduction

Indoor scene understanding and reconstruction have been extensively researched in computer vision. In recent years, the development of consumer RGB-D sensors has greatly facilitated 3D data capture and enabled high-quality reconstruction of indoor scenes. Although many methods have been proposed for continuous camera localization to register 3D depth data, it remains a challenge to scan some scenes in a single pass. The main difficulty is caused by interruptions in camera tracking, which results in a number of partial scans with little overlap. This frequently occurs in the following typical scenarios: (1) a large-scale scene is scanned region-by-region rather than in a single pass to reduce the workload or to meet the memory limit of a computer; (2) when scanning featureless areas or doorways, camera tracking often fails and so leads to several partial scans without sufficient overlap or feature points; (3) when a large scene is scanned using multiple robots, the scene is usually explored by different agents in disjoint sub-regions

Figure 1: We present a method to jointly align a set of unordered partial reconstructions and estimate a room layout.

which have little overlap [38], leading to a set of partial scans. The alignment of such unordered partial 3D data is an under-explored problem and it is challenging to the existing methods because of their requirements on the large overlap and dense feature points for scan registration.

In this paper, we propose a method for registering partial reconstructions of an indoor scene which may not have sufficient overlap, as shown in Fig. 1. Our key observation is that the local layouts of partial reconstructions can be viewed as the fragments of a global room layout which typically has the following two characteristics: (1) the room layout is a set of perpendicular or parallel walls, which is referred to the Manhattan World (MW) property; (2) the room layout forms a closed loop on a 2D floorplan. We exploit these properties to develop an efficient method for jointly predicting a room layout that has the above layout properties and estimating the transformations from a set of unordered partial reconstructions.

Most of the existing methods [2, 17] use boundary loop detection to estimate a room layout because their input is a long sequence of scans that have substantial overlap and complete coverage of the indoor scene. In contrast, the input to our method can be partially scanned data without clear boundaries. By taking noises and occlusions into consideration, our method is capable of reconstructing scenes with incomplete, disconnected or even occluded walls. Given such a set of partial scans with detected lay-

outs, we analyze the relationship between each local layout with the global layout to achieve successful alignment, while the existing methods would fail due to the lack of sufficient overlap and features for establishing correspondences. We formulate a novel optimal placement problem to determine the rotation and translation of each partial scan using the MW assumption and the layout properties, and then produce the final transformations to align the scans and predict a complete global room layout. The framework of our method is illustrated in Fig. 2.

Without relying on feature matching, our method not only works robustly when the partial reconstructions do not have substantial overlap, but also enables a series of novel applications, e.g., the reconstruction of featureless or large scenes, modeling from sparse input, RGB-D stream downsampling, to name a few (Sec. 5).

We validate our approach qualitatively and quantitatively on both real and synthetic scenes of various sizes and complexities, and compare it with the state-of-the-art methods. The evaluations and comparisons demonstrate that, given a set of partial reconstructions, our method is able to compute the accurate transformations to align them and reconstruct a high-quality scene layout by effectively estimating and combining local layouts of partial data.

2. Related Work

Indoor scene understanding has been a popular topic and accumulated rich literature in the past decades. We review the most relevant works and refer readers to the survey [25] to have an overview.

3D data registration. In the last decade, a number of simultaneous localization and mapping (SLAM) techniques are extensively employed to model 3D scenes using RGB-D sensors. Some typical works include Kinect Fusion [26], Elastic Fusion [36], ORB-SLAM [22] and so on. To establish robust correspondences between 3D data, a wide range of geometric feature descriptors [27, 42, 11] are proposed. Also, global registration approaches [41, 45] are developed to alleviate the local optimum issue when aligning point sets. These methods are effective for feature matching, surface alignment as well as 3D reconstruction. However, when it comes to the 3D data without sufficient overlap and correspondences, these algorithms are likely to fail or exhibit unacceptable inaccuracies (see Fig. 11 and Fig. 15).

Room layout estimation. Methods for room layout estimation can be roughly divided into three categories based on their inputs, i.e., single view RGB/RGB-D image, panoramic RGB/RGB-D image, and dense point cloud.

Many works focusing on layout estimation from a single image [16, 30, 8, 3, 29] have been continuously developed. Due to the limitation of the narrow field-of-view caused by a single standard image, researchers have tried to exploit

panoramic images [44, 2, 40] to recover the whole room context. With the success of deep learning in vision tasks, newest techniques [15, 46] rely on convolutional neural networks to map an RGB image to a room layout directly. These methods using standard or panoramic RGB images are highly dependent on feature points either for key structure detection or for pose estimation. Because of the instability of image feature points, these methods will suffer from inaccuracy as well as the incapability of handling complex (they usually recover "cuboid" or "L" shape [15]) and featureless scenes. Instead, our method uses depth data and is independent of feature points to avoid these drawbacks.

RGB-D images include 3D range information of each pixel, thus significantly improving the accuracy and the robustness of geometry reasoning. Some methods use a single RGB-D image [35, 43] to estimate room layout, which is also limited by the narrow field-of-view. With the superiority of panoramic RGB-D images, higher-quality layout analysis and structured modeling results have been achieved [10, 37]. There are also a few methods using densely scanned point clouds as input to estimate scene layouts [23, 17, 19]. Most of these methods target a complete scene in order to exploit the closed boundary nature of room layout, while our method is able to cope with the more challenging partial scans which lack clear outer boundaries.

Indoor scene constraints. Intrinsic properties of indoor scenes are widely used in indoor understanding and reconstruction. Manhattan World (MW) assumption is the predominant rule, thus Manhattan frame estimation is well researched for both RGB [16, 30] and RGB-D images [6, 12]. MW assumption serves as a guidance in many applications such as layout estimation [16, 30, 8, 3, 29, 40], camera pose estimation [33, 13] and reconstruction refinement [7, 9].

In addition to the MW assumption, indoor scenes have plentiful lines and planes which provide strong cues for many tasks. Elqursh and Elgammal [5] introduce a line-based camera pose estimation method, while Koch *et al*. [14] use 3D line segments to align the non-overlapping indoor and outdoor reconstructions. Planar patch detection and matching [34, 20, 4, 28, 31, 7, 17] are significantly used strategies to improve the reconstruction accuracy. Some works [34, 20, 4, 28] exploit plane correspondence to solve for frame-to-frame camera poses. Halber *et al*. [7] and Lee *et al*. [17] perform global registration leveraging structural constraints to elevate the scan accuracy. Shi *et al*. [31] use a CNN to learn a feature descriptor for planar patches in RGB-D images. These approaches all hinge on the success of feature matching at the overlapping areas, as opposed to the scenario in this paper.

3. Approach

The input to our system is a set of partially scanned fragments and we output the local layout of each fragment, the

(c) Global placement   (e) Partial data alignment

(a) Input partial data   (b) Local floorplan estimation   (d) Optimal placement and refinement   (f) Layout modeling
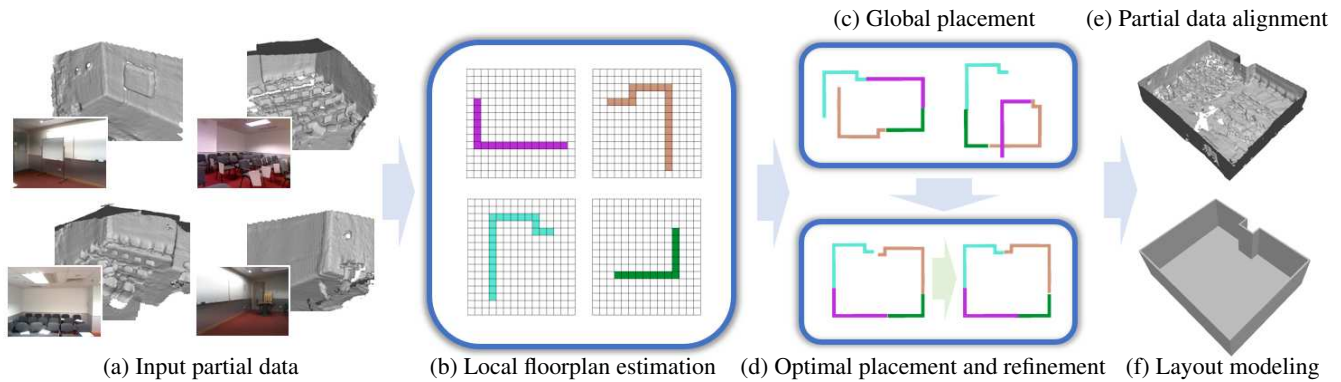
Figure 2: Overview of the proposed method. Given a set of unordered partial reconstructions (a), our algorithm first estimates their local floorplans (b) respectively. Then we compute the poses (c) of all the local floorplans to find a global optimal placement followed by a refinement process (d). Finally, we output the aligned complete reconstruction (e) and predict a final room layout (f) accordingly.

transformations to align them, and a global scene layout. As shown in Fig. 2, our approach consists of three main steps: (1) local layout estimation of each partial reconstruction; (2) optimal placement for global layout estimation; (3) pose refinement to make walls well-aligned. Before running our algorithm, we first extract point feature [27] to combine the partial scans that have more than 60% alignment inliers into one fragment; while the remaining scans can be considered as insufficiently overlapping.

3.1. Local Layout Estimation

We assume that walls obey the MW assumption. Inspired by Cabral and Furukawa [2], we formulate a graph-based shortest path problem to find a floorplan path. As opposed to their reliance on a complete point cloud with a closed-loop as input, we come up with new strategies dealing with partial input that may contain incomplete or partially occluded walls.

Preprocessing. We extract the planes using RANSAC and compute three MW directions $\{X_m, Y_m, Z_m\}$ [12]. For convenience, we set the X_m axis as the world up direction by assuming that the camera optical axis is roughly horizontal to the ground when the scanning begins, and the Y_m and Z_m axes are the wall directions. Then the local camera coordinates are aligned to the MW coordinates by the minimal rotation.

Wall estimation graph. We project all points of the fragment f_k onto a grid with cell size s. A cell that receives more than N projected vertices is considered as a high wall-evidence cell, where we use $s = 8cm$ and $N = 20$ in this paper. We search over the grid to look for contiguous sets of cells with high wall-evidence to extract candidate wall segments, such as w_1, w_2 and w_3 in Fig. 3.

Given a set of wall candidates, we build a wall estimation graph (*WE-graph*) where the nodes are the candidate keypoints of wall structures (e.g., wall corners) and the edges are the candidate walls. Due to noise and occlusion, the

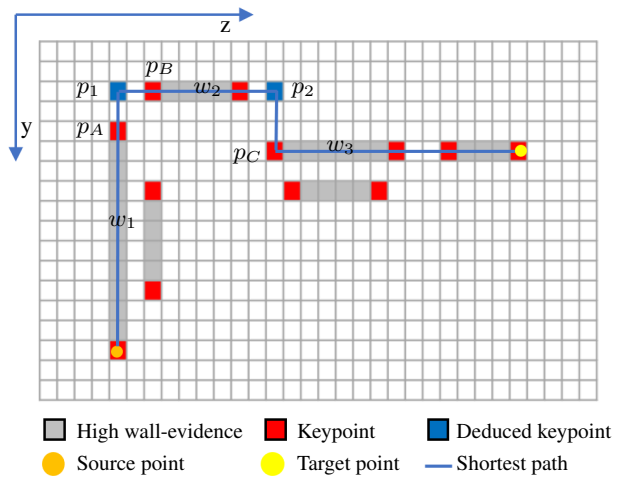| High wall-evidence | Keypoint | Deduced keypoint |
| Source point | Target point | Shortest path |

Figure 3: Local floorplan path determination. Points are projected onto the ground plane and discretized into a grid.

endpoints (red cells in Fig. 3) may not exactly be wall corners. We therefore need to reason out more candidate keypoints (e.g., p_1, p_2) to derive a complete wall structure.

Here we consider two typical cases: (1) two neighboring perpendicular candidate wall segments can be extended to an intersection point which may imply a potential wall corner, e.g., p_1 is deduced from w_1 and w_2 in Fig. 3; (2) two neighboring misaligned parallel candidate wall segments may imply an occluded wall in the invisible intermediate region. See w_2 and w_3 in Fig. 3, we project $p_C \in w_3$ to the line of w_2 to deduce a new keypoint p_2, and re-mark the cells between p_2 and p_C as high wall-evidence.

We set both of the deduced points (blue cells) and the wall endpoints (red cells) as the graph nodes. Then edges are added for every pair of the nodes as long as they are aligned to either Y_m or Z_m axis. The edge weight of a potential wall w is defined as

$$\frac{L(w) - H(w)}{H(w)} + \lambda, \tag{1}$$

where $L(w)$ is the length of w on the grid, and $H(w)$ is the number of high-evidence cells. The first term is to encourage edges to not only have fewer low wall-evidence cells but also be longer. The second term is a constant complexity penalty with $\lambda = 0.1$ (see the evaluation in Fig. 10). Through these two terms, we encourage the final path to have higher wall-evidence, be longer and simpler.
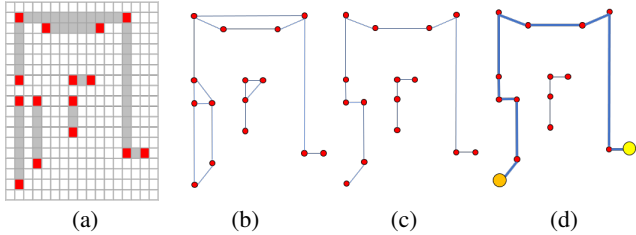

(a)   (b)   (c)   (d)

Figure 4: Source and target point determination for a partial scan. (a) Projection grid; (b) ST-graph; (c) minimal spanning forest (MSF); (d) source and target points derived from the longest path on the MSF.

Source and target determination. To solve for the floorplan path from an incomplete reconstruction that does not have a clear boundary, as shown in Fig. 4, we build another graph (*ST-graph*) to determine the source and target points. The edge weight in the *ST-graph* is the Euclidean distance between two nodes in the grid coordinate system. We compute the minimal spanning forest (MSF) of the graph to encourage the nodes to be connected by the minimal distance cost. Then we solve for the longest path on the MSF. The source and the target points are two endpoints of this longest path, where the first point in the clockwise sequence is considered as the source and the other as the target.

Finally, we find the minimum cost path from the source to the target on the *WE-graph* as the local layout estimation result.

3.2. Global Layout Placement

To determine the global layout, we need to find the rigid transformations for all partial fragments that do not have sufficient matched-overlap. We observe that under the MW assumption, the rotation of each partial fragment can be viewed as the alignment of its local MW coordinate to the world one; the translations of the small-overlapping fragments can be approximately viewed as the sequence in the global loop closure path where all of the local paths are concatenated end-to-end, see Fig. 5 for an example.

Given the local MW coordinate axes $\{X_m, Y_m, Z_m\}$ of a fragment and the world coordinate axes $\{X_w, Y_w, Z_w\}$, we first align the up direction X_m of the local MW coordinate to the world up direction X_w (see Preprocessing in Sec. 3.1) . Then the remaining correspondences from Y_m, Z_m to Y_w, Z_w have four different choices which compose the solution space of rotations. Let $f \in \{1, ..., N\}$ index all the

partial fragments, $R_f \in \{1, 2, 3, 4\}$ the candidate rotations of fragment f corresponding to the alignment from Y_m to Y_w, Y_m to $-Y_w$, Y_m to Z_w or Y_m to $-Z_w$ respectively, and $t_f \in \{1, ..., N\}$ the clockwise sequence of the fragment f on the floorplan loop.

A candidate placement is denoted as a tuple $\{f, R, t\}$ where the subscript is omitted for simplicity. It indicates the rotations and sequences for all the fragments as well as the room layout derived by the end-to-end concatenation of the local layout paths. We then define the binary variables $x_{f,R,t} \in \{0, 1\}$ to indicate whether the candidate placement exists in the solution set. The total energy is defined as

$$\min_{\boldsymbol{x}=\{x_{f,R,t}\}} E_l(\boldsymbol{x}) + E_c(\boldsymbol{x}) + E_b(\boldsymbol{x}), \qquad (2)$$

$$s.t. \quad \forall f \sum_{R,t} x_{f,R,t} = 1, \quad \forall t \sum_{f,R} x_{f,R,t} = 1, \qquad (3)$$

where E_l is the complexity of a layout, E_c the closure measurement, and E_b the similarity of the boundary between adjacent fragments. The constraints in Eq. (3) enforce mutual exclusion, i.e., each fragment and sequence index can only appear once in the final solution.

Layout complexity term. We form the complexity term E_l by summing up the number of wall corners and the number of edges in the convex hull of the floorplan, where the lowest energy encourages that the room not only contains fewer corners but also has simpler overall structure. See Fig. 5, (a) and (b) are two different placements for the same set of local layouts. Although they have the same number of wall corners, we prefer (a) since it has more aligned collinear wall segments which lead to fewer edges in the convex hull.
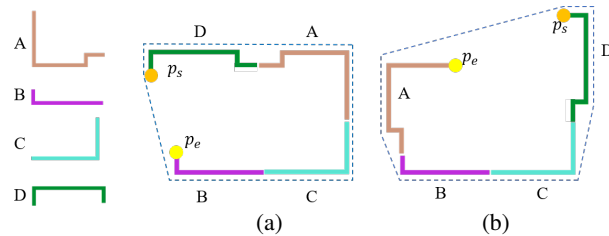

(a)   (b)

Figure 5: Two different placements via end-to-end local layout concatenation.

Closure term. The second term E_c denotes the closure of a layout path, by which we wish the gap between the start point and the endpoint on the final path to be as small as possible. See Fig. 5 for an example of computing this term, the closure is measured by the Manhattan distance (in meters) between the start point p_s of $x_{f,R,1}$ and the endpoint p_e of $x_{f,R,N}$.

Boundary similarity term. As shown in Fig. 6, the cutting plane going through the source or the target point on a local floorplan path is defined as the boundary plane (e.g., $\mathcal{B}_i$ and $\mathcal{B}_j$). The points within $10cm$ of the cutting plane are considered as the boundary points (e.g., $\mathcal{P}_i$ and $\mathcal{P}_j$). We refer to
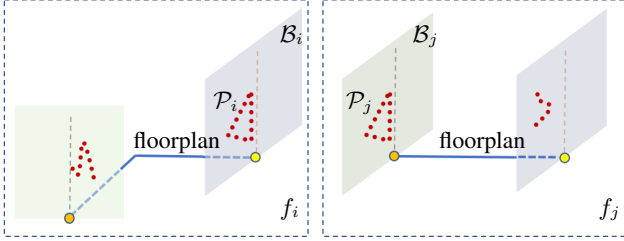
Figure 6: Analysis of the boundary similarity when f_j is placed next to f_i. $\mathcal{B}_i$ and $\mathcal{B}_j$ are two adjacent boundary planes; $\mathcal{P}_i$ and $\mathcal{P}_j$ are the boundary point sets around the planes, which are used for computing boundary similarity.

the probabilistic method [1] to analyze the match quality of the boundary points between two adjacent fragments, and obtain a mismatch score between 0 and 1. We sum up the mismatch scores of all adjacent pairs to compute E_b.

To solve this constrained 0-1 programming problem (Eq. (2)), we search for the global minima based on a DFS tree with alpha-beta pruning. Additionally, we also prune the invalid branches where walls incorrectly cross each other to further improve the efficiency.

3.3. Pose Refinement

The global layout placement encourages all fragments to form a loop closure without taking wall alignment into consideration. Thus in this step, we aim to refine the positions of all fragments by constraining the layout alignment.
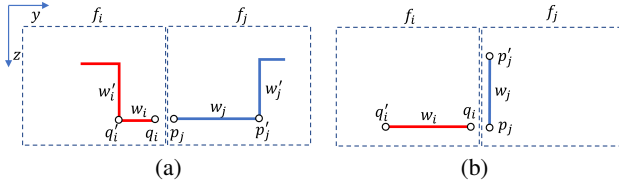

Figure 7: Two types of wall joints between two adjacent fragments f_i and f_j. (a) The connected walls are parallel; (b) the connected walls are perpendicular.

Let the sequence of local layouts be $\{f_1, f_2, ..., f_N\}$ on the loop. Since the walls are aligned to either Y or Z axis of the world coordinate system, we define $t_i = (y_i, z_i)$ to represent the translation to adjust the current position of the layout f_i. Meanwhile, we use q_i and p_j to denote the target point in f_i and the source point in f_j respectively, while p_i' and q_j' are their neighboring keypoints (corner-point or end-point) in the same local layout accordingly (see Fig. 7 for an illustration). There are two typical configurations of wall connection when f_j is placed next to f_i and the constraints are added accordingly as follows.

Parallel connection (Fig. 7 (a)). Two adjacent local layouts f_i and f_j are joined by two parallel walls. The walls are aligned along either the Y axis or the Z axis, while we only discuss the Y-aligned case which is shown in Fig. 7

(a). First, the Z coordinates of q_i and p_j should be equal or else the walls are misaligned. Second, given two joined walls w_i and w_j with the lengths l_{w_i} and l_{w_j} respectively, if $l_{w_i} \le l_{w_j}$, then p_j can not go across q_i' or else w_j will intersect with w_i' which is illegal. The constraints are defined as follows where $\alpha = min\{l_{w_i}, l_{w_j}\}$:

$$z_{q_i} + z_i = z_{p_j} + z_j,$$
$$(l_{w_i} + l_{w_j}) - |(y_{q_i'} + y_i) - (y_{p_j'} + y_j)| < \alpha. \quad (4)$$

Perpendicular connection (Fig. 7 (b)). Two adjacent local layouts f_i and f_j are jointed by two perpendicular walls. We only discuss the case of Fig. 7 (b) where w_i is aligned along the Y axis and w_j the Z axis. To avoid illegal crossing between w_i and w_j, p_j cannot go across w_i while q_i cannot go across w_j. The constraints are defined as:

$$y_{q_i} + y_i < y_{p_j} + y_j$$
$$z_{p_j} + z_j < z_{q_i} + z_i. \quad (5)$$

To solve for the adjustments $\mathbf{t} = \{(y_i, z_i)\}$ for all pairs of local layouts, we formulate an optimization problem to minimize the distance between the joints of the adjacent local layouts as follows:

$$\min_{\mathbf{t}} \sum_{(i,j)\in\mathcal{C}} ((q_i + t_i) - (p_j + t_j))^2. \quad (6)$$

Here $\mathcal{C}$ indicates the set of the pairs of the adjacent local layouts. Finally, we obtain the translations $\{(y_i, z_i)\}$ for all local layouts by solving Eq. (6) under the constraints (4) and (5), and update the final layout.

4. Experimental Results

We evaluate our algorithm using 101 scenes collected from SUNCG dataset [32], SUN3D dataset [39] and our real-world scanning. Each scene is given by a set of partial reconstructions derived from the region-by-region capturing or the failures of camera localization. A challenge in our testing data is, there may not be sufficient overlap among the partial data. Our dataset covers representative indoor layouts of which the scene area varies from $2m \times 6m$ to $18m \times 20m$, and the number of wall corners varies from 4 to 16. All the experiments are performed on a machine with Intel Core i7-7700K 4.2GHz CPU and 32GB RAM.

Evaluation metrics. We evaluate the performance of our method by the metrics defined below. A local or global layout estimation is correct if the average distance error between the estimated wall keypoints and the ground truth keypoints is below 5% relative to the length of the diagonal of the bounding box. A global placement is correct if the placement can lead to a correct global layout estimation. We use ACC_{local} to represent the percentage of the correct local estimations against all of the partial fragments in the dataset. Similarly, ACC_{global} represents the percentage of
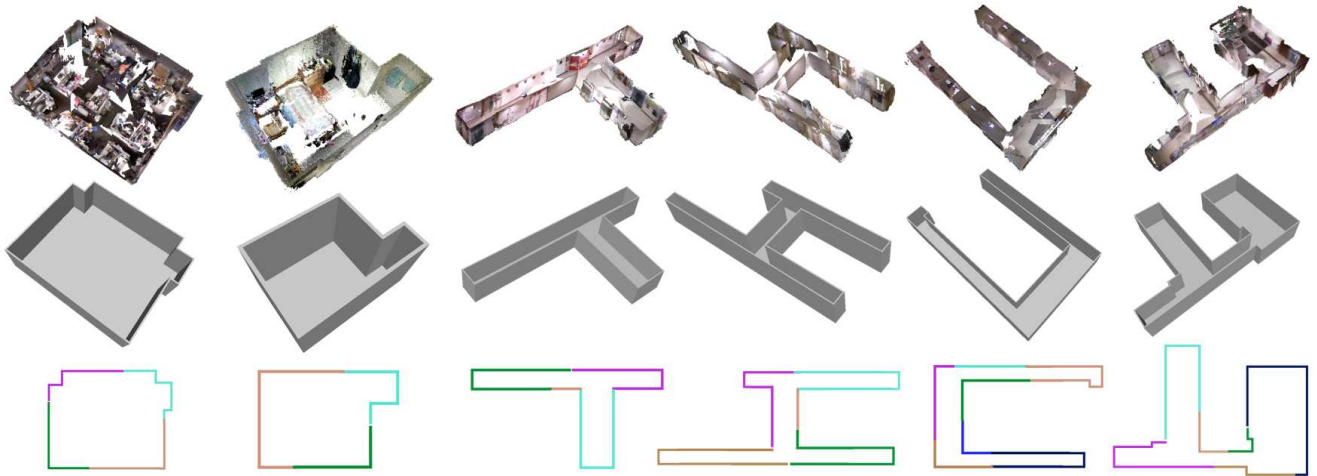
Figure 8: Results of the partial reconstruction alignment and the global layout estimation.

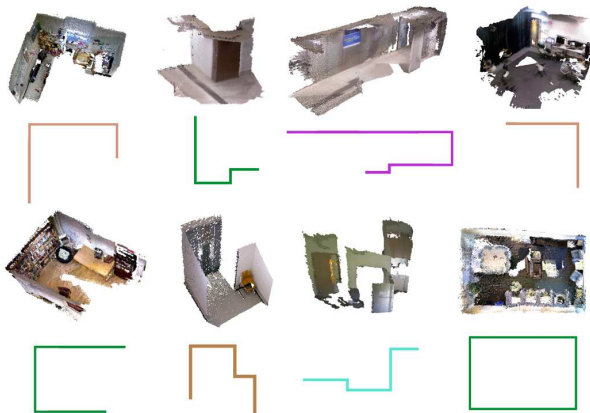the correct global placements against all scenes.

Figure 9: Results of the partial layout estimation.

Partial layout estimation. Our method is able to robustly estimate a partial room layout given an incomplete reconstruction without a closed boundary. Our testing data contains 401 various partial reconstructions, on which our method achieves $ACC_{local} = 98.3\%$. We also show some qualitative results in Fig. 9. Note that: (1) some walls are not captured in the point cloud but our method can still robustly estimate the correct layouts; (2) although our method targets partial data, it can be directly applied to estimate the layout of a complete reconstruction as well.

We evaluate the effect of different values of the complexity penalty λ in Eq. (1). Fig. 10 shows that a large λ tends to ignore the detailed structures and produce a simple layout. We fix λ to 0.1 to generate all of the results in this paper.

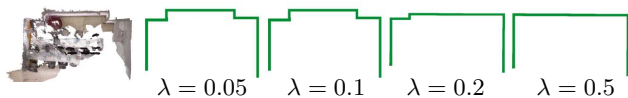$\lambda = 0.05$ $\lambda = 0.1$ $\lambda = 0.2$ $\lambda = 0.5$

Figure 10: The effect of the parameter λ of the penalty term.

Global layout placement. Fig. 8 shows some results of the partial scan alignment and the global layout estimation. Our method faithfully reconstructs some large-scale scenes by combining a set of partially scanned point clouds. We also quantitatively evaluate our method in Table 1. As an ablation study, Table 1 shows the performance given different configurations of the three terms in Eq. (2): (1) without closure term; (2) without complexity term; (3) without boundary similarity term; (4) full terms. The experiments demonstrate that the full configuration using all these three terms performs the best.

Configuration	$ACC_{global}(\%)$
w/o closure term	22.8
w/o complexity term	67.5
w/o boundary similarity term	80.2
full terms	**85.1**

Table 1: Performance of our method on global layout placement using different configurations.

Pose estimation error. We evaluate the pose estimation error on the synthetic scenes collected from SUNCG [32] dataset with ground truth camera poses. We also compare our method with the state-of-the-art 3D registration algorithms, including 3DMatch [42], Fast point feature histogram (FPFH) [27], and Orthogonal plane-based visual odometry (OPVO) [13]. Note that OPVO is also proposed under the MW assumption. Table 2 reports the angle error of rotation and the distance error of translation relative to the length of the diagonal of the bounding box. Since our testing data may not have sufficient overlap, we find that existing methods based on feature descriptors perform poorly under the same condition. Qualitative comparisons in Fig. 11 and quantitative comparisons in Table 2 both show that the other methods produce unacceptable inaccuracies, while our method achieves superior results.
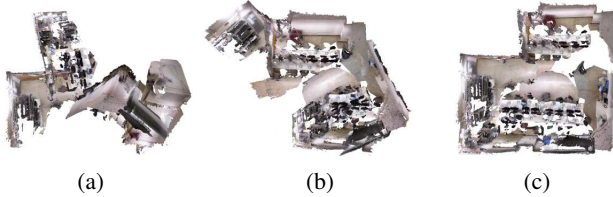
Figure 11: Qualitative comparison with point cloud alignment methods using feature descriptors. (a) 3DMatch [42]; (b) FPFH [27]; (c) ours.

Method	Rotation(°)	Translation(%)
3DMatch [42]	43.41	21.82
FPFH [27]	40.05	29.12
OPVO [13]	43.06	20.04
Ours	**8.79**	**9.15**

Table 2: Quantitative comparison on the SUNCG synthetic dataset [32] in terms of rotation angle error and translation distance error.

Method	Avg (%)	Max(%)
MW Modeler [18]	1.22	4.47
PolyFit [24]	1.31	5.01
RAPTER [21]	1.40	7.84
Ours	**0.90**	**2.57**

Table 3: Comparison with the state-of-the-art structured modeling methods in terms of layout reconstruction error.

Layout reconstruction quality. Manhattan-world Modeler [24], PolyFit [18] and RAPTER [21] are the state-of-the-art structured modeling methods for man-made scenes which take as input scanned point clouds. To compare with them in terms of layout reconstruction quality, we input to these methods the complete point clouds of the scenes in our dataset. Fig. 12 shows a set of qualitative comparison results. We are able to obtain considerably better results with accurate and high-quality wall structures.

Table 3 shows the quantitative comparison results with these methods. We uniformly sample points on the ground truth layout, and compute the distance error of the point samples to their nearest faces in the reconstructed model. We report the average and maximal error relative to the length of the diagonal of the bounding box. The results demonstrate that our method has smaller layout reconstruction errors than the other structured modeling methods.

Time efficiency. For the local layout estimation, on average our algorithm takes about 0.1s per 10k points. An exception is the scene of the last column in Fig. 8, where it takes about 200s to process a partial scan with 200k points. This is because a large number of small wall candidates are generated in the local layout estimation step due to heavy noises. For the pose determination and refinement, it takes less than 20s with an input of fewer than 10 fragments.
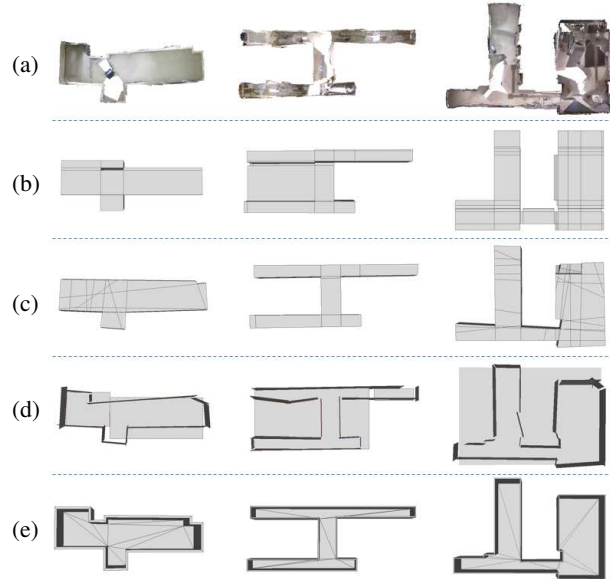

Figure 12: Qualitative comparison on layout reconstruction quality. (a) Input point clouds; (b) MW Modeler [18]; (c) PolyFit [24]; (d) RAPTER [21]; (e) ours.

Ambiguity and failure case. The optimal placement of the given local layouts may be ambiguous, which will result in an incorrect sequence (Fig. 13 (a)) or an incorrect layout (Fig. 13 (b)), although all the different results seem to be reasonable. The boundary similarity term in Sec. 3.2 is designed to alleviate this problem, however, if an ambiguity still occurs, more constraints need to be added to derive the correct result, e.g., user-specific fragment sequence.

Before running our algorithm, we first extract point feature [27] to combine the partial scans that have sufficient overlap into larger fragments. If there is large overlap between partial reconstructions but not detected successfully, our algorithm is likely to exhibit large error or output an incorrect result. We show a failure case in Fig. 14, where our result is not consistent with the ground truth.
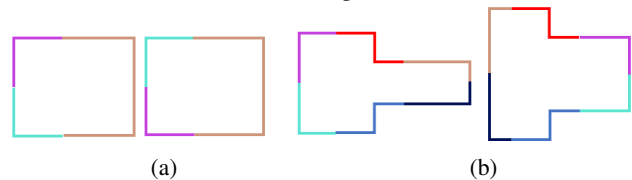

Figure 13: Ambiguity of placements. (a) Different placements produce the same layout; (b) different placements produce different layouts but both are reasonable.

5. Applications

Since our method does not depend on feature matching to align 3D data, it facilitates several novel applications. In this section, we demonstrate the following three.

Featureless scene reconstruction. For scenes that have a

Ground truth   Overlap is large but not detected   Estimated layout

Figure 14: A failure case where the input fragments have large overlap but not successfully detected by feature descriptor matching.

large expanse of featureless walls, it is very difficult for the existing methods to reconstruct them by continuous feature tracking. Fig. 15 shows the advantage of our method in reconstructing this kind of scene, while we directly align a set of partial scans caused by camera interruptions without using feature matching.

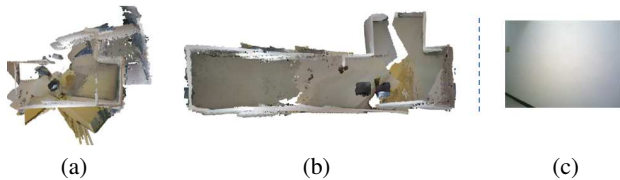
(a)   (b)   (c)

Figure 15: Reconstruction results of a scene with a large expanse of featureless walls. (a) The reconstruction result by continuous camera tracking using ORB-SLAM visual odometry [22]; (b) our result by aligning partial scans; (c) a featureless wall that fails camera localization in this scene.

Large scene reconstruction. As aforementioned, scanning a large scene region-by-region is easier than in a single pass due to the heavy workload, the accumulation error and the memory limit of a computer. Fig. 16 shows the reconstruction results for a large scene using different strategies. In practice, we pay more efforts to maintain the uninterrupted scanning, but it still exhibits large accumulative errors. Instead, using region-based scanning, the scene is first divided into sub-regions and scanning each one separately is easier. Also, this strategy achieves better accuracy as illustrated.
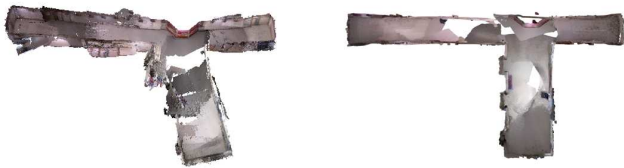

Figure 16: Reconstruction results of a large scene. Left: the result by continuous camera localization using ORB-SLAM visual odometry [22]; right: our result by aligning a set of partial scans.

Modeling from sparse input and down-sampling. The proposed method can recover a room layout from a small number of RGB-D images without adequate overlap, which can be used to model a scene given sparse input and down-

sample the RGB-D stream in a scanning system (e.g., Matterport scanning system) for efficiency. As shown in Fig. 17, our method successfully aligns the RGB-D sequences and estimates the room layouts accordingly, which shows the ability of our method in modeling from sparse input.

Figure 17: Room layout modeling and camera pose estimation by stitching sparse RGB-D frames.

6. Conclusion

In this work, we propose a novel approach to jointly align a set of partial reconstructions caused by camera interruptions and predict a room layout. Instead of relying on feature descriptor matching, our method is able to estimate the transformations of the partial 3D data without sufficient overlap, which is proved to be a challenge for the existing methods. Technically, we first estimate a local layout for each partial data and further formulate an optimal placement problem to combine these local layouts into a global loop closure under certain constraints. We have evaluated our algorithm quantitatively and qualitatively and compared it with the state-of-the-art methods, all of which demonstrate the effectiveness of our method on the alignment of small-overlapping partial scans as well as the global (partial) room layout estimation.

References

[1] Igor Bogoslavskyi and Cyrill Stachniss. Analyzing the quality of matched 3d point clouds of objects. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6685–6690. IEEE, 2017.

[2] Ricardo Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 628–635. IEEE, 2014.

[3] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Understanding indoor scenes using 3d geometric phrases. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 33–40. IEEE, 2013.

[4] Alejo Concha Belenguer and Javier Civera Sancho. Dpptam: Dense piecewise planar tracking and mapping from a monocular sequence. In *Proc. IEEE/RSJ Int. Conf. Intell. Rob. Syst.*, number ART-2015-92153, 2015.

[5] Ali Elqursh and Ahmed Elgammal. Line-based relative pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3049–3056. IEEE, 2011.

[6] Bernard Ghanem, Ali Thabet, Juan Carlos Niebles, and Fabian Caba Heilbron. Robust manhattan frame estimation from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3772–3780, 2015.

[7] Maciej Halber and Thomas Funkhouser. Fine-to-coarse global registration of rgb-d scans. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[8] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1849–1856. IEEE, 2009.

[9] Jingwei Huang, Angela Dai, Leonidas Guibas, and Matthias Nießner. 3dlite: towards commodity 3d scanning for content creation. *ACM Transactions on Graphics, 2017*, 2017.

[10] Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. Structured indoor modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1323–1331, 2015.

[11] Andrew E. Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on pattern analysis and machine intelligence*, 21(5):433–449, 1999.

[12] Kyungdon Joo, Tae-Hyun Oh, Junsik Kim, and In So Kweon. Globally optimal manhattan frame estimation in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1763–1771, 2016.

[13] Pyojin Kim, Brian Coltin, and Hyoun Jin Kim. Visual odometry with drift-free rotation estimation using indoor scene regularities. In *BMVC*, 2017.

[14] Tobias Koch, Marco Körner, and Friedrich Fraundorfer. Automatic alignment of indoor and outdoor building models using 3d line segments. *Proceedings of Computer Vision and Pattern Recognition 2016*, pages 10–18, 2016.

[15] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4875–4884. IEEE, 2017.

[16] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2136–2143. IEEE, 2009.

[17] Jeong-Kyun Lee, Jaewon Yea, Min-Gyu Park, and Kuk-Jin Yoon. Joint layout estimation and global multi-view registration for indoor reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 162–171, 2017.

[18] Minglei Li, Peter Wonka, and Liangliang Nan. Manhattanworld urban reconstruction from point clouds. In *European Conference on Computer Vision*, pages 54–69. Springer, 2016.

[19] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floornet: A unified framework for floorplan reconstruction from 3d scans. *arXiv preprint arXiv:1804.00090*, 2018.

[20] Lingni Ma, Christian Kerl, Jörg Stückler, and Daniel Cremers. Cpa-slam: Consistent plane-model alignment for direct rgb-d slam. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 1285–1291. IEEE, 2016.

[21] Aron Monszpart, Nicolas Mellado, Gabriel J Brostow, and Niloy J Mitra. Rapter: rebuilding man-made scenes with regular arrangements of planes. *ACM Trans. Graph.*, 34(4):103–1, 2015.

[22] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[23] Srivathsan Murali, Pablo Speciale, Martin R Oswald, and Marc Pollefeys. Indoor scan2bim: Building information models of house interiors. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 6126–6133. IEEE, 2017.

[24] Liangliang Nan and Peter Wonka. Polyfit: Polygonal surface reconstruction from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2353–2361, 2017.

[25] Muzammal Naseer, Salman H Khan, and Fatih Porikli. Indoor scene understanding in 2.5/3d: A survey. *arXiv preprint arXiv:1803.03352*, 2018.

[26] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[27] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217. IEEE, 2009.

[28] Renato F Salas-Moreno, Ben Glocken, Paul HJ Kelly, and Andrew J Davison. Dense planar slam. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 157–164. IEEE, 2014.

[29] Alexander G Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 353–360. IEEE, 2013.

[30] Alexander G Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2815–2822. IEEE, 2012.

[31] Yifei Shi, Kai Xu, Matthias Niessner, Szymon Rusinkiewicz, and Thomas Funkhouser. Planematch: Patch coplanarity prediction for robust rgb-d reconstruction. *arXiv preprint arXiv:1803.08407*, 2018.

[32] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[33] Julian Straub, Nishchal Bhandari, John J Leonard, and John W Fisher. Real-time manhattan world rotation estimation in 3d. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 1913–1920. IEEE, 2015.

[34] Yuichi Taguchi, Yong-Dian Jian, Srikumar Ramalingam, and Chen Feng. Point-plane slam for hand-held 3d sensors. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 5182–5189. IEEE, 2013.

[35] Camillo J Taylor and Anthony Cowley. Parsing indoor scenes using rgb-d imagery. In *Robotics: Science and Systems*, volume 8, pages 401–408, 2013.

[36] Thomas Whelan, Renato F Salas-Moreno, Ben Glocker, Andrew J Davison, and Stefan Leutenegger. Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016.

[37] Erik Wijmans and Yasutaka Furukawa. Exploiting 2d floorplan for building-scale panorama rgbd alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 308–316, 2017.

[38] Kai M Wurm, Cyrill Stachniss, and Wolfram Burgard. Coordinated multi-robot exploration using a segmentation of the environment. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 1160–1165. IEEE, 2008.

[39] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[40] Hao Yang and Hui Zhang. Efficient 3d room shape recovery from a single panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5422–5430, 2016.

[41] Jiaolong Yang, Hongdong Li, and Yunde Jia. Go-icp: Solving 3d registration efficiently and globally optimally. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1457–1464, 2013.

[42] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017.

[43] Jian Zhang, Chen Kan, Alexander G Schwing, and Raquel Urtasun. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1273–1280. IEEE, 2013.

[44] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European Conference on Computer Vision*, pages 668–686. Springer, 2014.

[45] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.

[46] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2051–2059, 2018.