3D Room Layout Recovery Generalizing across Manhattan and Non-Manhattan Worlds

Haijing Jia[1], Hong Yi[1], Hirochika Fujiki[2], Hengzhi Zhang[1]
Wei Wang[1], Makoto Odamaki[2]
[1]Ricoh Software Research Center (Beijing) Co., Ltd. Beijing, China
[2]Ricoh Company, Ltd. Japan
{Haijing.Jia, Hong.Yi, Hengzhi.Zhang, Wei.Wang5}@cn.ricoh.com
{hirochika.fujiki, makoto.odamaki}@jp.ricoh.com

Figure 1. 3D room layout recovery results. The first column shows the predicted boundaries (the ceiling and floor boundaries in green, and the wall-wall boundary in blue) and surface normal (Manhattan surface in blue and non-Manhattan surface in yellow) by our method. The second and third columns are the final room layouts and the generated 3D models. Qualitative comparisons against two competing methods under the floor plan view are shown in the last column. HorizonNet [37] fails in non-Manhattan layouts and AtlantaNet [28] fails in Manhattan layouts with occlusions. Only our method is capable of generalizing across Manhattan and non-Manhattan layouts.

Abstract

Recent 3D room layout recovery approaches mostly concentrate on Manhattan layouts, where the vertical walls are orthogonal with respect to each other, even though there are many rooms with non-Manhattan layouts in the real world. This paper presents a room layout recovery method generalizing across Manhattan and non-Manhattan worlds. Without introducing additional supervision, we extend current Manhattan layout recovery methods by predicting an extra surface normal feature, which is further used for an adaptive post-processing to reconstruct layouts of arbitrary shapes. Experimental results show that ***our method has a great improvement on non-Manhattan layouts while being capable of generalizing across Manhattan and non-Manhattan layouts.***

1. Introduction

The room layout recovery is a lightweight indoor 3D reconstruction [21, 30] technology which aims to predict the indoor 3D geometric structure from images. It plays an important role in many indoor scene understanding applications, such as virtual tour roaming [29], floor plan reconstruction [9] and navigation guidance [20]. The methods [2, 11, 28, 37, 44–46, 51, 57] in recent years address the

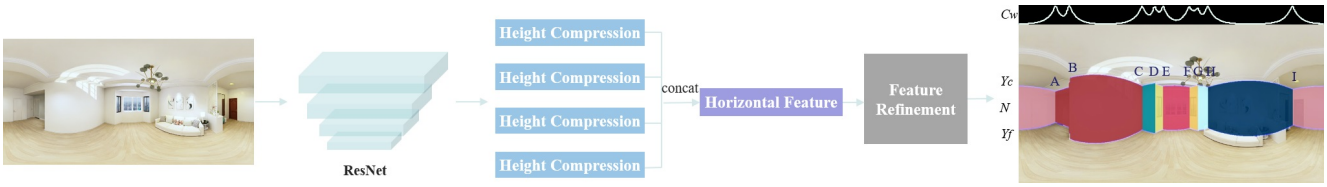

Figure 2. The network architecture. The network takes a 360° panorama as the input and passes it to a ResNet [14]. Then the pyramid features are compressed and concatenated to generate the initial horizontal feature. A feature refinement block is followed for further feature extraction. Our network jointly predicts the ceiling boundary *Yc*, the floor boundary *Yf*, the wall-wall existence *Cw* and the wall surface normal *N*.

room layout recovery problem from a single panorama image, which captures the complete 360° FOVs of the environment.

In room layout recovery research, a room is assumed to be an Atlanta world [34] layout with a horizontal floor, ceiling and vertical walls. It is further divided into Manhattan [3] and non-Manhattan world layouts based on whether the vertical walls are orthogonal with respect to each other.

Previous works [11, 37, 46, 57] mostly concentrate on Manhattan layouts. Relying on the strong Manhattan assumption [3], they can provide excellent accuracy for the simplest Manhattan layout, the cuboid (with 4 walls), as well as more complex general Manhattan layouts such as "L"-shape and "T"-shape. The **Manhattan assumption is great, but it fails once encountering non-Manhattan layouts**.

The non-Manhattan room layout recovery is a challenging topic due to its complexity and the lack of non-Manhattan datasets. A few researchers [2, 28] try to learn non-Manhattan features from the rare non-Manhattan data picked out from the public datasets, and they remove the strict Manhattan constraints in the post-processing and approximate a simple polygon as the room layout, which relaxes the Manhattan constraints to arbitrary layouts, but this method introduces **problems** such as: (1) it's impossible to recover the occluded areas and refine the right-angle structures, and (2) there are many false alarms for walls recovery in the final room layout.

The motivation of this work stems from the idea that if a room is known to be Manhattan or non-Manhattan layout, then the corresponding post-processing can be applied. The layout type can be obtained indirectly through the surface normal of the walls. This paper presents a new room layout recovery method that generalizes across Manhattan and non-Manhattan worlds. The **contribution** of this work includes:

1). Without introducing additional supervision, we extend current Manhattan room layout recovery methods to the Atlanta world.

2). Our network not only estimates regular layout elements like the ceiling, floor, and wall-wall boundaries, but also estimates additional surface normals to indicate the rough structure of the room, such as which walls are Manhattan or non-Manhattan surfaces (see Figure 2), which are further used in an adaptive post-processing to reconstruct the layout of arbitrary shapes (see Figure 3d).

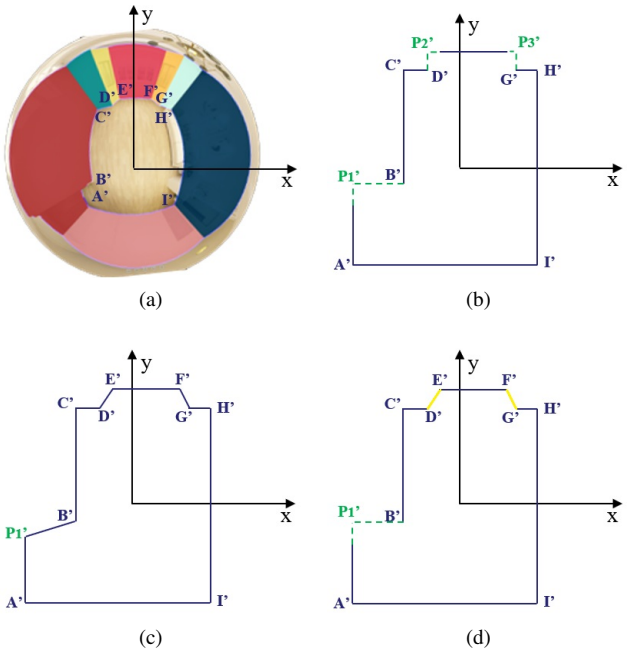

Figure 3. (a) is the ceiling boundary *Yc* projected floor plan on the ceiling plane; (b), (c) and (d) are the generated floor plans by Manhattan assumption based post-processing, non-Manhattan post-processing and our adaptive post-processing methods. For the input panorama in Figure 2, apparently (d) is correct. Boundaries and points added by post-processing are shown in green, and non-Manhattan boundaries are shown in yellow.

3). Different from the traditional pixel-wise surface normal, our model predict column-wise surface normal of the walls. The surface normal is encoded as 1D representations, then learned together with the ceiling, floor, and wall-wall boundaries in a single network.

4). The experimental results demonstrate that our method

has **a great improvement** (a boost of 21.0% in 2D IoU and 21.37% in 3D IoU than the state-of-the-art AtlantaNet [28]) **on non-Manhattan layouts while being capable of generalizing across Manhattan and non-Manhattan layouts.**

2. Related works

Room layout recovery is an active research topic in both computer vision and computer graphics fields. Various works have been proposed in recent years which differ in the inputs (perspective images or 360° panoramas), features (edge-based, corner-based or hybrid) and methods (traditional image processing or deep learning).

In terms of the input images, some methods target at estimating the room layout from a single-view perspective image. The typical pipeline is to generate a set of layout hypotheses from the extracted features such as: line segments [15, 25, 31], semantic segmentation [17, 18], and volumetric reasoning [24], and apply iterative optimization or voting to rank, so the top ranked layout is selected. Then the great success of deep networks motivates the wide application of deep learning features in room layout restoration tasks. Mallya et al. [27] and Ren et al. [33] predict the informative edges separating the geometric classes (walls, the floor and the ceiling) with FCN-based models. [50, 53] follow this scheme but exploit more effective features to improve the prediction quality. Dasgupta et al. [5] estimate labels for each of the surfaces of the room instead of edge maps. [12] exploits the combination of geometry line segments and the deep learning edge map. Later, RoomNet [23] proposes a new pipeline which extracts a set of room layout keypoints with an end-to-end encoder-decoder network and connects the keypoints in a specific order.

Recently, with the increasing popularity of 360° cameras, many computer vision tasks, such as object classification [8], detection [13, 47], depth estimation [42] and image enhancement [52], etc., operate on panoramic images which covers the complete 360° FOVs of the environment. PanoContext [51] firstly proposes to estimate the layout from a 360° panorama while inherits the earlier pipeline of vanishing point estimation, room layout hypothesis generation and room layout sampling based on Geometric Context (GC) [18], Orientation Map (OM) [15] and the objects in the room. Pano2CAD [44] follows this pipeline and combines surface normal estimation. Yang et al. [36] recover the 3D room shape from a collection of partially oriented super pixel facets and line segments. Beside geometric cues, [48] adds semantic cues (saliency and object detection maps) to enrich image features. All these works extract image features from a set of projected perspective images instead of the original panoramas.

LayoutNet [57] is the first work to learn the layout features directly on a single RGB panorama together with the Manhatan line map. The encoder -decoder network jointly predicts the dense corner and boundary probability maps. [11] uses a deeper encoder ResNet-50 [14] to extract better low-level features. [33] extends the input to the combination of panorama and a perspective celling-view image and the input of AtlantaNet [28] is the combination of the projected floor and ceiling views. All the above methods exploit CNNs as the feature extractor, while [10] applies EquiConvs, a convolution applied directly on the spherical projection, therefore invariant to the distortions in panorama images. Unlike the traditional dense prediction on the 2D image in [11, 33, 57], [6, 19, 37] extract layout elements on each image column, so that the room layout is encoded as several 1D vectors. This kind of per-column modality is shown to be effective for improving the room layout recovery performance [58], and latter works [2, 32, 38, 43] all inherit this idea. Most recent papers [7, 38, 49] focus on multi-task collaborative optimization which offers redundant and complementary information from different perspectives. HoHoNet [38] models the layout structure, dense depth and semantic segmentation in one framework and employs multi-head self-attention [40] to run faster and improve accuracy. [49] first predicts the coarse depth and semantic segmentation to enforce the layout depth estimation, then uses the layout depth to recover the 3D layout. [56] relies on spherical coordinate localization using geodesic heatmaps and directly infers Manhattan-aligned outputs without any post-processing.

The main limitation of the existing methods [11, 33, 37, 49, 57, 58] is their heavy reliance on the powerful Manhattan assumption. They often fail once encountering non-Manhattan layouts. In order to relax this geometric constraint, [10] extracts layout corners from the corner map directly, but requires that all the corners are visible. [2] focuses on refining the visible layouts by detecting the discontinuity from the predicted boundaries and post-processing the discontinuity on 2D panoramic image and 3D layout respectively. AtlantaNet [28] considers the largest connected region of the ceiling segmentation mask resulting from the network as the room shape. These methods can handle non-Manhattan cases, but lacks accuracy on the rectangular corners because they do not assume any perpendicularity between walls. Many works [2, 33, 58] propose to relax Manhattan constrains to the general layouts in their future works, and they all mention the need to introduce additional information or predict the layout from multiple views.

3. Approach

This paper presents a room layout recovery method **generalizing across Manhattan and non-Manhattan worlds**. Without introducing additional supervision, we extend current Manhattan layout recovery methods by predicting an extra surface normal feature, which is further used in an adaptive post-processing to handle the reconstruction of the

layout of arbitrary shapes.

3.1. Surface normal representation

In three dimensions, the surface normal, to a surface at one point is a vector perpendicular to the tangent plane of the surface at the point. Surface normals are difficult to collect and annotate, generally calculated from the depth [26] information. For the layout recovery problem, we don't need the surface normal at each point, and what we care is just the surface normal of walls. We propose to calculate the surface normal ground truth from the layout corner annotations, which reduces time-consuming and labor intensive surface normal labeling work.

Specifically, we project walls' surfaces to the ceiling (or floor) plane, then each wall surface is represented as a line, and the camera center is the origin of the coordinate system as shown in Figure 4b and 4d. Then the problem of predicting 3D normal of wall surfaces is transformed into the problem of predicting 2D normal of the projected wall lines.

In Figure 4, points $A, B, \dots G$ are the ceiling ground truth corners, and $A', B', \dots G'$ are their projected points on the ceiling plane. Given the projected points $A'\,(x_{a'}, y_{a'})$ and $B'\,(x_{b'}, y_{b'})$, the normal of wall $\overrightarrow{A'B'}$ is expressed as

$$N_{\overrightarrow{A'B'}} = \left(\frac{y_{b'} - y_{a'}}{\left|\overrightarrow{A'B'}\right|}, \frac{-(x_{b'} - x_{a'})}{\left|\overrightarrow{A'B'}\right|} \right)$$

and the angle between the normal vector and the X axis (in Figure 4b and 4d) can always be calculated. The surface normal for the wall AB on the panorama is equal to $N_{\overrightarrow{A'B'}}$.

3.2. 1D representations

As Figure 2 shows, we predict 3 layout elements (the ceiling boundary, floor boundary, wall-wall boundary probability) and an additional surface normal for each input panorama. Generally, the layout elements are encoded as three parameters: *Yc*, *Yf* and *Cw*, which represents the position of the ceiling and floor boundaries, and the existence of wall-wall boundary for each image column. This kind of 1D representations (also called horizontal features) have shown successes in room layout recovery problem [6, 19, 37]. To learn room layout features and the surface normal jointly, the walls surface normal is encoded column by column, and the surface normal for each image column is represented as two 1D representations *N(nx, ny)*. *Yc*, *Yf*, *Cw*, *nx* and *ny* vectors each are with shape 1xW.

3.3. Framework

Figure 2 shows an overview of our network. The input is a RGB panorama with shape of 3×H×W. Following the basic architecture of extracting horizontal features in [37, 38],

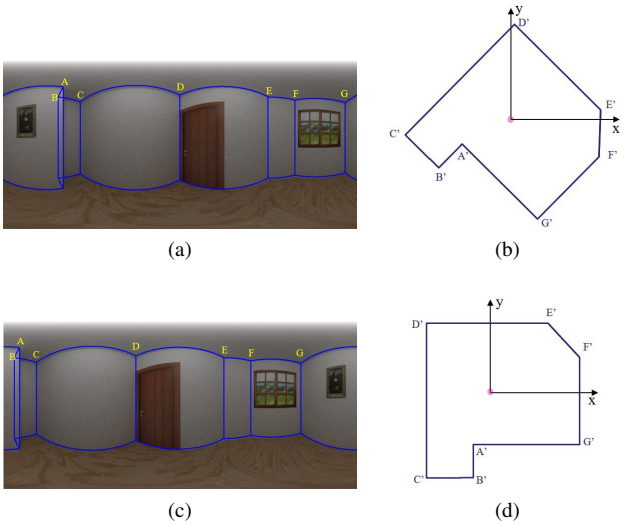

Figure 4. (a) and (c) are two panoramas, and (c) is the result of (a) after alignment. The blue boundaries are room layout ground truths, and $A, B, \dots G$ are ceiling corners. (b) and (d) are the ceiling boundaries in (a) and (c) projected floor plans on the ceiling plane.

the backbone of our network is a ResNet50 with four blocks which extracts features from the input panorama on four different scales. For each feature from the backbone's pyramid, a sequence of convolution layers is applied to gradually compress the height to one and upsample the spatial width to W/4. Then the resulting features are concatenated along channel to the shape of W×1×W/4.

Because any layout corner can be inferred from the positions of other corners, bidirectional LSTM (Bi-LSTM) [16, 35] is adopted to further capture the global feature and long-term dependencies from the horizontal features. And considering the computational and time cost, our LSTM predicts *Yc*, *Yf*, *Cw*, *nx* and *ny* for every four columns, resulting in a 5x4 matrix.

In this work, L1 loss is applied to the regression of *Yc*, *Yf*, *nx* and *ny*, while for *Cw*, the binary cross entropy loss is used. The losses are equally weighted.

3.4. Adaptive post-processing

The success of Manhattan room layout recovery benefits a lot from the Manhattan assumption based post-processing. Figure 3a shows the ceiling boundary projected floor plan on the ceiling plane. Assuming adjacent walls to be perpendicular to each other, the Manhattan post-processing adjusts the walls positions slightly and predicts the hidden corners (see Figure 3b). For non-Manhattan layout, the post-processing method [28] is to approximate a polygon from the projected floor plan as the room shape (see Figure 3c). Apparently, the above post-processing methods are in-

effective for the panoramas with both Manhattan and non-Manhattan surfaces. Manhattan surfaces need to be optimized to right-angle structures but non-Manhattan surfaces don't. In this work, we predict an additional information, the walls surface normal, which indicates the rough structure of the room, such as which walls are Manhattan or non-Manhattan surfaces, then the adaptive post-processing strategies are applied to different surfaces.

The first step of our adaptive post-processing is to find the wall-wall boundary candidates $A, B, \ldots, I$. As Figure 2 shows, the predicted wall-wall probability *Cw* is a line with many peaks and troughs, and each peak corresponds to a wall-wall boundary. Secondly, calculate the surface normal of each wall surface. The network predicts the surface normal for each wall column, then the surface normal is expressed as an angle relative to the X axis. Given the angle of each wall column, we can calculate the average angles of wall surfaces $AB, BC, \ldots, HI$ and IA. For the input panorama in Figure 2, the average angles of wall surfaces AB, BC, CD, EF, GH, HI and IA are 6°, 0°, 267°, 273°, 264°, 181° and 92° respectively. These 7 walls are almost perpendicular or parallel to X axis, and they are Manhattan surfaces. The average angles of wall surfaces DE and FG are 323° and 216°, and they are non-Manhattan surfaces. In this paper, the angle threshold for Manhattan surface is ±10°. Thirdly, we project the ceiling boundary *Yc*, wall-wall boundary candidates $A, B, \ldots, I$ and surface normal to the ceiling plane as Figure 3a shows. Finally, the Manhattan assumption based post-processing is applied to Manhattan surfaces, so that the occluded corner $P1'$ is predicted. For non-Manhattan surfaces $D'E'$ and $F'G'$, we adopt the projected results directly. Figure 3d shows the floor plan result after our adaptive post-processing. Finally, the room height is calculated by averaging over the predicted ceiling and floor positions in each column.

4. Experiments

The input panoramas are of the size of 3×512×1024 and pre-processed by the panoramic image alignment algorithm mentioned in [57], so that whether a wall is Manhattan or non-Manhattan surface is directly determined by the predicted surface normal and angle. The network is trained with Adam optimizer [22] on one Nvidia GeForce Titan X GPU for 400 epochs with batch size 8 and learning rate 0.0004. The data augmentation techniques in [37] are also adopted to our layout elements and the surface normal.

4.1. Datasets

Most of the public room layout datasets are Manhattan datasets, such as PanoContext [51], Stanford 2d-3d [57], Realtor360 [10] and MatterportLayout [58]. In this paper, we evaluate the Manhattan room layout recovery performance on PanoContext and Standford 2d-3d datasets, and follow the dataset split (train, valid and test) adopted by LayoutNet [57], DulaNet [46] and HorizonNet [37].

Non-Manhattan datasets are difficult to collect and many non-Manhattan data are blendered by the synthetic technology. Structured3D [55] is a synthetic dataset including panoramas with mostly Manhattan layouts, and a small amount of non-Manhattan layouts. AtlantaLayout [28] is a small non-Manhatttan dataset with a hundred panoramas selected from Matterport3D [1] and Structured3D [55] datasets.

To train a model generalizing across Manhattan and non-Manhattan worlds with no bias, we organize the Strd3D_non_M dataset, providing 939 training, 75 validation, and 75 test panoramas, which is a subset of Structured3D [55] dataset. Unlike the mostly Manhattan layouts in public datasets, Strd3D_non_M has a balanced quantity distribution of the most common real world layouts (cuboid, general Manhattan and non-Manhattan).

The synthetic data does alleviate the problem of insufficient real world data. However its photo-realism is far inferior from the real world data. To evaluate the layout recovery performance and generalization ability of different methods on the real world non-Manhattan layouts, we collect a RealWorldNonManhattan dataset consisting of 52 panoramas with Ricoh Theta V camera.

4.2. Evaluation metrics

Following Zou et al. [58], we evaluate the room layout recovery performance with six standard metrics, 2D IoU, 3D IoU, CE (Corner Error), PE (Pixel Error), RMSE and $\delta 1$. However, these metrics do not take into account the false alarm results of walls recovery in the generated layouts. These walls do not seem to affect the room shape too much but affect the original room structure (they don't belong to the room). So three new metrics, Junction, Wireframe and Plane, are introduced in General Room Layout Estimation Competition [54], and used to calculate the F-measure of the predicted corners, lines and planes.

4.3. Non-Manhattan results

We compare the non-Manhattan performance of our method and the state-of-the-art non-Manhattan method AtlantaNet [28]. The lack of training code for AtlantaNet prevents it from being retrained, so we evaluate AtlantaNet (a pre-trained model provided in its project page [4] which is trained on MatterportLayout [58] cleaned Manhattan dataset and fine-tuned on AtlantaLayout [28] training set) and our model (trained on Strd3D_non_M) on AtlantaLayout dataset. The results are shown in Table 1 and Figure 5. The AtlantaLayout dataset is rather small, so we put the test and valid sets together as the evaluation set here. The layout labels with serious errors have been corrected and the new labels will be open to public.

Method	AtlantaLayout					RealWorldNonManhattan				
	Junction	Wireframe	Plane	2D IoU	3D IoU	Junction	Wireframe	Plane	2D IoU	3D IoU
AtlantaNet [28]	0.40	0.18	0.54	69.37	**64.92**	0.32	0.12	0.52	64.18	59.60
Ours	**0.42**	**0.22**	**0.82**	**69.93**	63.53	**0.41**	**0.25**	**0.75**	**85.18**	**80.97**

Table 1. Quantitative evaluation of AtlantaNet [28] and our model on the AtlantaLayout [28] and RealWorldNonManhattan dataset.

Figure 5. Qualitative results and comparison of our method and AtlantaNet [28] on the AtlantaLayout [28] (the first two rows) and RealWorldNonManhattan (the third to the last row) datasets. The results of our method are shown in blue boundaries and AtlantaNet [28] are shown in red boundaries.

Method	Manhattan				Non-Manhattan				Overall			
	2D IoU	3D IoU	RMSE	$\delta 1$	2D IoU	3D IoU	RMSE	$\delta 1$	2D IoU	3D IoU	RMSE	$\delta 1$
HorizonNet [37]	**92.78**	**91.28**	**0.08**	**0.99**	90.8	89.95	0.12	0.94	92.1	90.81	0.09	0.97
AtlantaNet [28]	60.65	56.58	0.52	0.71	62.14	59.4	0.52	0.76	60.98	57.4	0.52	0.72
Ours	91.95	90.41	0.09	0.98	**94.24**	**93.56**	**0.06**	**0.98**	**92.69**	**91.42**	**0.08**	**0.98**

Table 2. Quantitative evaluation of HorizonNet [37], AtlantaNet [28] and our method on Manhattan and non-Manhattan layouts. Our method and HorizonNet [37] are trained on Strd3D_non_M, and AtlantaNet [28] model is from its project page.

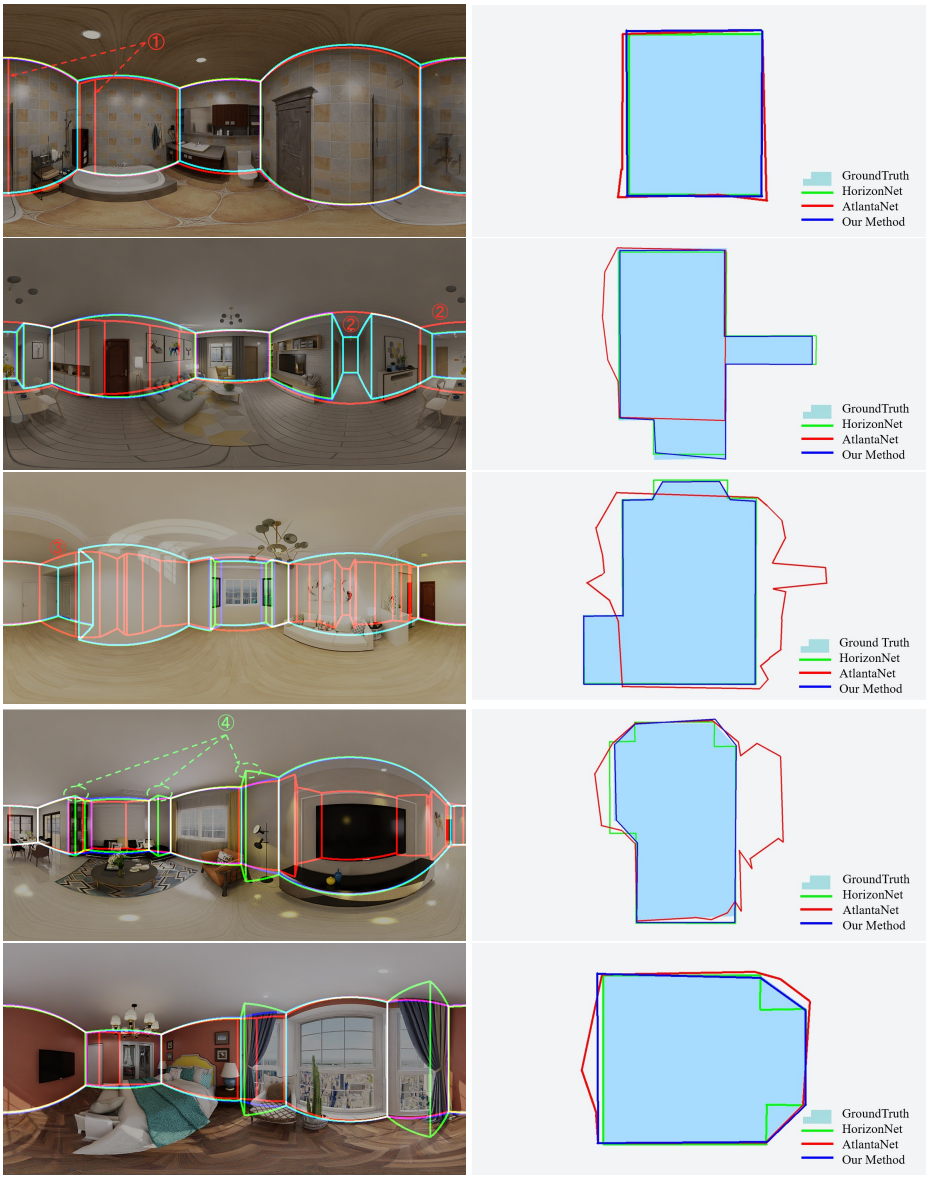

Figure 6. Qualitative comparison. The first column shows the layout recovery results of our approach (blue), HorizonNet [37] (green) and AtlantaNet [28] (red). The second column shows the comparison under the floor plan view. The typical problems of AtlantaNet [28] and HorizonNet [37] are pointed out. ①: false alarm results of the walls recovery; ②: small structures; ③: occluded areas; ④: non-Manhattan areas. Repeated problems are pointed out only once.

Besides, we compare the generalization abilities of AtlantaNet [28] and our method on the RealWorldNonManhattan dataset. The qualitative and quantitative comparisons are shown in Table 1 and Figure 5.

Basically, AtlantaNet [28] is supposed to have an advantage over us in these comparisons. Our model is trained on the synthetic dataset Strd3D_non_M. AtlantaNet's training data comes from the real world datasets (MatterportLayout [58] and AtlantaLayout [28]). For AtlantaNet, there is no domain gap between the training and evaluation data. However, from Table 1, we find that: (1) the layout recovery accuracy of our method is significant higher than AtlantaNet [28]; (2) the Junction by our method is better than AtlantaNet [28], which means there is less false alarm walls in our results. The qualitative results in Figure 5 also support that our method **has a great improvement on non-Manhattan layouts and generalizes to the real world data better.**

4.4. Generalization across Manhattan and non-Manhattan worlds

We evaluate our method, HorizonNet [37] (the best Manhattan layout recovery method) and AtlantaNet [28] on Strd3D_non_M test set, which includes both Manhattan and non-Manhattan layouts. Our method and HorizonNet [37] are trained on Strd3D_non_M, and AtlantaNet [28] model is from its project page.

We provide a quantitative comparison in Table 2. The results demonstrate that: (1) our method achieves better accuracy on non-Manhattan layouts and gets a boost of 0.59% in 2D IoU and 0.61% in 3D IoU overall, indicating that it generalizes across Manhattan and non-Manhattan worlds very well; (2) the Manhattan layout recovery accuracy of our method is aligned with the state-of-the-art HorizonNet [37], and (3) our method outperforms AtlantaNet [28] by a large margin. The qualitative comparison is shown in Figure 6. As discussed previously, two obvious problems of our competing method AtlantaNet [28] are (1) false alarm results of walls recovery, and (2) incapability of recovering small structures and occluded areas. Apparently, the Manhattan layout recovery method HorizonNet [37] is not applicable to non-Manhattan cases. **Our method is the only way to tackle these problems.**

4.5. Manhattan results

Manhattan layout recovery is a relatively simpler task than non-Manhattan. Various works concentrate on the most common cuboid layout and all of them including our method provide good results on PanoContext [51] + Stanford 2d-3d [57] dataset. The detailed performance is presented in Table.3. Our method ranks best in CE and PE, while second in 3D IoU, which shows that it is capable of handling Manhattan layouts as the state-of-the-art methods.

Method	CE (%)	PE (%)	3D Iou (%)
LayoutNet [57]	0.83	2.59	82.66
DulaNet [46]	**0.67**	2.48	**86.60**
HorizonNet [37]	0.69	2.27	82.66
Ours	**0.67**	**2.21**	82.75

Table 3. Cuboid layout performance. All methods are trained and evaluated on the PanoContext [51] + Stanford 2d-3d [57] dataset. Our method is aligned with the performance of the state-of-the-art methods.

5. Conclusions

This work introduces a room layout recovery method generalizing across Manhattan and non-Manhattan worlds. Without introducing additional supervision, we extend current Manhattan layout recovery methods to Atlanta world by predicting an extra surface normal feature, which is further used in the adaptive post-processing to handle the reconstruction of the lay-out of arbitrary shapes. Our experimental results clearly demonstrate that our method has great improvement on non-Manhattan layouts while being capable of generalizing across Manhattan and non-Manhattan layouts.

In the future, we will continue to try more advanced technologies such as Transformer [41] and semi-supervision [39], and extend our method to more complicated layouts like curved wall, curved ceiling, and even arbitrary geometric shapes.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5

[2] Dongho Choi. 3d room layout estimation beyond the manhattan world assumption. *arXiv preprint arXiv:2009.02857*, 2020. 1, 2, 3

[3] James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 941–947. IEEE, 1999. 2

[4] crs4. Atlantanet. https://github.com/crs4/AtlantaNet. 5

[5] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 616–624, 2016. 3

[6] Erick Delage, Honglak Lee, and Andrew Y Ng. A dynamic bayesian network model for autonomous 3d reconstruction

from a single indoor image. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2418–2428. IEEE, 2006. 3, 4

[7] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In *2019 International Conference on 3D Vision (3DV)*, pages 76–84. IEEE, 2019. 3

[8] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018. 3

[9] Dirk Farin, Wolfgang Effelsberg, and Peter HN de With. Floor-plan reconstruction from panoramic images. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 823–826, 2007. 1

[10] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2):1255–1262, 2020. 3, 5

[11] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cedric Demonceaux, and Jose J Guerrero. Panoroom: From the sphere to the 3d layout. *arXiv preprint arXiv:1808.09879*, 2018. 1, 2, 3

[12] Clara Fernandez-Labrador, Alejandro Perez-Yus, Gonzalo Lopez-Nicolas, and Jose J Guerrero. Layouts from panoramic images with geometry and deep learning. *IEEE Robotics and Automation Letters*, 3(4):3153–3160, 2018. 3

[13] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J Guerrero. What's in my room? object recognition on indoor panoramic images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 567–573. IEEE, 2020. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[15] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *2009 IEEE 12th international conference on computer vision*, pages 1849–1856. IEEE, 2009. 3

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[17] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 654–661. IEEE, 2005. 3

[18] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007. 3

[19] Chi-Wei Hsiao, Cheng Sun, Min Sun, and Hwann-Tzong Chen. Flat2layout: Flat representation for estimating layout of general room types. *arXiv preprint arXiv:1905.12571*, 2019. 3, 4

[20] Umit Isikdag, Sisi Zlatanova, and Jason Underwood. A bim-oriented model for supporting indoor navigation requirements. *Computers, Environment and Urban Systems*, 41:112–123, 2013. 1

[21] Zhizhong Kang, Juntao Yang, Zhou Yang, and Sai Cheng. A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information*, 9(5):330, 2020. 1

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4865–4874, 2017. 3

[24] David C Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. 2010. 3

[25] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2143. IEEE, 2009. 3

[26] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 4

[27] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 936–944, 2015. 3

[28] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. Atlantanet: Inferring the 3d indoor layout from a single 360° image beyond the manhattan world assumption. In *European Conference on Computer Vision*, pages 432–448. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[29] Giovanni Pintore, Valeria Garro, Fabio Ganovelli, Enrico Gobbetti, and Marco Agus. Omnidirectional image capture on mobile devices for fast automatic generation of 2.5 d indoor maps. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 1

[30] Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. State-of-the-art in automatic 3d reconstruction of structured indoor environments. In *Computer Graphics Forum*, volume 39, pages 667–699. Wiley Online Library, 2020. 1

[31] Srikumar Ramalingam and Matthew Brand. Lifting 3d manhattan lines from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 497–504, 2013. 3

[32] Shivansh Rao, Vikas Kumar, Daniel Kifer, C Lee Giles, and Ankur Mali. Omnilayout: Room layout reconstruction from indoor spherical panoramas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2021. 3

[33] Yuzhuo Ren, Shangwen Li, Chen Chen, and C-C Jay Kuo. A coarse-to-fine indoor layout estimation (cfile) method. In *asian conference on computer vision*, pages 36–51. Springer, 2016. 3

[34] Grant Schindler and Frank Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 2

[35] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 4

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

[37] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 1, 2, 3, 4, 5, 7, 8

[38] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. *arXiv preprint arXiv:2011.11498*, 2020. 3, 4

[39] Phi Vu Tran. Sslayout360: Semi-supervised indoor layout estimation from 360deg panorama. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15353–15362, 2021. 8

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 8

[42] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 3

[43] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12956–12965, 2021. 3

[44] Jiu Xu, Björn Stenger, Tommi Kerola, and Tony Tung. Pano2cad: Room layout from a single panorama image. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 354–362. IEEE, 2017. 1, 3

[45] 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. 1

[46] Shang Ta Yang, Fu En Wang, Chi Han Peng, Peter Wonka, Min Sun, and Hung Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 8

[47] Wenyan Yang, Yanlin Qian, Joni-Kristian Kämäräinen, Francesco Cricri, and Lixin Fan. Object detection in equirectangular panorama. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2190–2195. IEEE, 2018. 3

[48] Yang Yang, Shi Jin, Ruiyang Liu, Sing Bing Kang, and Jingyi Yu. Automatic 3d indoor scene modeling from single panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3926–3934, 2018. 3

[49] Wei Zeng, Sezer Karaoglu, and Theo Gevers. Joint 3d layout and depth prediction from a single indoor panorama image. In *European Conference on Computer Vision*, pages 666–682. Springer, 2020. 3

[50] Weidong Zhang, Wei Zhang, Kan Liu, and Jason Gu. Learning to predict high-quality edge maps for room layout estimation. *IEEE Transactions on Multimedia*, 19(5):935–943, 2016. 3

[51] 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. 1, 3, 5, 8

[52] Yupeng Zhang, Hengzhi Zhang, Daojing Li, Liyan Liu, Hong Yi, Wei Wang, Hiroshi Suitoh, and Makoto Odamaki. Toward real-world panoramic image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 628–629, 2020. 3

[53] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10–18, 2017. 3

[54] Jia Zheng. Structured3d general room layout estimation track @ eccv 2020. https://competitions.codalab.org/competitions/24183#learn_the_details-evaluation. 5

[55] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. *arXiv preprint arXiv:1908.00222*, 2(7), 2019. 5

[56] Nikolaos Zioulis, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Single-shot cuboids: Geodesics-based end-to-end manhattan aligned layout estimation from spherical panoramas. *Image and Vision Computing*, 110:104160, 2021. 3

[57] 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. 1, 2, 3, 5, 8

[58] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. 3d manhattan room layout reconstruction from a single 360 image. *arXiv preprint arXiv:1910.04099*, 2019. 3, 5, 8