

HorizonNet: Learning Room Layout with 1D Representation and Pano Stretch Data Augmentation

Cheng Sun

Chi-Wei Hsiao

Min Sun

Hwann-Tzong Chen

National Tsing Hua University

{chengsun, chiweihsiao}@gapp.nthu.edu.tw

sunmin@ee.nthu.edu.tw

htchen@cs.nthu.edu.tw



Figure 1: Some examples of 3D reconstructed room layouts by our HorizonNet.

Abstract

We present a new approach to the problem of estimating the 3D room layout from a single panoramic image. We represent room layout as three 1D vectors that encode, at each image column, the boundary positions of floor-wall and ceiling-wall, and the existence of wall-wall boundary. The proposed network, HorizonNet, trained for predicting 1D layout, outperforms previous state-of-the-art approaches. The designed post-processing procedure for recovering 3D room layouts from 1D predictions can automatically infer the room shape with low computation cost—it takes less than 20ms for a panorama image while prior works might need dozens of seconds. We also propose Pano Stretch Data Augmentation, which can diversify panorama data and be applied to other panorama-related learning tasks. Due to the limited data available for non-cuboid layout, we re-label 65 general layout from the current dataset for fine-tuning. Our approach shows good performance on general layouts by qualitative results and cross-validation.

1. Introduction

The goal of this work is to predict the room layout from a panoramic image. Most of the state-of-the-art methods

solve this problem by adopting more effective deep network architectures for their models to learn from different cues in the image. Assumptions about the room structures are often made to constrain the solution space so that the predictions of the deep model would not deviate from the common cases too much. Post-processing steps can further be performed to refine the predictions. Given a number of images with annotated layouts for training, state-of-the-art methods are able to achieve good results on the test data. However, acquiring high-quality room-layout annotations for panoramic images is labor-demanding. The annotations done by different people might be inconsistent due to ambiguities about the locations of wall boundaries, especially for well-decorated rooms. Moreover, currently available datasets do not include more images of complex room layouts. The annotation for a complex layout would just be approximated as a cuboid-shaped or L-shaped layout, introducing even more ambiguities for training and testing.

Two important and correlated issues may be further addressed for improving state-of-the-art methods. The first issue is the lack of more training and validation data with precise annotations. The second issue is that, without more annotated data for training, the deep networks cannot be too large, otherwise the test accuracy might be low due to over-fitting. Collecting more data to train a more sophisticated

model is indeed beneficial and doable, but a more efficient way to improve the performance should also be welcome. We argue that, if we have some better understanding of the problem and make good use of domain knowledge, we may improve the performance without acquiring a lot more annotated data or using a larger deep network. Data augmentation is a common procedure in deep learning to generate more data for training. Standard data augmentation heuristics such as random cropping or luminance change for image classification or object detection might not be effective for layout prediction. Our idea is to take account of the underlying geometric constraints and design a better data augmentation mechanism specifically for training layout-predicting deep networks. On the other hand, instead of increasing the model complexity, we aim to enhance the model by devising a compact representation with respect to the geometric constraints. We can, therefore, remove redundant degrees of freedom and force the model to focus more on learning critical properties for layout prediction.

We characterize our contributions as follows:

- We introduce a 1D $\mathcal{O}(W)$ representation that encodes the whole-room layout for a panoramic scene. Training with such a representation allows our method to outperform previous state-of-the-art results, yet requires fewer parameters and less computation time.
- We propose a data augmentation mechanism called *Pano Stretch Data Augmentation*, which generates panorama images on the fly during training and improves the accuracy under all settings in our experiments. This data augmentation mechanism also has the potential for boosting other tasks (*e.g.*, semantic segmentation, object detection) that directly work on a panorama.
- We show that leveraging RNNs in a layout prediction task is helpful for improving the accuracy. RNNs are able to capture the long-range geometric pattern of room layouts.
- Owing to the 1D representation and our efficient post-processing procedure, the computation cost of our model is very low, and the model can be easily extended to handle complex scenes with layouts other than cuboid-shaped or L-shaped.

Code and data are available at: <https://sunset1995.github.io/HorizonNet/>.

2. Related Work

Room layout estimation from a single-view RGB image is an active research topic over the past decade. Many approaches have been developed in this field. Most of them

exploit the Manhattan world assumption that the room layouts, and even the furniture, are aligned with the three principal axes [3]. The Manhattan world assumption imposes constraints on the layout estimation problem, and, based on the assumption, the Manhattan aligned vanishing points could also be used to rectify the image and extract features for inferring the layout.

Delage *et al.* [6] train a dynamic Bayesian network to recognize the floor-wall boundary in each column of the perspective image. Many approaches search the Manhattan aligned layout based on extracted geometric cues. Lee *et al.* [18] test the hypothesis using Orientation Map (OM) while Hedau *et al.* [12] using Geometric Context (GC) [14]. Hedau *et al.* [10] further jointly inference the room layout with 3D objects, *e.g.* beds. Similar strategies have also been used by later methods, such as introducing an improved scoring function [26, 27], generating layout hypothesis with Manhattan junction [22], and modeling the interaction between objects and layout [5, 10, 34].

The aforementioned methods only deal with perspective images. Zhang *et al.* [32] propose to estimate the layout from a 360° H-FOV panoramic image. They extend the previous methods of vanishing point detection, hypothesis generation, and scoring hypotheses based on OM, GC and object interaction, and apply all of them to panoramas. Xu *et al.* [28] also use the OM, GC, object detection, and object orientation to reconstruct 3D layout. Yang *et al.* [29] use superpixels and Manhattan aligned line segments as features, and formulate the problem by constraint graphs. The method of [31] follows a similar approach using more geometric and semantic features. Other approaches attempt to recover the floor plan from a panorama using image gradient cues [21] or from multiple panorama images [2].

Recent methods rely more on deep networks to improve layout estimation. Most of them leverage dense prediction models to classify geometric or semantic label for each pixel. For perspective images, common ways are to predict the boundary probability map [19, 23], classes of boundaries [33, 23], classes of layout surface [4, 15], and corner keypoints heatmaps [17]. The predicted dense maps can be post-processed to generate layouts. A few deep learning methods have been developed for panorama-based layout estimation. Zou *et al.* [35] predict the corner probability map and boundary map directly from a panorama. They also extend Stanford 2D-3D dataset [1] with annotated layouts for training and evaluation. Fernandez-Labrador *et al.* [9] train the deep network on perspective images. During testing, they stitch the predicted perspective boundary maps into a panorama and combine them with geometric cues to infer the layout. Two concurrent works DuLa-Net [30] and CFL [8] show improved quantitative results with the ability to produce general room shape not limited to cuboid shape. DuLa-Net [30] combines the surface semantic mask from

conventional equirectangular view and the projected floor and ceiling view. CFL [8] proposes convolution kernel specialized for equirectangular image.

Unlike all the existing methods that use neural networks to perform dense prediction for layout estimation, we leverage the property of aligned panorama image to predict the positions of floor-wall and ceiling-wall boundaries, as well as the existence of wall-wall boundary for each column of an equirectangular image. Our model only produces three values for each column of an image, and thus the output size of the model is reduced from $\mathcal{O}(HW)$ to $\mathcal{O}(W)$. The proposed output representation is similar to [6] but they only predict floor-wall boundary for each column of a perspective image using a Dynamic Bayesian Network. In contrast, our work can handle panoramas and recognize floor-wall, ceiling-wall and wall-wall boundaries using a deep neural network. Existing works [35, 9, 30, 8] on the same task learn to make dense $\mathcal{O}(HW)$ predictions over the entire image while our model predicts only three values for each image column. RoomNet [17] imitates RNN’s recurrent structure with “time steps” equal to refinement steps. We use RNN where each “time step” is responsible for estimating the result across a few image columns.

3. Approach

The goal of our approach is to estimate Manhattan room layout from a panoramic image that covers 360° H-FOV. Unlike conventional dense prediction (target output size = $\mathcal{O}(HW)$) for layout estimation using deep learning [4, 9, 7, 15, 19, 23, 33], we formulate the problem as regressing the boundaries and classifying the corner for each column of image (target output size = $\mathcal{O}(W)$). The proposed HorizonNet trained for predicting the $\mathcal{O}(W)$ target is presented in Sec. 3.1. In Sec. 3.2, we introduce a simple yet fast and effective post-processing procedure to derive the layout from output of HorizonNet. Finally in Sec. 3.3, we introduce *Pano Stretch Data Augmentation* which effectively augments the training data on-the-fly by stretching the image and ground-truth layout along x or z axis (Fig. 5).

All training and test images are pre-processed by the panoramic image alignment algorithm mentioned in [35]. Our approach exploits the properties of the aligned panoramas that the wall-wall boundaries are vertical lines under equirectangular projection. Therefore, we can use only one value to indicate the column position of wall-wall boundary instead of two (each for a boundary endpoint).

3.1. HorizonNet

Fig. 2 shows an overview of our network, which comprises a feature extractor and a recurrent neural network. The network takes a single panorama image with the dimension of $3 \times 512 \times 1024$ (channel, height, width) as input.

1D Layout Representation: The size of network output is $3 \times 1 \times 1024$. As illustrated in Fig. 3, two of the three output channels represent the ceiling-wall (y_c) and the floor-wall (y_f) boundary position of each image column, and the other one (y_w) represents the existence of wall-wall boundary (*i.e.* corner). The values of y_c and y_f are normalized to $[-\pi/2, \pi/2]$. Since defining y_w as a binary-valued vector with 0/1 labels would make it too sparse to detect (only 4 out of 1024 non-zero values for simple cuboid layout), we set $y_w(i) = c^{dx}$ where i indicates the i th column, dx is the distance from the i th column to the nearest column where wall-wall boundary exists, and c is a constant. To check the robustness of our method against the choice of c , we have tried 0.6, 0.8, 0.9, 0.96, 0.99 and get similar results. Therefore, we stick to $c = 0.96$ for all the experiments. One benefit of using 1D representation is that it is less affected by zero dominant backgrounds. 2D whole-image representations of boundaries and corners would result in 95% zero values even after smoothing [35]. Our 1D boundaries representation introduces no zero backgrounds because the prediction for each component of y_c or y_f is simply a real-valued regression to the ground truth. The 1D wall-wall (corners) representation also changes the peak-background ratio of ground truth from $\frac{2N}{512 \cdot 1024}$ to $\frac{N}{1024}$ where N is the number of wall-wall corners. Therefore, the 1D wall-wall representation is also less affected by zero-dominated background. In addition, computation of 1D compact output is more efficient compared to 2D whole-image output. As depicted in Sec. 3.2, recovering the layout from our three 1D representations is simple, fast, and effective.

Feature Extractor: We adopt ResNet-50 [11] as our feature extractor. The output of each block of ResNet-50 has half spatial resolution compared to that of the previous block. To capture both low-level and high-level features, each block of the ResNet-50 contains a sequence of convolution layers in which the number of channels and the height is reduced by a factor of 8 ($= 2 \times 2 \times 2$) and 16 ($= 4 \times 2 \times 2$), respectively. More specifically, each block contains three convolution layers with $4 \times 1, 2 \times 1, 2 \times 1$ kernel size and stride, and the number of channels after each Conv is reduced by a factor of 2. All the extracted features from each layer are upsampled to the same width 256 (a quarter of input image width) and reshaped to the same height. The final concatenated feature map is of size $1024 \times 1 \times 256$. The activation function after each Conv is ReLU except the final layer in which we use Sigmoid for y_w and an identity function for y_c, y_f . We have tried various settings for the feature extractor, including deeper ResNet-101, different designs of the convolution layers after each ResNet block, and upsampling to the image width 1024, and find that the results are similar. Therefore, we stick to the simpler and computationally efficient setting.

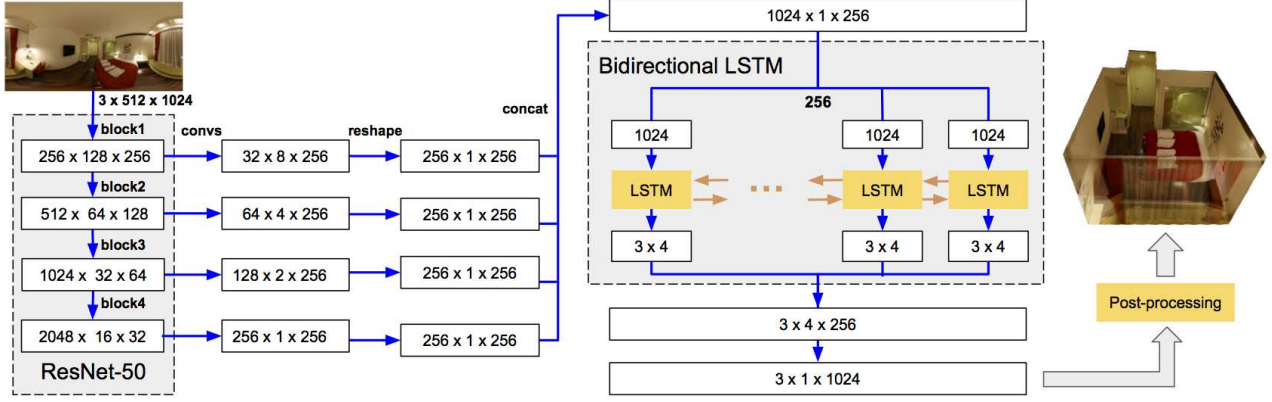


Figure 2: An illustration of the HorizonNet architecture.



Figure 3: Visualization of our 1D ground truth representations. y_w denotes the existence probability of wall-wall boundary. y_c, y_f (plotted in green and blue) denote the positions of the ceiling-wall boundary and floor-wall boundary respectively. For better visualization, we plot y_w, y_c, y_f with line width greater than one pixel.

Recurrent Neural Network for Capturing Global Information:

Recurrent neural networks (RNNs) are capable of learning patterns and long-term dependencies from sequential data. Geometrically speaking, any corner of a room can be roughly inferred from the positions of other corners; therefore, we use the capability of RNN to capture global information and long-term dependencies. Intuitively, because LSTM [13], a type of RNN architecture, stores information about its prediction for other regions in the cell state, it has the ability to predict for occluded area accurately based on the geometric patterns of the entire room. In our model, RNN is used to predict y'_c, y'_f, y'_w column by column. That is, the sequence length of RNN is proportional to the image width. In our experiment, RNN predicts for four columns instead of one column per time step, which requires less computational time without loss of accuracy. As the y_c, y_f, y_w of a column is related to both its left and right neighbors, we adopt the bidirectional RNN [25] to capture the information from both sides. Fig. 7 and Table 1 demon-

strate the difference between models with or without RNN.

3.2. Post-processing

We recover general room layouts that are not limited to cuboid under following assumptions: *i*) intersecting walls are perpendicular to each other (Manhattan world assumption); *ii*) all rooms have the one-floor-one-ceiling layout where floor and ceiling are parallel to each other; *iii*) camera height is 1.6 meters following [32]; *iv*) the pre-processing step correctly align the floor orthogonal to y-axis.

As described in Sec. 3.1, raw outputs of our deep model $y'_f, y'_c, y'_w \in \mathcal{R}^{1024}$ contain the layout information for each image column. Each value in y'_f and y'_c is the position of floor-wall boundary and ceiling-wall boundary at the corresponding image column. y'_w represents the probability of wall-wall existence of each image column.

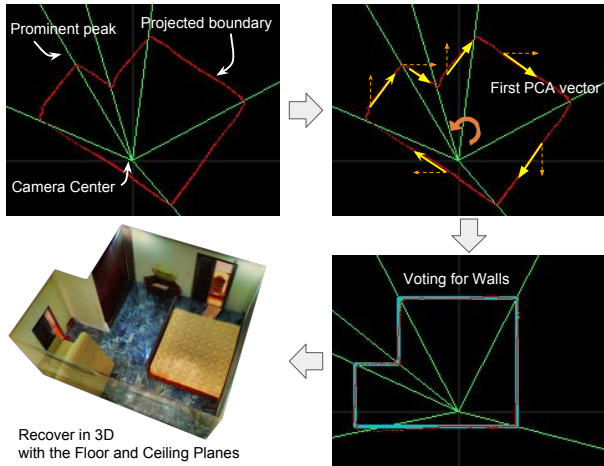
Recovering the Floor and Ceiling Planes: For each column of the image, we can use the corresponding values in y'_f, y'_c to vote for the ceiling-floor distance. Based on the assumed camera height, we can project the floor-wall boundary y'_f from image to 3D XYZ position (they all shared the same Y). The ceiling-wall boundary y'_c shares the same 3D X, Z position with the y'_f on the same image column, and therefore the distance between floor and ceiling can be calculated. We take the average of results calculated from all image columns as the final floor-ceiling distance.

Recovering Wall Planes: We first find the prominent peaks on the estimated wall-wall probability y'_w with two criteria: *i*) the signal should be larger than any other signal within 5° H-FOV, and *ii*) the signal should be larger than 0.05.

Fig. 4a shows the projected y'_c (red points) on ceiling plane. The green lines are the detected prominent peaks which split the ceiling-wall boundary (red points) into multiple parts. To handle possibly failed horizontal alignment in the pre-processing step, we calculate the first principal component of each part, then rotate the scene by the aver-

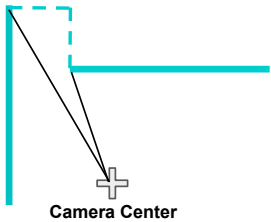
age angle of all first principal components (top right figure in Fig. 4a). So now we have two types of walls: *i*) X-axis orthogonal walls and *ii*) Z-axis orthogonal walls. We construct the walls from low to high variance suggested by the first principal component. Adjacency walls are forced to be orthogonal to each other, thus only walls whose two adjacent walls are not yet constructed have the freedom to decide the orthogonal type. We use a simple voting strategy: each projected red point votes for all planes within 0.16 meters (bottom right figure in Fig. 4a). The most voted plane is selected. Two special cases are depicted in Fig 4b which occur when the two adjacency walls are already constructed and they are orthogonal to each other. Finally, the XYZ positions of all corners are decided according to the intersection of three adjacent Manhattan junction planes.

The time complexity of our post-processing procedure is $\mathcal{O}(W)$, where W is the image width. Thus the post-processing can be efficiently done; in average, it takes less than 20ms to finish.

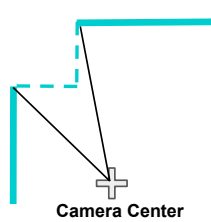


(a) Depicting how we recover the wall planes from our model output.

Occluded Corner



False Negative



(b) Two special cases: Instead of voting for a wall, we add a corner according to the two prominent peaks and the positions of two walls.

Figure 4: Visualization of wall planes recovering. Fig. 4a is an example that the pre-processing algorithm fails to correctly align the horizontal rotation of panorama.

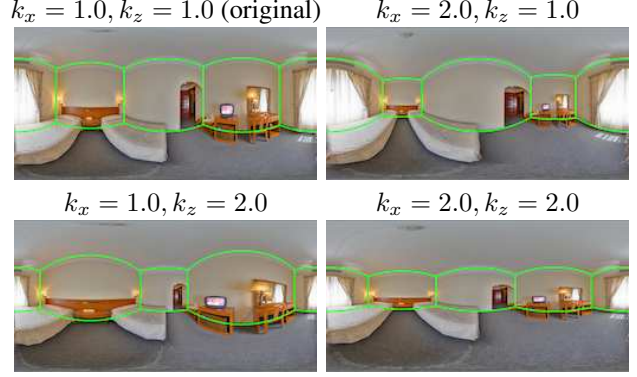


Figure 5: Visualization of the proposed *Pano Stretch Data Augmentation*. The image and ground-truth layout (green lines) are stretched along x or z axis (the effect of scaling y can be covered by x and z). This can augment the data by changing the room's length and width. This augmentation strategy improves our quantitative results under all experiment settings (Table 3).

3.3. Pano Stretch Data Augmentation

For a 360° H-FOV panoramic image, we propose to stretch along axes in 3D space to augment training data. To achieve this goal, we first represent each pixel under UV space as (u, v) where $u \in [-\pi, \pi]$, $v \in [-\pi/2, \pi/2]$. The coordinate (u, v) can be easily computed as the column and row of an equirectangular image, subject to a rotation angle of the camera. Here we introduce an additional variable d , which denotes the depth of a pixel. We will show that d can be eliminated later so our final equation does not depend on it.

We project the pixels to 3D space and multiply their x, y, z by k_x, k_y, k_z . The equation of stretched x', y', z' are shown in Eq. 1.

$$\begin{cases} x' = k_x \cdot x = k_x \cdot d \cdot \cos(v) \cdot \cos(u); \\ y' = k_y \cdot y = k_y \cdot d \cdot \sin(v); \\ z' = k_z \cdot z = k_z \cdot d \cdot \cos(v) \cdot \sin(u). \end{cases} \quad (1)$$

We can then project the stretched points back to the sphere by Eq. 2 for further equirectangular projection. atan2 in the equation is 2-argument arctangent. The depth d is eliminated since it exists in both terms of atan2 . We fix $k_y = 1$ because setting k_y to a value other than one is equivalent to multiplying k_x, k_z by the same value.

$$\begin{cases} u' = \text{atan2}(k_z \cdot \sin(u), k_x \cdot \cos(u)); \\ v' = \text{atan2}(k_y \cdot \sin(v), \sqrt{k_x^2 \cos^2(u) + k_z^2 \sin^2(u)} \cdot \cos(v)). \end{cases} \quad (2)$$

In our implementation, we do the inverse mapping by Eq. 3. For each pixel in the target image, we compute

the corresponding coordinate and sample its value from the source image via bilinear interpolation. Fig. 5 shows a visualization sample.

$$\begin{cases} u = \text{atan2}(k_x \cdot \sin(u'), k_z \cdot \cos(u')); \\ v = \arctan(k_z \cdot \tan(v') \cdot \csc(u') \cdot \sin(u)). \end{cases} \quad (3)$$

Note that our Pano Stretch Data Augmentation procedure could also be used on other tasks (*e.g.*, ground-truth map of semantic segmentation, bounding box for object detection) that directly work on panoramas. The augmentation procedure has the potential to boost the accuracy of those tasks.

4. Experiments

4.1. Datasets

We train and evaluate our model using the same dataset as LayoutNet [35]. The dataset consists of PanoContext dataset [32] and the extended Stanford 2D-3D dataset [1] annotated by [35]. To train our model, we generate $3 \times 1 \times 1024$ ground truth from the annotation. We follow the same training/validation/test split of LayoutNet.

4.2. Training Details

The Adam optimizer [16] is employed to train the network for 300 epochs with batch size 24 and learning rate 0.0003. The L1 Loss is used for the ceiling-wall boundary (y_c) and floor-wall boundary (y_f). The Binary Cross-Entropy Loss is used for the wall-wall corner (y_w). The network is implemented in PyTorch [20]. It takes four hours to finish the training on three NVIDIA GTX 1080 Ti GPUs.

The data augmentation techniques we adopt include standard left-right flipping, panoramic horizontal rotation, and luminance change. Moreover, we exploit the proposed Pano Stretch Data Augmentation (Sec. 3.3) during training. The stretching factors k_x, k_z are sampled from uniform distribution $U[1, 2]$, and then take the reciprocals of sampled values with probability 0.5. The process time of Pano Stretch Data Augmentation is roughly 130ms per 512×1024 RGB image. Therefore, it is feasible to be applied on-the-fly during training.

4.3. Cuboid Room Results

We generate cuboid room by only selecting the four most prominent peaks in the post-processing step (Sec. 3.2).

Quantitative Results: Our approach is evaluated on three standard metrics: *i)* **3D IoU**: intersection over union between 3D layout constructed from our prediction and the ground truth; *ii)* **Corner Error**: average Euclidean distance between predicted corners and ground-truth corners (normalized by image diagonal length); *iii)* **Pixel Error**: pixel-wise error between predicted surface classes and ground-truth surface classes.

Method	3D IoU(%)	Corner error(%)	Pixel error(%)
Train on PanoContext dataset			
PanoContext [32]	67.23	1.60	4.55
LayoutNet [35]	74.48	1.06	3.34
DuLa-Net [30]	77.42	-	-
CFL [8]	78.79	0.79	2.49
ours	82.17	0.76	2.20
Train on PanoContext + Stnfd.2D3D datasets			
LayoutNet [35]	75.12	1.02	3.18
ours	84.23	0.69	1.90

Table 1. Quantitative results of cuboid layout estimation evaluated on the PanoContext [32] dataset. Our method outperforms all existing methods under all settings.

The quantitative results of different training and testing settings are summarized in Table 1 and Table 2. To clarify the difference, the input resolution of DuLa-Net [30] and CFL [8] are 256×512 while LayoutNet [35] and ours are 512×1024 . Other than conventional augmentation technique, CFL [8] is trained with Random Erasing while ours is trained with the proposed Pano Stretch. DuLa-Net [30] did not report corner errors and pixel errors. Our approach achieves state-of-the-art performance and outperforms existing methods under all settings.

Qualitative Results: The qualitative results are shown in Fig. 6. We present the results from the best to the worst based on their corner errors. Please see more results in the supplemental materials.

Computation time: The 1D layout representation is easy to compute. Forward passing a single 512×1024 RGB image takes 8ms and 50ms for our HorizonNet with and without RNN respectively. The post-processing step for extracting layout from our 1D representation takes only 12ms. We evaluate the result on a single NVIDIA Titan X GPU and an Intel i7-5820K 3.30GHz CPU. The reported execution time is averaged across all the testing data.

4.4. Ablation Study

Ablation experiments are presented in Table 3. We report the result averaged across all the testing instances. For a fair comparison, we also experiment with dense $\mathcal{O}(HW)$ prediction following LayoutNet [35] but replace the U-Net [24] with the same backbone as our architecture.¹ The results of this setting are presented in the first two rows. We do not try dense $\mathcal{O}(HW)$ output with RNN since it would consume

¹To output dense (full-image) probability map, we change the Conv layer after each ResNet block from reducing both height and channels to reducing only channels, and then upsample to the same spatial dimension as the input image. Finally, the processed features of four blocks are concatenated and passed through a Conv layer to generate the final result.

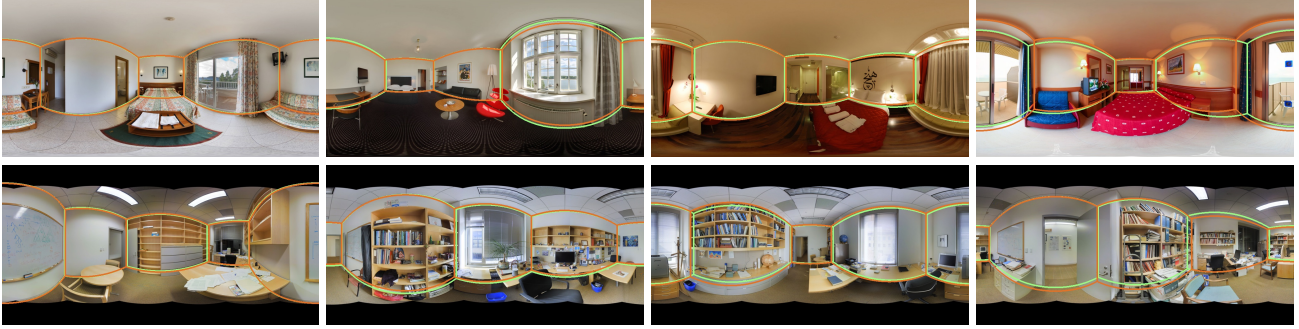


Figure 6: Qualitative results of cuboid layout estimation. The results are separately sampled from four groups that comprise results with the best 0–25%, 25–50%, 50–75% and 75–100% corner errors (displayed from the first to the fourth columns). The green lines are ground truth layout while the orange lines are estimated. The images in the first row are from PanoContext dataset [32] while second row are from Stanford 2D-3D dataset [1].

Method	3D IoU(%)	Corner error(%)	Pixel error(%)
Train on PanoContext dataset			
CFL [8]	65.13	1.44	4.75
ours	75.57	0.94	3.18
Train on Stnfd.2D3D dataset			
LayoutNet [35]	76.33	1.04	2.70
DuLa-Net [30]	79.36	-	-
ours	79.79	0.71	2.39
Train on PanoContext + Stnfd.2D3D datasets			
LayoutNet [35]	77.51	0.92	2.42
ours	83.51	0.62	1.97

Table 2. Quantitative results of cuboid layout estimation evaluated on the Stanford-2D3D [1] dataset. Our method outperforms all existing methods under all settings.

too many computing resources. We can see that learning on our 1D $\mathcal{O}(W)$ layout representation is better than conventional dense $\mathcal{O}(HW)$ layout representation.

We observe that training with the proposed Pano Stretch Data Augmentation can always boost the performance. Note that the proposed data augmentation method can also be adopted in other tasks on panoramas and has the potential to increase their accuracy as well. See supplemental material for the experiment using Pano Stretch Data Augmentation on semantic segmentation task.

For the rows where RNN columns are unchecked, the RNN components shown in Fig 2 are replaced by fully connected layers. Our experiments show that using RNN in network architecture also improves performance. Fig. 7 shows some representative results with and without RNN. The raw output of the model with RNN is highly consistent with the Manhattan world even without post-processing, which

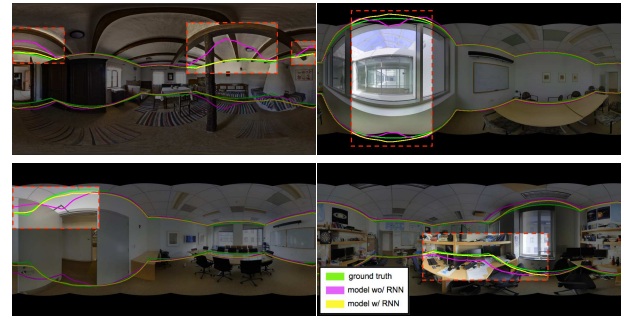


Figure 7: Visualization of model outputs with and without RNN. We plot the ground truth (green), outputs of the model with RNN (yellow), and outputs of the model without RNN (magenta). Both predictions are raw network outputs without post-processing. The model with RNN performs better than the model without RNN in images contain ceiling beam, black missing polar region caused by smaller camera V-FOV, and occluded area.

demonstrates the ability of RNN to capture the geometric pattern of the entire room.

4.5. Non-cuboid Room Results

Since the non-cuboid rooms in PanoContext and Stanford 2D-3D dataset are labeled as cuboids, our model is never trained to recognize non-cuboid layouts and concave corners. This bias makes our model tend to predict complex-shaped rooms as cuboids. To estimate general room layouts, we re-label 65 rooms from the training split to fine-tune our trained model. We fine-tune our model for 300 epochs with learning rate $5e-5$ and batch size 2.

To quantitatively evaluate the fine-tuning result on general-shaped rooms, we use 13-fold cross validation on the 65 re-annotated non-cuboid data. The results are sum-

Output Shape	Stretch Aug.	RNN	3D IoU(%)	Corner error(%)	Pixel error(%)	#params	FPS
dense $\mathcal{O}(HW)$			77.87	1.02	2.73	67M	98
dense $\mathcal{O}(HW)$	V		79.64	0.74	2.39	67M	98
our $\mathcal{O}(W)$			80.65	0.80	2.43	25M	119
our $\mathcal{O}(W)$	V		81.22	0.71	2.28	25M	119
our $\mathcal{O}(W)$		V	81.23	0.72	2.20	57M	20
our $\mathcal{O}(W)$	V	V	83.74	0.65	1.95	57M	20

Table 3. Ablation study demonstrates the effectiveness of each component in our approach. We show that all of our proposed designs can improve the quantitative result. Besides, our proposed 1D layout representation significantly reduces the number of parameters. FPS is measured for forward-pass of a $3 \times 512 \times 1024$ image on an NVIDIA TITAN X GPU.



Figure 8: Qualitative results of non-cuboid layout estimation. The occluded walls are filled with black. The blue lines in the equirectangular images are the estimated room layout boundary.

marized in Table 4. We depict some examples of reconstructed non-cuboid layouts from the testing and validation splits in Fig.1 and Fig.8. See supplemental material for more reconstructed layouts. The results show that our approach can work well on general room layout even with corners occluded by other walls.

Method	Finetuning	3D IoU(%)
LayoutNet		74.1
LayoutNet	V	75.1
ours		77.4
ours	V	82.5

Table 4. Quantitative results on the 65 re-annotated non-cuboid datas. The result of fine-tuning is evaluated by 13-fold validation.

5. Conclusion

We have presented a new 1D representation for the task of estimating room layout from a panorama. The proposed HorizonNet trained with such 1D representation outperforms previous state-of-the-art methods and requires fewer computation resources. Our post-processing method which recovers 3D layout from the model output is fast and effective, and it also works for complex room layouts even with occluded corners. The proposed Pano Stretch Data Augmentation further improves our results, and can also be applied to the training procedure of other panorama tasks for potential improvement.

Acknowledgement: This research was partially supported by iStaging and by MOST grants 106-2221-E-007-080-MY3, 107-2218-E-007-047, and 108-2634-F-001-007.

References

- [1] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, Feb. 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] James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 941–947. IEEE, 1999.
- [4] 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.
- [5] Luca Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. Understanding bayesian rooms using composite 3d object models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 153–160, 2013.
- [6] Erick Delage, Honglak Lee, and Andrew Y Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2418–2428. IEEE, 2006.
- [7] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cedric Demonceaux, and Jose J Guerrero. Panoram: From the sphere to the 3d layout. *arXiv preprint arXiv:1808.09879*, 2018.
- [8] Clara Fernandez-Labrador, Jos M Fcil, Alejandro Perez-Yus, Cdric Demonceaux, Javier Civera, and Jos J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *arXiv:1903.08094*, 2019.
- [9] Clara Fernandez-Labrador, Alejandro Perez-Yus, Gonzalo Lopez-Nicolas, and Jose J Guerrero. Layouts from panoramic images with geometry and deep learning. *arXiv preprint arXiv:1806.08294*, 2018.
- [10] Abhinav Gupta, Martial Hebert, Takeo Kanade, and David M Blei. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *Advances in neural information processing systems*, pages 1288–1296, 2010.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.
- [12] 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.
- [13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [14] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.
- [15] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *CVPR*, 2017.
- [16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [21] 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 *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.
- [22] Srikumar Ramalingam, Jaishanker K Pillai, Arpit Jain, and Yuichi Taguchi. Manhattan junction catalogue for spatial reasoning of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3065–3072, 2013.
- [23] 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.
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [25] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [26] Alexander G Schwing and Raquel Urtasun. Efficient exact inference for 3d indoor scene understanding. In *European Conference on Computer Vision*, pages 299–313. Springer, 2012.
- [27] R Urtasun, M Pollefeys, T Hazan, and AG Schwing. Efficient structured prediction for 3d indoor scene understanding. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2815–2822. IEEE, 2012.
- [28] Jiu Xu, Björn Stenger, Tommi Kerola, and Tony Tung. Pano2cad: Room layout from a single panorama image. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 354–362. IEEE, 2017.

- [29] 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.
- [30] 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. *arXiv preprint arXiv:1811.11977*, 2018.
- [31] 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.
- [32] 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.
- [33] 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. *arXiv preprint arXiv:1707.00383*, 2017.
- [34] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3119–3126, 2013.
- [35] Chuhan 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.