GEOMETRIC BASED CONVOLUTIONAL NEURAL NETWORKS FOR EFFECTIVE INDOOR SCENE LAYOUT ESTIMATION

Juntao Feng, Xuejin Chen

University of Science and Technology of China, Hefei Anhui, China.

ABSTRACT

Layout estimation from a single RGB image is a fundamental and indispensable problem for indoor scene understanding, which models the inner space as a 3D cuboid, including floor, ceiling and walls and their boundaries. However, it is significantly challenging to extract layout structure with large clutter and occlusions. In this paper, we propose a geometric based networks for a single RGB image which encodes depth and normal information from image itself. We have demonstrated that using geometric information jointly works better than using only RGB images for indoor scene layout estimation with fully convolutional neural networks(FCNN). Then an optimization framework takes full advantages of spacial labelling results and layout boundary relations from networks to generate final layout estimates. The proposed method has proven to achieve competitive accuracy of layout estimation on two commonly used benchmark datasets.

Index Terms— Scene understanding, layout estimation, geometric embedding.

1. INTRODUCTION

The main purpose of indoor scene layout estimation is to extract semantic boundaries among walls, ceiling and floor, and to obtain different planes which provide strong spacial expression of the scene, for cluttered indoor scenes from a single RGB image, as shown in Fig. 1.

(a) Input image (b) Layout estimation (c) Layout visulization

Fig. 1. Examples of indoor scene images. (a) Input images. (b) Layout estimation with semantic labels. (c) Layout visulization superimposed on the input image.

Thanks to XYZ agency for funding.

Layout estimation is a fundamental and indispensable problem for indoor scene understanding, and it plays a critical role in a diverse range of hot applications, such as scene reconstruction, robot navigation, virtual reality, and so on. Moreover, on the Large-scale Scene Understanding Challenge(LSUN), indoor scene layout estimation has become one of the popular tasks towards scene-centric challenges. However, the most important clues such as room corners and layout boundaries are often occluded by a large amount of clutter occuring everywhere in daily life(see Fig.3 (a)), which make the layout task tough to estimate. Moreover, illumination variations existed in indoor scene(see Fig.3 (b-c)), may increase the difficulty of visual understanding. Last but not the least, the views of indoor scene images are in a wide range(see Fig.3 (d-f)), and such factor can cause appearance diversity for an indoor scene image.

(a) (b) (c)

(d) (e) (f)

Fig. 2. Examples of indoor scene images. (a) Much clutter. (b) Too bright. (c) Too dark. (d-f) Different views.

(cxj: Related work: from estimation using vanishing points/low-level features, then two-step (FCN+post-processing), to end-to-end network.) In recent decades, many researchers have made massive attempts to estimate spacial layout from a single image automatically. One popular framework using a 3D cuboid to approximately express indoor scene layout was introduced by [1] in 2009. The author generated several layout candidates by using vanish point detection. Then it gathered mid-level features like the line membership features

and the geometric context features for each candidate and ranked them by using a structured SVM. Unfortunately, the process of layout candidiate generation is highly sensitive and fragile towards a large amount of clutter. Based on this milestone work, several strategies focused on layout hypothesis generation and ranking were considered. Wang et al.[2] introduced latent variables to model indoor clttuer. Moreover, in [3][4], Schwing et al. modeled cllutterd indoor scenes with higher-order potentials and jointly generated layout and objects with box shapes. Some researchers introduced low-level information as geometric restrictions as supplementary in indoor scene problem. Lee et al.[5]. proposed several physically valid structure hypotheses by geometric reasoning and verified to find the best fitting model to line segments. Ramalingam et al.[6] employed Manhattan Junction grouping to select best layout. As mentioned above, traditional methods need to design image features manually, which greatly increase the complexity and are weekly capable of adapting to dealing with all complex indoor scenes. Besides, there exists a large time consuming for layout candidate generation, which is stongly against our intention to rely on computers to understand the room layout as fastly as possible.

Since the development of the 3D-data capture devices like kinects and RGBD depth cameras, more and more 3D-based methods towards indoor scene field have been studied. Moreover, some researches using 3D-based methods for indoor layout estimation have been studied. Geiger et al.[7] proposed a high-order graphical model and jointly reasoned about the layout, objects and superpixels from RGBD image. Ren et al.[8] proposed a cloud of oriented gradient descriptor and Manhattan voxel that links the 2D appearance and 3D pose of object categories to better capture the 3D room layout geometry and object detection. Guo et al.[9] interpreted layout and 3D model jointly from a RGBD image by aligning to 3D model dataset. 3D information provide additional geometric cues based on planes and objects, which merge big planes and obejects together and weaken the effects of light and complex textures, and thus lead to robust semantic understanding. With 3D-based methods, not only can layout estimation work turn out to be more robust, but also more problems we can solve even 3D modeling problem for whole indoor scenes.

Recently, the deep learning methods and convolution neural network(CNN) have achived impressive progresses in various computer vision tasks, such as semantic segmentation [10][11], object detection [12][13], scene understanding [14][15], and so on. Towards layout estimation problem, several researchers have achieved to adopt deep learning methods to solve it. Mallya et al.[16], presented a Fully Convolutional Networks (FCNs)[10] framwork for learning informative edge maps from a single image, which provided as a new information to sample vanishing lines for layout candidates generation and ranking. This work is the first to train CNN to produce robust features used to replace hand-crafted features towards layout estimation problem. However, in the framwork of layout candidates generation and ranking, their algorithm still remained time-consuming. Dasgupta et al.[17] used the FCN to learn semantic surface labels including left wall, front wall, right wall, ceiling, and ground. Initial layout generation and optimization were all based on suface belif labels. Unfortunately, Their algorithm relies entirely on FCN's precision for semantic surface labels. In other words, if FCN result gains two much error, we may get totally wrong prediction. Moreover, optimization procedure does not utilize any restrictions based on edge information, just taking random combinations of edges instead, which is geometric inconformity compared with RGB images. Ren et al.[18] adopted a multi-task fully convolutional neural network (MFCN) to jointly predict the room edges and semantic labels. Then a coarse-to-fine method was adopted which enforced several constraints such as layout contour straightness, surface smoothness and geometric constraints to generate fine layout from coarse prediction of room edges. They have considerd geometric constraints to predict high quality estimation results. They need complex geometric computation and rules to generate useful critical lines, however, such low-level features are exactly what we avoid to extract. Zhang et al.[19] used another deconvolution network which has multi-layer deconvolution and a receptive field as large as the entire image compared to FCN. As the result, they can obtain highly reliable edge maps. Then they follow the framework of layout candidates generation using vanish line samplling with an adaptive line sample strategy for robustness and time reduction. A measurement of similarity is proposed between edge map and layout candidates for ranking. Unexpectedly, their results are less precise compared with [17][18]. One explanation may be that the latter two optimization strategies are more effective compared to vanish line sampling.

(cxj: compared to [18], what is our advantage? They apply geometric constraints as an optimization problem. [17] also apply a post-processing step to enforce geometric constraints. One big disadvantage of the post-processing is that it takes seconds to optimize the layout. [17] requires 30 seconds for the layout optimization.)
(drf: compared to [18], we have better performance on layout estimation before optimization. ie. the output of our FCN-MC are more reliable because we apply additional information(depth and normals). We use the same post-processing step in [17], so we have the same problem, or maybe worse.)

(cxj: [20] presents an end-to-end trainable network that predicts the layout corners and room type A RNNN framework is employed to refine the layout. Different from previous pixel-based representation of the layout, they use a keypoint-based representation.)

In this work, our algorithm is based on framework of [17], which does not need to extract low features of lines and vanish points from a single RGB image. To enhance the existed FCN's ability, we use networks to estimate depth and normal information from one single RGB image. Then depth

and normal information serves as geomtric embedded into FCN networks to jointly generate high quality suface maps and edge maps. Based on[17] optimization framwork, we introduce more geometric constraints from predicted edge maps to optimize surface labels to generate high quality layout estimation. Experimental results demonstrate that our method is robust and effective for layout estimation even facing a high clutter on two popular room layout benchmark datasets.

2. RELATEDWORK

Layout estimation of interior scenes has been drawn substantial attention in computer vision, especially in recent several decades. According to the methodologies used to recover the layout or structure, we mainly discuss two categories of layout estimation methods: geometric reasoning based on geometric constraints like vanishing points, and neural network-based methods in recent five years.

Most of existing approaches recover the scene structure from a single image under the Manhattan world assumption, under which most objects are aligned with three dominant orthogonal directions. Under this assumption, a three-step pipeline is widely used: vanishing points are estimated from detected line segments, layout hypotheses are created, and then each hypothesis is evaluated according various evidences in the image to find the best one. (cxj: list each method..) More detailed building models are created from line segments in [5] where they consider corner types, edge orientations, and geometric relationships in hypothesis evaluation. However, these methods usually spend several minutes on the hypothesis evaluation while there are huge number of hypotheses in a cluttered scene.

Besides of recovering the main wall-floor structure, more geometric inference of objects in the scene are added to generate more accurate estimation in cluttered scenes. [1] reply on cuboid representations to .., while [7] employs more precise 3D models to represent objects, taking the advantage of the depth information.

First, only use vanishing lines to generate hypotheses. Then, add surface orientation constraints. Finally, add analysis on the clutter objects or occluding objects in the scene.

With the development of depth camera, a large number of methods for structure recovery from RGBD images appeared. [7] jointly infers 3D objects and the scene layout from a single RGBD image via a high-order CRF model. A novel representation called Manhattan voxel is proposed in [8] to capture more detailed 3D room layout. By employing a cascade of classifiers, the contextual relationships among object categories and scene layout are considered to improve the accuracy of both object categorization and layout estimation. However, more complicated models brings higher time costs varying from 10 to 30 minutes to analyze a single RGBD image.

Deep learning methods and convolution neural network(CNN) have achieved impressive progresses in various computer vision tasks, such as semantic segmentation [10, 11], object detection [12, 13], scene understanding [14, 15], and so on. Towards the layout estimation problem, Mallya et al.[16] presented a fully convolutional networks (FCNs) framework for learning informative edge maps from a single image, to provide additional hints for sampling vanishing lines for layout candidates generation and ranking. This work firstly uses CNN to produce robust features instead of hand-crafted features for layout estimation problem. However, the layout hypothesis generation and ranking parts still play as post-processing steps, costing much time as traditional methods. Dasgupta et al. [17] use FCN to learn semantic surface labels including left wall, front wall, right wall, ceiling, and ground. Initial layout generation and optimization were all based on surface belief labels. Unfortunately, Their algorithm relies entirely on FCN's precision for semantic surface labels. Moreover, the optimization procedure does not utilize any restrictions based on edge information, just taking random combinations of edges instead, which is geometric inconformity compared with RGB images. Ren et al. [18] adopted a multi-task fully convolutional neural network (MFCN) to jointly predict the room edges and semantic labels. Then a coarse-to-fine method was adopted which enforced several constraints such as layout contour straightness, surface smoothness and geometric constraints to generate fine layout from coarse prediction of room edges. They have considered geometric constraints to predict high quality estimation results. They need complex geometric computation and rules to generate useful critical lines, however, such low-level features are exactly what we avoid to extract. Zhang et al.[19] used another deconvolution network which has multi-layer deconvolution and a receptive field as large as the entire image compared to FCN. As the result, they can obtain highly reliable edge maps. Then they follow the framework of layout candidates generation using vanishing line sampling with an adaptive line sample strategy for robustness and time reduction. A measurement of similarity is proposed between edge map and layout candidates for ranking. Unexpectedly, their results are less precise compared with [17, 18]. One explanation may be that the latter two optimization strategies are more effective compared to vanish line sampling. (cxj: in comparison, we provide ..)

3. OUR METHOD

3.1. System Overview

Under the Manhattan world assumption, a room layout is represented as cube having at most five walls (Left, Front, Right, Ceiling, Ground) visible in the image. Given an RGB image I with (cxj: arbitrary?) (drf: yes, but it will be resized to 320×240 and then input into the network [21])size $w \times h$, our

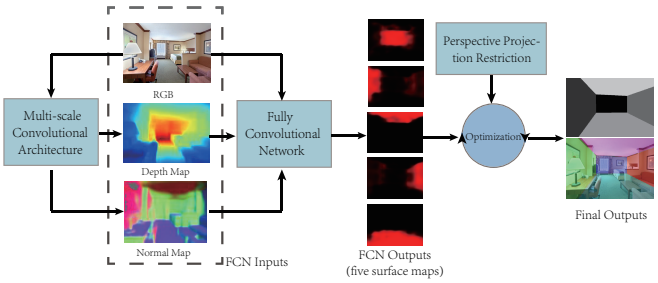


Fig. 3. An overview of our layout estimation algorithm pipeline. First we adopt a multi-scale CNN achitecture to predict geometric information from RGB image, including depth and normals. Then we encode abovementioned information into FCNN , which help to accurately estimate the layout. Optimization framwork based on perspective projection restriction is adopted to generate final precise layout estimates.

algorithm generates a room layout **L** consisting of a surface label for each pixel $L_{ij} \in$ {left wall, front wall, right wall, ceiling, ground}. Fig. 3 shows our algorithm pipeline. Different from [17], we first estimate the depth depth D_I and normal map N_I from the input color image to generate *geometric hints* using a multi-scale convolutional architecture [21], as described in Sec. 3.2. Integrating the original RGB image, the estimated depth and the normal map, a fully convolutional network is used to predict five surface maps, each of which describes the belief for each specific layout surface. Details will be described in Sec. 3.3. To generate more clear and straight boundaries in the final layout, an optimization step is adopted to (cxj: what does this optimzation do?)(drf: The output of the FCN-MC maybe not consistent with the model that we use to parameterize the layout. For example, the boundaries are not straight, there may be multiple disjoint components per label, and it may contain spurious regions like spurious front wall. The optimization do some preprocessing first to prune the extra disjoint components and fill in the hole caused by pruning and judge whether the conbination of the plane is reasonable. Then apply an iterative refinement process to obtain L. So the ultimate goal of the optimization step is to get L which parameterize the straight boundaries of layout.), as described in Sec. **??**.

(cxj: several questions here)

1. **input image size**: as claimed in [8], the receptive field of VGG16 is 404x404, why do we use 321 $\times$ 321?
   (drf: Need to do experiment. As has been revealed by [8], if the input image size is smaller than the receptive field size, it is padded with zeros and spatial resolution is lost in this case. So 321$\times$321 may be not appropriate, we should change it to 404 $\times$ 404)

(cxj: Figure 3 is very similar with [17]..)

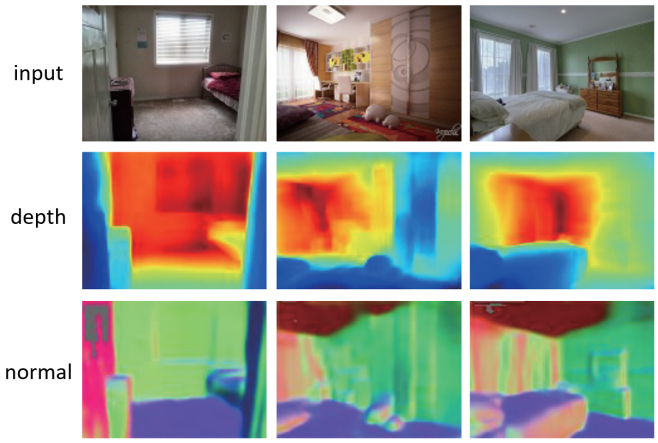


Fig. 4. Estimation of depth and normal from a single RGB image using the multi-task FCN in [21].

3.2. Geometric Fusion FCNN for Coarse Layout Estimation

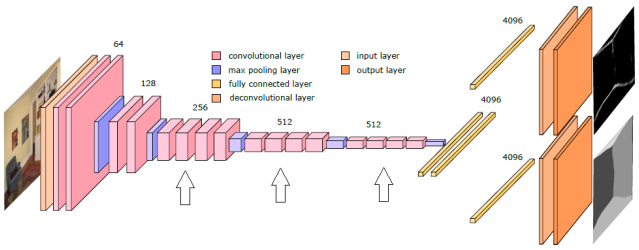
We use the multi-scale convolutional network proposed in [21] to estimate the depth and normal map from a single RGB image, as Figure 4 shows. (cxj: More analysis on the results of depth map and normal map.)

Though the predicted depth and normal are not accurate, they provide valuable 3D information for high-level structure estimation, especially in cluttered scenes. Normals can serve as clues tending to merge big planes together, with interference factors like clutter, textures and illumination eliminated to varying degrees. These mid-level geometric information can be used to adapt and improve performance for layout estimation compared to using RGB only.

3.3. Surface Label Prediction using MFCN

In previous work, [17, 18] achieved to use fully convolutional neural network(FCNN) or multi-task fully convolutional neural network (MFCNN) to predict coarse semantic layout surfaces and layout edges. However, due to much clutter, complex textures and illumination variations existed, semantic surfaces like walls are visually separated in to pieces, making whole surfaces difficult to aggregate together. Fig. 8(b) shows layout estimation results using FCNN architecture for prediction using a general FCN widely used in previous methods [17, 18]. Under the comparison between the predicted results and the ground truth, we can see that due to the environmental effect of indoor scenes, layout results estimated from FCNN are not reliable. When there exits clutter right on the boundaries, such critical clues are partially or entirely excluded, and thus we can not tell location of each plane precisely. Moreover, an entire plane may be predicted separately into pieces due to clutter lay in the plane.

Network Architecture (cxj: Explain the network archi-



(drf:

Fig. 5. The network architecture of [18], illustration of the FCN-VGG16 with two output branches. We only use the second branch to learn semantic surface and prune the first branch. By the way, the number of convolution layers in this figure may be wrong, i check the prototxt of VGG16 and FCN-32s, there should be three layers in the part which are pointed by black arrow.

)

tecture.) (drf: We use the network architechure proposed by [18] but with only one output branch while [18] is a multi-task FCN and learn both layout and semantic surface. The architechure is based on VGG16, in Fig. 3.3 We remove the first branch after fc7 and in turn we don't have to use a joint loss. The network is trained specifically for semantic surface segmentation. And we fuse the depth map and the normal map with rgb image as additional geometric information in the network to improve the performance of segmentation.)

Given the input RGB image and the estimated geometric information including depth and normal, there are several possible ways to integrate them in a network to predict the surface labels. We introduce two architectures here. We also resize them to 321 as input to the following FCN.

The first way is treading the depth and normal map as four additional channels associated with the input RGB image. Given the seven-channel input (3 channels from RGB, 1 channel from depths and 3 channels from normals), we train a VGG16 FCN to predict the five belief maps for the five surfaces in a room. The network architecture is shown in Fig. 6. This architecture is named as FCN-MC in this paper. However, it is intuitively seen that different channels have a wide range of variance and they probably can not be directly fused at the beginning convolutional layers.

The second way to integrate them is to later fuse the features that are extracted from different channels separately with a number of convolutional layers. This network is named as FCN-GF (geometric fusion), whose architecture is shown in Fig. 7.

For both networks to predict the five surface maps, the loss function is defined as the softmax ...(cxj: check wiht Fengjuntao. thesis page 18.) (drf: According to fengjuntao, the loss function was written by himself but probably not stan-

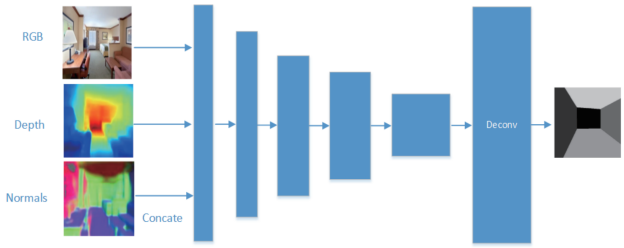


Fig. 6. A simple network architecture taking the RGB, depth and normal together as input to a VGG16 FCN. (cxj: add more details of each layer.)

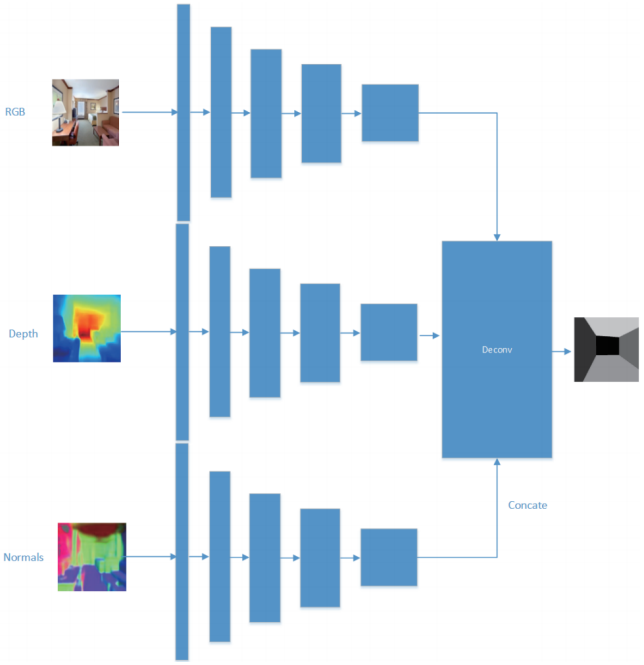
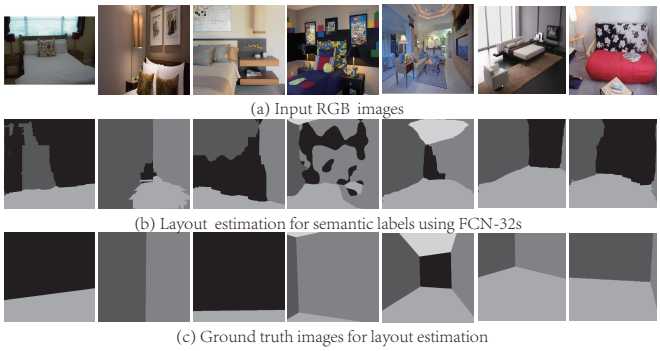


Fig. 7. A network architecture that fuses the RGB, depth and normal together later. (cxj: add more details of each layer.)

dard. I check the softmaxwithloss function on the internet and i think the formulation should be written as...or ...) (cxj: The output is five maps or a single map with five surface labels?) (drf: The output of the FCN-MC/FCN-GF is a $w \times h \times 5$ multidimensional array T, where w and h is the width and length of the input rgb image, and each of the 5 slices can be interpreted as a classification map for a specific label.
A single map with five surface labels can be obtained by simply picking the label with the highest score for each pixel among those 5 slices. and we use this single map for step3: optimization.)

Training (cxj: How do you train this network with additional input? Any option to change the network? Do you modified the network parameters? say number of neurons or layers?) (drf: For FCN-MC, depth map and normals map are merged with the rgb image as new channels and input to the FCN.)



(a) Input RGB images

(b) Layout estimation for semantic labels using FCN-32s

(c) Ground truth images for layout estimation

Fig. 8. Layout estimation results using different architectures. Row from top to bottom: (a) the input RGB image. (b)(c)(d)Surface predictions using the FCNN architecture used in [17, 18], our FCN-MC and FCN-GF respectively. (e) The ground truth.

(cxj: Result analysis for different architures.) (drf: For FCN-MC, We treat the normal and the depth as homogeneous information of the rgb image and are a supplement to the rgb image. These addtitional information can serve as context constraints to improve the performance of the network.
For FCN-GF, we treat the normal and the depth as unhomogeneous information of the rgb image and may disrupt the network parameters as they share weights with the rgb image in the following convolution layers during training. So we learn these information sepertately and fuse them in high-level layer. ie. conv5-3. (more detailed result analysis coming soon..)) From Table 1, we can see that ...(cxj: FCN-GF is better than FCN-MC?) (drf: We expect that the performance of FCN-GF should be better than FCN-MC, but it turns out that FCN-MC slightly outperforms FCN-GF. Furthermore, the structrue of FCN-MC have less paramters than FCN-GF, so it's time saving and memory saving compared to FCN-MC.

Table 1. Pixelwise accuracy for surface label prediction.

Network	Accuracy
FCN-32s	0.8109
FCN-MC	0.8392
FCN-GF	0.8350

Table 2. Performance benchmarking on Hedau's dataset

Method	Pixel Error (%)
Proposed FCN-GF	xxx

Although the result reveals FCN-MC is better, the structure of FCN-GF is still worth trying, as the depth maps used as input to the network are not standard. ie. They are encoded to 3 channels rgb images and should be decoded to gray scale image. And the batch size in training stage may be not appropriate due to hardware limitation.) Our method generate much more clear edges, less holes.

(cxj: Do test on RGBD images...)

3.4. Layout Generation

We adopt a popular model that several researchers [1, 17, 18] have been used to parameterize indoor layout based on the Manhattan world assumption. Indoor scene layout can be modeled as

$$L = (l1, l2, l3, l4, v) \tag{1}$$

where l_i stands for i^{th} vanishing line and v stands for the specific vanishing point. The whole scene is equivalent to be labeled to five semantic surfaces, coresponding to (front, left, right, ceiling, ground), as described in Fig. **??**. Based ob Eq. (**??**), each surface can be reconstructed with vanishing lines, extension lines between vanishing point and Intersection point, and image boundaries. Due to the camera pose, not five surfaces are always visible, and such layout can still be modeled by Eq. (**??**). Different examples are given in Fig. **??**.

4. RESULTS

Compared with [8], our error is 9.51, their error is 9.31. (cxj: How about time cost? Is our method faster?) (drf: Need to do some research. We use the optimization method in [17], but is realized in cpu mode. They takes approximately 30 seconds per frame on an Nvidia Titan X.)

Table 3. Performance benchmarking on the LSUN dataset

Method	Corner Error (%)	Pixel Error (%)
Hedau et al. (2009) [1]	15.48	24.23
Mallya et al. (2015) [16]	11.02	16.71
Dasgupta et al. (2016) [17]	8.20	10.63
Ren et al. (2016) [8]	7.95	9.31
Proposed FCN-GF	xxx	9.51

5. CONCLUSION

6. REFERENCES

7. REFERENCES

[1] Varsha Hedau, Derek Hoiem, and David Forsyth, "Recovering the spatial layout of cluttered rooms," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1849–1856.

[2] Huayan Wang, Stephen Gould, and Daphne Roller, "Discriminative learning with latent variables for cluttered indoor scene understanding," *Communications of the ACM*, vol. 56, no. 4, pp. 92–99, 2013.

[3] 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*. IEEE, 2012, pp. 2815–2822.

[4] Alexander G Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun, "Box in the box: Joint 3d layout and object reasoning from single images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 353–360.

[5] 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*. IEEE, 2009, pp. 2136–2143.

[6] 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*, 2013, pp. 3065–3072.

[7] Andreas Geiger and Chaohui Wang, "Joint 3d object and layout inference from a single rgb-d image," in *German Conference on Pattern Recognition*. Springer, 2015, pp. 183–195.

[8] Zhile Ren and Erik B Sudderth, "Three-dimensional object detection and layout prediction using clouds of oriented gradients," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1525–1533.

[9] Ruiqi Guo, Chuhang Zou, and Derek Hoiem, "Predicting complete 3d models of indoor scenes," *arXiv preprint arXiv:1504.02437*, 2015.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[12] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[14] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[15] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[16] Arun Mallya and Svetlana Lazebnik, "Learning informative edge maps for indoor scene layout prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 936–944.

[17] 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*, 2016, pp. 616–624.

[18] Yuzhuo Ren, Chen Chen, Shangwen Li, and C-C Jay Kuo, "A coarse-to-fine indoor layout estimation (cfile) method," *arXiv preprint arXiv:1607.00598*, 2016.

[19] Weidong Zhang, Wei Zhang, Kan Liu, and Jason Gu, "Learning to predict high-quality edge maps for room

layout estimation," *IEEE Transactions on Multimedia*, 2016.

[20] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich, "Roomnet: End-to-end room layout estimation," *CoRR*, vol. abs/1703.06241, 2017.

[21] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.