Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

# Review on Room Layout Estimation from a Single Image

Bincy P Mathew
M.Tech Image Procesing
College of Engineering Chengannur
Kerala, India

Smitha Dharan
Professor in CSE
College of Engineering Chengannur
Kerala,India

***Abstract*—Goal of room layout estimation is to predict the spatial layout of a room from a monocular image. It aims at finding the wall-floor, wall-wall, wall-ceiling boundaries from a single room image. The room layout can provide a better understanding of the 3D scenes.**

**The room layout can be integrated into mixed reality games to supply a far better immersiveness experience, or utilized in other related augmented reality applications such room redecoration. Various studies have been introduced for room layout estimation. This paper review several works which deal with room layout estimation.**

***Keywords*—*Room layout estimation, scene understanding, deep learning.)***

## I. INTRODUCTION

Room layout estimation is one of the main tasks of 3D indoor scene understanding. It is the problem of recovering the structure of an indoor scene from a single image. It generates a drawing or a digital model to scale of an existing room. Most indoor structures consist of planar surfaces. Layout estimation models the room space with a best-fit 3D box. In other words, room layout estimation extract semantic boundaries among walls, ceiling and floor from a single image. Each image pixel is classified as belonging to floor, ceiling, or wall surface. Fig.1 shows the 3D box layout of rooms. Estimation of room spatial structure can help us to easily interact with environments.

Indoor scene understanding is a main task for many real-world applications such as robotics and augmented reality [1,2]. Room layouts are used to generate floor plans or 3D Computer Aided Design (CAD) models for the building construction industry. The model is used to sketch the renovation of an apartment or to estimate the quality and the correctness of an ongoing construction with respect to the initial model. In real-estate industry, these models are used for virtual tour. Room layouts are also used for indoor navigation to help people localize themselves in large areas such as shopping malls and for robot navigation in indoor environments. It can be integrated with Augmented Reality(AR) to find AR applications such as gaming and room furniture visualizations.

However, room layout estimation in highly cluttered and occluded indoor scenes is very challenging. As indoor scenes primarily contain a lot of furniture, strong occlusions usually appear between furniture and walls. In the presence of occlusion and clutter, it is difficult to locate the room boundaries. Recent studies are trying to resolve these issues to an extent.

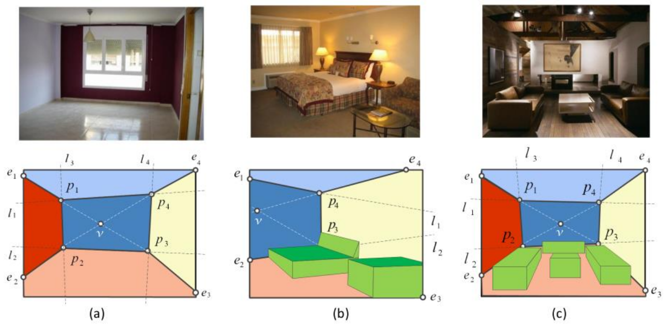

Fig: 1. 3D box layout of rooms: (a) an easy setting where all five surfaces are present; (b) a setting where some surfaces are outside the image; (c) a setting where key boundaries are occluded [12]

The paper is organized as follows. Section II reviews some methodologies of room layout estimation. Section III presents a discussion on these methods.

## II. METHODS OF ROOM LAYOUT ESTIMATION

Most of the room layout estimation methods are based on the common assumption of Manhattan worlds [3]. The Manhattan assumption is that the visual scene conforms to a 3-D grid (right-angle structure). Recently, edge maps or semantic labels learned from FCNs have become popular for this task, and their use has significantly enhanced the layout estimation performance. An edge map is a heat map that represents the boundaries of the ceiling, walls, and floor. The semantic labels are five belief maps, each of which represents the region of a semantic surface of the room (ceiling, floor, front wall, left wall, and right wall). Most of the CNN-based approaches for estimating room layout edges employ an encoder-decoder topology with a standard classification network for the encoder and utilize a series of deconvolutional layers for up sampling the feature maps.

The task of spatial layout estimation was first introduced by Hedau *et al.* [4], where the problem of recovering spatial layouts of indoor cluttered scenes from single images is addressed by modelling the scenes jointly in terms of 3D box layouts and surface labels of pixels. The 3D box layout coarsely models the space of the room as if it were empty. The surface labels provide precise positioning of visible object, wall, floor and ceiling surfaces. First, three mutually orthogonal vanishing points are estimated and a series of layout hypotheses were generated by uniformly sampling rays from the vanishing points. Then, using a learned structured regressor each layout hypothesis is assigned a score, and the layout with the highest score is selected as the result (room layout). However, here candidate 3D boxes are generated and inference is formulated in terms of single high dimensional discrete random variable.

Triplet of vanishing points corresponding to the three principal orthogonal directions of a room are estimated, which specifies the box layout orientation. Rother's algorithm [5] is modified for finding mutually orthogonal vanishing points with more robust voting and search schemes. Rother ranks all triplets employing a voting strategy, scoring angular deviation between the line and the point and using RANSAC driven search. Among the intersection points of all detected lines, triplets are selected as candidate points. Hedau used an alternate greedy strategy where the candidate point with the highest vote are selected, and then remove lines that cast high votes for this point. Quantized the remaining intersection points using variable bin sizes in the image plane. Extend the linear voting scheme utilized in [5] to a more robust exponential voting scheme. The vote of a line segment *l* for a candidate point *p* is defined as below,

$$v(l,p) = |l| * \exp - \left(\frac{\alpha}{2\sigma^2}\right) \tag{1}$$

Once the vanishing points are estimated, sample the space of translations for getting a box layout. A layout is specified by two rays through each of two vanishing points, which give four corners and edges, and the remaining edges of the box follow by casting rays through the third vanishing point and these corners. The layout generation is illustrated in Fig. 2.

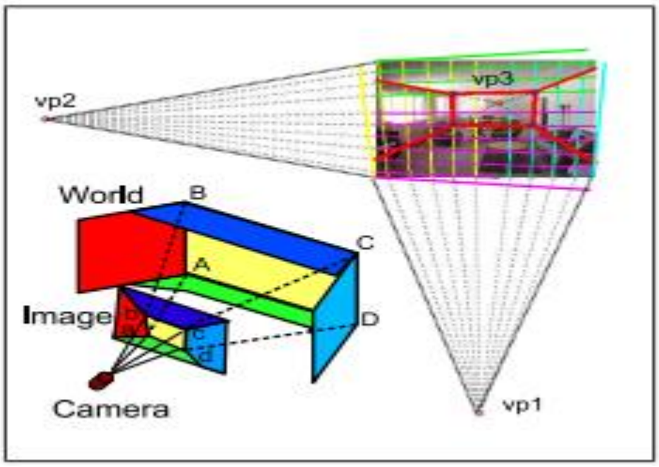

Fig: 2. layout generation method used by Hedau et al. [4]

Rank the box layouts based on how well they fit the ground truth layout. Using a learned structured regressor each layout hypothesis is assigned a score, and the layout with the highest score is selected as the result. All experiments are performed on a dataset of 308 indoor images collected from the web and from LabelMe [7].

Chao *et al.* [6] employed a method to estimate the vanishing points more accurately. Besides low level features, high level feature like human poses are utilized to estimate the vanishing points. This method is built on top of the fact that people are often the focus of indoor scenes. It exploits the 3D geometric relationships between people and room box to estimate vanishing points, camera height, and 3D locations of the people together. It enhances vanishing point estimation by considering humans in the scene. It follows the same procedure of [4] to generate layout hypothesis and identify the best-fit candidate layout.

Mallya *et al.* [8] used the FCN to predict the edge maps. The coarse output maps of FCN are fed to a deconvolutional layer to obtain dense pixel-wise outputs. An adaptive vanishing line sampling method is used to generate the layout hypotheses. Once the edge maps and vanishing points were estimated, uniformly spaced sectors are generated from horizontal vanishing point. Then the sector with highest average edge strength is sampled to get the layout hypotheses. These hypotheses were ranked by the learned edge maps along with the line membership (LM) and geometric context (GC) features. Here the layout prediction is achieved by just using edge based features which is more simpler than earlier works. The FCNN is used solely for generating an intermediate feature.

The FCN is jointly trained for prediction of informative edge map and prediction of geometric context labels. Joint training is performed by sharing all layers of FCN except for deconvolutional layers which produce the softmax probability maps for the respective types of output. The approach of [4] is used for vanishing point estimation. Once the informative edge maps and vanishing points are estimated, uniformly spaced sectors are drawn from horizontal vanishing points.

Rank all the resulting sectors by the average informative edge strength and retain top K sectors. N rays are sampled uniformly from each of the selected sectors. This strategy is illustrated in Fig.3. These hypotheses were ranked by the learned edge maps along with the line membership (LM) and geometric context (GC) features.

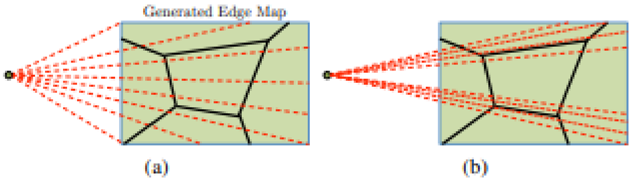

Fig: 3. Adaptive layout generation. (a) Uniformly spaced sectors originating from the horizontal vanishing point are first generated. K = 1 sectors with highest average edge strength per pixel are chosen, both above and below the horizontal line through the vanishing point. (b) N = 2 rays are sampled from each selected sector. [8]

Dasgupta *et al.* [9] used the FCN to predict the semantic labels. The belief map generated by FCN are dense classification maps at lower resolution than original image. Initial layout hypothesis is generated by logistic regression. To obtain final layout an iterative refinement process is used. The layout is further optimized by the four vanishing lines and one vanishing point to generate the final layout. It is having a fall short of exploiting end to end learning ability of CNNs.

For any given layout the scoring function is defined as:

$$S(L = f(\tau)|T) = \frac{1}{wh}\sum_{i,j} T_{i,j}^{(L_{i,j})} \tag{2}$$

To obtain the final layout τ* an iterative refinement process is used. The optimization algorithm is described in Algorithm 1. This algorithm greedily optimizes each parameter in τ sequentially, and repeats until no further refinement in the score can be obtained.

```
Algorithm 1: Layout Optimization[9]
Input:    T                      // Output of CNN
          $(l_1,l_2,l_3,l_4,v)$  // Initialization
Output:   Layout $\tau^* = (l_1,l_2,l_3,l_4,v)$
repeat
    foreach Candidate vanishing point p do
        evaluate $S(\tau = (l_1,l_2,l_3,l_4,p)|\mathbf{T})$
        if Score improved then
            v = p
        end
    end
    foreach $I \in (1\ldots4)$ do
        foreach Candidate line l do
            evaluate $S(\tau = (l_1,l_2,l_3,l_4,v)|\mathbf{T})$
            if Score improved then
                $l_i = l$
            end
        end
    end
until  Score did not improve
```

Lee *et al.* [10] adopted a direct formulation of room layout estimation problem by predicting the locations of the room layout key-points with an end-to-end network. It is a simple and direct formulation of layout estimation problem as a key-point localization problem. The network infer both the room layout corners (key-points) and room type. Then connect the key-points in specific order of room type to obtain the spatial room layout. Here a convolutional encoder-decoder architecture is used. The semantic segmentation of the layout surfaces is simply attainable because of this connectivity. Used a key-point based room layout representation to train the model. Fig. 4 shows a list of room types with their respective key-point definition as defined by [25]

All the experiments are evaluated in two challenging benchmark datasets: Hedau [4] dataset and Large-scale Scene Understanding Challenge (LSUN) room layout dataset [25].

Yang *et al.* [11] presented a system for reliable real-time corridor layout understanding from a single image. This is applicable for robot navigation. It contains a learning algorithm to detect ground and wall planes. An efficient and accurate CNN+CRF classifier is used to segment indoor images into two geometric classes. Then used geometric constraints to compute the relative orientations of the wall and ground to pop up ground and wall planes into a simplified 3D model.

It focuses on corridor environments, which mobile robots operating indoors often have to traverse. The dataset contains 967 images from three sources: 349 images from the SUN RGBD [12] (category “corridor”); 327 images from SUN database [13] (category “corridor”) and 291 images from the video taken around the Carnegie Mellon University campus.

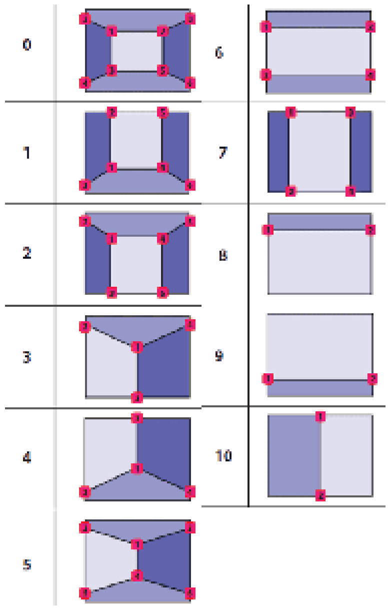

Fig: 4. Defenition of room layout types. The type is indexed from 0 to 10.

Ren *et al.* [14] used the learned edge maps as a hint for generating layout hypotheses based on the vanishing lines, undetected lines, and occluded lines. Adopted a multi-task fully convolutional neural network to jointly predict the surface labels and boundaries. This CFILE (coarse-to-fine indoor layout estimation) system consists of two stages. In the first stage, a multi-task fully convolutional neural network (MFCN) is used to get a coarse but robust layout estimation. Since the CNN is weak in accomplishing spatial smoothness and conducting geometric reasoning, it cannot provide a fine-scale layout result. In the second stage, the coarse layout from MFCN is used as the guidance to detect a set of critical lines. Then, a small set of high quality layout hypotheses are generated based on these critical lines. Finally, a score function is defined to select the best layout as the desired output.

It has an architecture that employs the VGG-16 network for the encoder followed by fully-connected layers and deconvolutional layers that up sample to one quarter of the input resolution. The fully-connected layers enables the network to have a large receptive field but at the cost of loosing the feature localization ability. The layout refinement consists of two steps: generate a hypothesis set and define a score function.

The score function is defined as:

$$S(L|P) = \frac{1}{N}\sum_{i,j} P_{(i,j)}\,, \forall L_{(i,j)} = 1 \qquad (3)$$

where P is the output from the MFCN, L is a layout from the hypotheses set, N is a normalization factor that is equal to the total number of layout pixels in L. Then, the optimal layout is selected by:

$$L^* = \arg\max_{L} S(L|P) \qquad (4)$$

The layout with highest score is chosen to be the final layout. This method is evaluated on two popular datasets; namely, Hedau's dataset [4] and the LSUN dataset [26].

Lin *et al.* [15] introduced a deep-learning based approach for estimating the layout of a given indoor image in real-time. It motivates the generalization ability of the network and the smoothness of estimated layout edges without deploying post-processing techniques. They used a stronger ResNet-101 backbone and model the network in a fully convolutional manner.

This model inputs a colour image and outputs the planar semantic segmentation of the same. It followed the layout representation proposed in DeLay [9] in which the layout estimation can be regarded as a five-class planar semantic segmentation problem.

Zhang *et al.* [16] proposed a deconvolution network which has multi-layer deconvolution to attain highly reliable edge maps. The deconvolution network predict the edge map of a room image. Then candidate layouts are generated from the edge map using an adaptive sampling strategy. The best fit layout is selected from ranking.

The deconvolution network is trained to estimate the room edge maps. The network has feature extraction and map generation part. The feature extraction part is similar to AlexNet [17] and map generation part consist of four successive deconvolution layers. This part generate high quality edge maps out of the features extracted from feature extraction part.

For box layout estimation candidate layouts are generated from the edge map and then select the best-fit layout by ranking. It models a room as box as in Hedau [4]. The three vanishing points are estimated using the method of [4]. These are ordered as vertical, farther horizontal and closer horizontal vanishing points. An adaptive sampling strategy is used in this. The edge mapo is divided into several uniformly spaced sectors by the sampled rays from the vanishing points. The total number of sectors is denoted as M. $s_i$ is the average edge strength of each sector where i = 1,2,...M. The $i^{th}$ sector is selected if it satisfies the below conditions.

1. $s_i > s_{i+1}$ and $\mathrm{s_i} > \mathrm{s_{i-1}}$,
2. $s_i - s_{i+1} > D$ and $s_i - s_{i-1} > D$.

The criterion to find the best fit edge map is as below.

$$\mathrm{s_i} = \frac{\mathrm{m_i} \cdot \mathrm{M}}{\|\mathrm{m_i}\|_\mathrm{F} \|\mu\|_\mathrm{F}} - \mu \|m_i - M\|_F \qquad (5)$$

Where $m_i$ denotes the edge map produced by the $\mathrm{i}^{th}$ parameterized candidate layout. M is the predicted edge map of deconvolution net. $\|.\|_\mathrm{F}$ indicates the Frobenious norm.

Most recently, Zhang *et al.* [18] proposed an architecture based on the VGG-16 backbone for simultaneously estimating the layout edges as well as predicting the semantic segmentation of the walls, floor and ceiling. It is an encoder-decoder network with shared encoder and two separate decoders, which are having multiple transposed convolution layers. The network predictions are combined in a scoring function to evaluate the quality of candidate layouts. Candidate layouts are generated through ray sampling and from a predefined pool.

In the encoder-decoder network, the encoder is same with VGG-16 model. Then the network is divided into two branches of decoders. Each branch consist of four successive deconvolution layers. The first branch outputs edge map and second branch outputs five heat maps.

The layouts are generated in two different methods : Ray sampling and Predefined pool. Ray sampling method is same as [16]. For the predefined pool the training samples of LSUN dataset are all included. The final layout hypotheses L are produced by combining selected layouts from ray sampling and predefined pool. Layout optimization is applied to the layout hypotheses L. The optimization algorithm is as below.

**Algorithm 2 : Layout optimization[18]**

```
Input:      layout hypotheses L, segmentation map M, edge
            map E
Output:     optimized layout $l$*
for each    $l_k \in$ L, k = 1,2,…,K do
    $l = l_k$, where  $l = (t, p_1, p_2, \ldots, p_n)$
    $s = \mathrm{S}(\mathbf{M}_l, \mathbf{E}_l \mid \mathbf{M}, \mathbf{E})$
    while True do
        for each $p_i$ do
            generate neighbor points $\Pi_i = \{ p_i^1, p_i^2, \ldots \}$
            for each $p_j^i \in \Pi_i$ do
                replace $p_i$ with $p_j^i$, obtain a new layout
                l'
                $s' = \mathrm{S}(\mathrm{M}_l', \mathrm{E}_l' \mid \mathrm{M}, \mathrm{E})$
                if s' > s then
                    L= l',s= s'
        if score does not increase then
            break
    $l_k^* = l$
return    $l^* = \max(l_k^*)$
```

Boniardi *et al.* [19] introduced a more parameter efficient encoder with dilated convolutions and incorporate the eASPP for capturing large context, complemented with an iterative training strategy that enables to predict thin layout edges without discontinuities.

| Research Paper | Discussions |
|---|---|
| Hedau *et al.*[4] | Spatial layout of cluttered rooms can be modelled jointly in terms of a 3D box layout and surface labels of pixels. Candidate 3D boxes are generated and inference is formulated in terms of single high dimensional discrete random variable. |
| Chao *et al.*[6] | Exploits 3D geometric relationships between people and room box to jointly estimate vanishing points. |
| Mallya *et al.*[8] | Layout prediction by just using edge based features which is more simpler than earlier works. Used FCNN solely for generating an intermediate feature. |
| Yang *et al.*[11] | Real time understanding by labelling frames at real time. Excessive simplification of segmentation boundaries. |
| Dasgupta *et al.*[9] | Demonstrated FCNs can be adapted to estimate layout labels directly from RGB images. Use CNN to generate a set of low level features. Fall short of exploiting end to end learning ablility of CNNs. |
| Ren *et al.*[14] | Train deep network features to classify pixels into combination of layout surfaces and boundary surfaces. |
| Lee *et al.*[10] | Simple and direct formulation of room layout estimation as a keypoint localization problem. |
| Zhang *et al.*[16] | Intoduced an adaptive sampling strategy. |
| Zou *et al.*[22] | Works well with non cuboid layouts. |
| Yang *et al.*[23] | Good performance in cuboid shaped rooms. Accuracy drops significantly as number of corners increases. |
| Deng *et al.*[20] | Deep network that combines textures and geometric hints |
| Lin *et al.*[15] | Strongest RestNet-101 backbone and model the network in a fully convolutional manner. |
| Boniardi *et al.*[19] | The network has large receptive field which helps to capture good layout edges in occlusions also. |
| Zhang *et al.*[18] | semantic labels are joinly learned with edge maps where they could mutually benefit each other. Discountinous edges in case of significant clutter. |

Deng *et al.* [20] employed a deep network that combines textures and geometric hints to predict the surface layout from a single image. First, depths and normals are estimated from the input RGB image. It used muti-scale convolutional network proposed in [21] to estimate the depth and normal maps from RGB images. Secondly, a multi-channel FCN is used to integrate these for semantic surface segmentation. Then, an optimization framework is used to refine the layout estimation. For any candidate layout $L$ with $r$ semantic surfaces, the score function is defined as:

$$s(\bar{L}|T) = \frac{1}{w*h}\sum_{r} T_r^{(\overline{L_r})} \tag{6}$$

where r represents the pixels in a certain region for the corresponding semantic surface. The proposal with highest score is selected as $L^*$, namely:

$$L^* = arg\underset{\bar{L}}{\underline{m}}ax\; S(\bar{L}|T) \tag{7}$$

Zou *et al.* [22] estimate the room layout from a single panorama or perspective image. It trains a FCN from panoramas and vanishing lines and generate the layout models from edge and corner maps. First, it analyzes the vanishing points and align the image to be level with the floor. Then corner and boundary probability map are predicted directly on the image using a CNN. Finally, the layout parameters are optimized to fit the predicted corners and boundaries. It relaxes the commonly assumed cuboid layout limitation and works well with non-cuboid layout.

Yang *et al.* [23] used an end-to-end deep learning framework, for estimating 3D room layouts from a single panorama image. It has the ability to produce general room shape not limited to cuboid shape. The network architecture consists of two encoder-decoder branches for analyzing features from two distinct views of the input panoramas. Integrate the surface semantic mask from conventional equi-rectangular view and the projected floor and ceiling view. The panorama and ceiling branch are connected through a feature fusion scheme through which information can be shared between them. The network is jointly trained to output a floor-ceiling and floor-plan probability map and a layout height. But here the accuracy drops significantly as number of corners increases.

## III. DISCUSSIONS AND CONCLUSIONS

Room layout estimation is one of the main tasks of 3D indoor scene understanding. Various studies have been introduced for room layout estimation. Most of the room layout estimation methods are based on the common assumption of Manhattan worlds [3]. Recently edge maps or semantic labels learned from FCN has significantly enhanced the layout estimation performance. Most of the CNN based approaches for estimating room layout edges employ an encoder-decoder topology. Discussions on the papers mentioned above are shown in Table I.

ACKNOWLEDGMENT

We would like to thank our Director (IHRD) and Principal of our institution for providing us the facilities to support this work. In addition, we thank Jyothi R L, Asst Professor in CS of opur institution for the valuable comments that greatly improved the work.

## REFERENCES

[1] S. L. Joseph “Semantic indoor navigation with a blind-user oriented augmented reality,” in Proc. IEEE Int. Conf. Syst. Man Cybern.,2013, pp. 3585–3591.

[2] F. Boniardi, A. Valada, R. Mohan, T. Caselitz, and W. Bur-gard. “ Robot localization in floor plans using a room layoutedge extraction network.”, arXiv preprint arXiv:1903.01804,2019.

[3] Coughlan, J.M., Yuille, A.L. “Manhattan World: Compass Direction from a Single Image by Bayesian Inference.” in Int. Conf. on Comp. Vis., pp 941–947, 1999.

[4] V. Hedau, D. Hoiem, and D. Forsyth, “Recovering the spatial layout of cluttered rooms,” in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 1849–1856

[5] C. Rother. “A new approach to vanishing point detection in architectural environments” IVC, 20, 2002.

[6] Y.-W. Chao, W. Choi, C. Pantofaru, and S. Savarese. “Layout estimation of highly cluttered indoor scenes using geometric and semantic cues.” In Image Analysis and Processing–ICIAP 2013, pages 489–499. Springer, 2013.

[7] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe:A database and Web-based tool for image annotation,” Int. J. Comput. Vis., vol. 77, nos. 1–3, pp. 157–173, 2008.

[8] A. Mallya and S. Lazebnik, “Learning informative edge maps for indoor scene layout prediction,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 936–944.

[9] S. Dasgupta, K. Fang, K. Chen, and S. Savarese, “DeLay: Robust spatial layout estimation for cluttered indoor scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 616–624.

[10] C.-Y. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich, “RoomNet: End-to-end room layout estimation,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 4875–4884.

[11] Shichao Yang, Daniel Maturana, and Sebastian Scherer. “Real-time 3Dscene layout from a single image using convolutional neural networks”.In Robotics and automation (ICRA), IEEE international conference on,pages 2183 – 2189. IEEE, 2016.

[12] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. “Sun RGBD: A rgb-d scene understanding benchmark suite.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 567–576, 2015.

[13] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio Torralba, et al. “Sun database: Large-scale scene recognition from abbey to zoo.” In Computer vision and pattern recognition (CVPR) 2010 IEEE conference on , pages 3485-3492.

[14] Y. Ren, S. Li, C. Chen, and C.-C. J. Kuo, “A coarse-to-fine indoor layout estimation (CFILE) method,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 36–51.

[15] H. J. Lin, S.-W. Huang, S.-H. Lai, and C.-K. Chiang, “Indoor scene layout estimation from a single image,” in Proc. of the IEEE International Conference on Pattern Recognition, 2018.

[16] W. Zhang, W. Zhang, K. Liu, and J. Gu, “ Learning to predict high quality edge maps for room layout estimation,” IEEE Transactions on Multimedia, vol. 19, no. 5, pp. 935–943, 2017.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[18] W. Zhang, W. Zhang and J. Gu, “Edge-semantic learning strategy for layout estimation in indoor environment,” IEEE transactions on cybernetics, (2019).

[19] F. Boniardi, A. Valada, R. Mohan, T. Caselitz, and W. Bur-gard. “ Robot localization in floor plans using a room layoutedge extraction network.” arXiv preprint arXiv:1903.01804,2019.

[20] Ruifeng Deng,Xuejin Chen, ”Robust Room Layout Estimation from a Single Image with Geometric Hints” in Proc. of the IEEE International Conference on Image Processing, 2018.

[21] D.Eigen and R.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, pages 2650-2558

[22] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. “Layoutnet: Reconstructing the 3d room layout from a singlergb image.” InThe IEEE Conference on Computer Vision andPattern Recognition (CVPR), June 2018.

[23] S.-T. Yang, F.-E. Wang, C.-H. Peng, P. Wonka, M. Sun, and H.-K. Chu. “Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama.” arXiv preprintarXiv:1811.11977, 2018.

[24] C. Liu, A. G. Schwing, K. Kundu, R. Urtasun, and S. Fidler, “Rent3D:Floor-plan priors for monocular layout estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3413–3421.

[25] S. L. Joseph et al., “Semantic indoor navigation with a blind-user oriented augmented reality,” in Proc. IEEE Int. Conf. Syst. Man Cybern., 2013, pp. 3585–3591.

[26] L. Li et al., “A wearable virtual usher for vision-based cognitive indoor navigation,” IEEE Trans. Cybern., vol. 47, no. 4, pp. 841–854, May 2017.

[27] H. Qiao, Y. Li, F. Li, X. Xi, and W. Wu, “Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning,” IEEE Trans. Cybern., vol. 46, no. 10, pp. 2335–2347, Oct. 2016.

[28] Y. Zhang et al.” Largescale Scene Understanding Challenge: Room Layout Estimation”. Accessed: Sep. 15, 2015. [Online]. Available: http://lsun.cs.princeton.edu/2016/

[29] V. Hedau, D. Hoiem, and D. Forsyth, “Thinking inside the box: Using appearance models and context based on room geometry,” in Proc. Eur. Conf. Comput. Vis., pp. 224–237, 2010.

[30] J. M. Coughlan and A. L. Yuille, “The manhattan world assumption: Regularities in scene statistics which enable Bayesian inference,” in Proc. NIPS, vol. 2, 2000, p. 3.