DeepRoom: 3D Room Layout and Pose Estimation from a Single Image

Hung Jin Lin and Shang-Hong Lai(✉)

Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
vtsh.jn@gmail.com, lai@cs.nthu.edu.tw

Abstract. Though many deep learning approaches have significantly boosted the accuracy for room layout estimation, the existing methods follow the long-established traditional pipeline. They replace the front-end model with CNN and still rely heavily on post-processing for layout reasoning. In this paper, we propose a geometry-aware framework with pure deep networks to estimate the 2D as well as 3D layout in a row. We decouple the task of layout estimation into two stages, first estimating the 2D layout representation and then the parameters for 3D cuboid layout. Moreover, with such a two-stage formulation, the outputs of deep networks are explainable and also extensible to other training signals jointly and separately. Our experiments demonstrate that the proposed framework can provide not only competitive 2D layout estimation but also 3D room layout estimation in real time without post-processing.

Keywords: Room layout estimation · Deep learning · Pose estimation

1 Introduction

The research on 3D scene understanding dates back to 1960s' simple Block World assumption [20], with the vision of reconstructing the global scene with local evidences, and nowadays it has become one of the most pivotal research area in the era of artificial intelligence and deep learning. The goal for scene understanding is to know the semantic meaning of each single object and also the environments constructed the scene. For the case of indoor scene, it is often referred to the topics like object detection and semantic segmentation at the object-level, and structure-level information such as spatial layout estimation. The effectiveness of the indoor layout estimation can be applied to applications such as the indoor navigation, localization, and the virtual object arrangement in the rooms.

The layout estimation for an interior room from an image can be represented in several levels of structure with different parameterization; for example, pixel-wise classification labeling, crossing rays originating from vanishing points, and the projection of 3D solid geometry models. Many layout estimation works are based on the underlying assumption of "Manhattan World" proposed by Coughlan [3] in 1999. It means scenes are composed of three dominant orthogonal orientations, and the walls are perpendicular to each other as well as the

S. Palaiahnakote et al. (Eds.): ACPR 2019, LNCS 12047, pp. 719–733, 2020.
https://doi.org/10.1007/978-3-030-41299-9_56

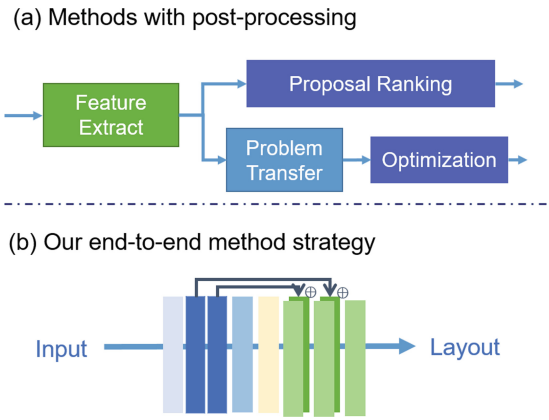


Fig. 1. Difference between our proposed framework versus the previous methods in terms of system pipeline: (a) previous systems usually require post-processing, and (b) our end-to-end approach.

ceiling and floor. The cuboid model is applied to represent the room in most of the cases, in which the room enclosed by four walls, floor and ceiling. The earlier researches with machine learning approach tailored optimization and post-processing for the geometry reasoning on the hand-crafted features from the single-view image. In the succeeding methods with deep learning, existed works still built the framework as a two-tier pipeline with the deep neural networks for feature discription and the optimization step for the final estimate. These extra procedures, refining the layout estimation or extracting the 3D representation for the modeling of layout, usually take considerable amounts of computation and make them far from real-time applications. On the other hand, the widely used room layout datasets do not provide appropriate 3D annotated information, and it makes the problem of 3D modeling more challenging. Furthermore, very few deep learning models have been proposed for such 3D geometry recovery problems, and we simplify this with a deep neural network solution (Fig. 1).

This is not to deny the astonishing results achieved by the aforementioned methods. However, the time consumption of post-process rendered these methods unsuitable for time-efficient applications. And, many deep learning methods make the layout as the pixel-wise dense representation which is not enough to describe the 3D structure of layouts. To address these issues, in this paper, we propose a novel framework for predicting the 2D as well as 3D room layout estimation with efficient deep neural networks in a row. Under this framework, we will estimate the 2D layout as the intermediate representation, and then predict the 3D cuboid modeling parameters from the 2D representation. As the result, our method can estimate the layout in 2D and also 3D space completely through deep networks and provide the state-of-the-art results in real-time without post-processing.

The main contributions of this paper are threefold (Fig. 2),

– We decouple the layout estimation into two neural networks with the explainable intermediates separately in conjunction with effective training strategies;
– We believe that we are the first to model the 3D layout estimation task with two efficient end-to-end networks, and thus achieve real-time estimation;
– We demonstrate how to make good use of the existing datasets with the limitation of only the 2D layout annotations available to achieve the capability of 3D layout estimation from a single image.

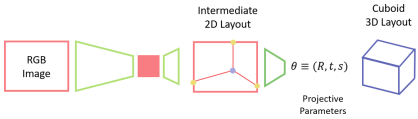


Fig. 2. An overview of our framework composed of two-stage networks.

2 Related Work

2.1 Room Layout Estimation

With the Manhattan assumption, Hoiem [8] proposed to estimate the outdoor scene geometry through learning appearance-based models of surfaces at different orientations and geometric context [7]. On the other hand, Hoiem [9] made the concept of geometric context into indoor scenes, and modified the labels into six classes for the indoor case: left-wall, right-wall, front-wall, ceiling, floor and objects, and which is also the most common classification modeling for the indoor layout estimation that inspired many later researches. In [9], they took features from color, texture, edges and vanishing point cues computed over each superpixel-segment, and applied a boosted decision tree classifier to estimate the likelihood of each possible label. Lee [14] alternatively used the orientation maps for the feature description which takes the layout and the objects oriented with three orthogonal orientations. In the traditional layout estimation approaches, the researchers extract several meaningful evidence from images, such as line segments [20,24], orthogonal orientations of line segment [14], superpixel [21], and contextual segments [7], or the volume form like geons [2]. However, these evidences fail in the cases of highly cluttered and occluded scenes containing less meaningful local features for the structure, and thus, some researched on inferring estimations through hyper volumetric reasoning [5,6,22].

With the rise of machine learning, structured learning [18] has been developed for the task, and its goal is to model the environment structure by generating hypotheses with incomplete low-level local features [5,6,17].

2.2 Room Layout Estimation with Deep Learning

With the successful modeling in the previous works, many have resorted to the deep learning approach due to its superior performance in several computer vision tasks. Some adopted the end-to-end supervised FCN (fully convolutional network) model [16] to the perspective of room layout estimation as a task of critical line detection, for instance the estimation of informative edges in [17] and coarse and fine layout joint prediction in [19]. Dasgupta [4], the winner of LSUN Room Layout Challenge 2015 [23], tackled the task with a two-tier framework: segment the planes and walls of the input image with a deep neural network first, and then optimize the output with the vanishing point estimation. The promising result of Dasgupta *et al.* inspired several subsequent works [19,25,26] to follow their two-tier pipeline, which consists of a FCN-like network for semantic segmentation and a layout optimization technique (e.g., layout hypotheses proposal-ranking pipeline in [4,17,19,25], and special optimization modules in [26]) for post-processing. For real-world application, however, the post-processing may be impractical if it is time-consuming.

2.3 Camera Pose and Geometry Learning

In early works, researchers viewed the locating 6-DoF camera pose task as a Perspective-N-Point (PnP) problem. For example, Arth [1] estimated pose of the query camera by solving a modified 3 point perspective (3PP) pose estimation problem after extracting the epipolar geometry of a query images and its nearby images. In recent works, researchers tend to estimate the 6-DoF camera pose via deep learning. PoseNet [12] is the first approach to regress the 6-DoF camera pose from a single RGB image through an end-to-end CNN. After that, a sequential works extended the PoseNet. To deal with the uncertainty of output predictions, Kendall [10] changed the network into a Bayesian model. And in [11], they notably improved the performance of PoseNet by introducing noval loss functions according to the geometry and scene reprojection error.

3 Layout Estimation in 2D Space

Our layout estimation framework can be decoupled into two stages, the 2D layout estimation in Sect. 3, and 3D cuboid model representation through projective parameters estimation in Sect. 4 and these two stage can be applied either jointly or separately.

3.1 Multi-purpose Layout Network

Under the Manhattan world assumption, we can consider each indoor scene as being composed of multiple orthogonal planes and the layout of regular room can be further simplified into the cuboid model. From this perspective, the layout estimation can also be regarded as a region segmentation problem on each surface

of cuboid. To describe the segments of these regions, it can be parameterized by the densely segmentation, or the borders or the points of these polygons. In the previous deep learning methods, researchers proposed several representations, such as planar segment with semantic labels [4,15,19], scoring heatmap on the layout edges [17,26], and corner heatmaps [13]. From these works, we found that each representation has its pros and cons for the later usage, i.e. post-process methods. Consequently, we take the success of Lin [15] for fully convolutional network adopted ResNet 101 as the back-bone, which is the state-of-the-art layout estimation without post-process. We further advance the estimation for 2D layout in multiple representations through the multi-stream decoders, including the forms of *corners*, *edges* and the *semantic planes* simultaneously (Fig. 3).

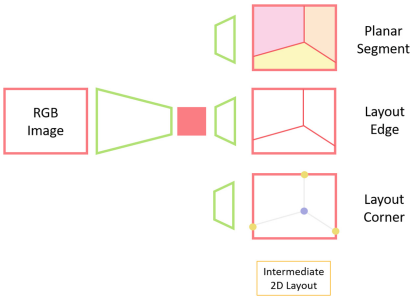


Fig. 3. The 2D layout estimation in multi-task network. The basic component is the ResNet101 back-bone network for feature extraction, and we make an independent decoders (up-sampling module) for each distinct targets, for planar segmentation, layout edge, and also layout corners.

Layout Segment. We refer the dense region segment to semantic planar described in [4] for five classes: front-wall, right-wall, left-wall, ceiling, and floor. It can then be formed as one semantic segmentation like problem with the labels on larger structure scale rather than object-level segments.

Layout Corner. RoomNet [13] estimates the corners for each possible layout structure in nearly fifties-channel heatmaps which is computationally inefficient. However, the corners labeled in captured scenes are given by two kinds of points, one is the room corner (*inner corner*), the other is the intersected points on the borders of image (*outer corner*), as illustrated in Fig. 4. In other words, we can categorize them into two classes rather than abundant channels nor as single one.

Layout Edge. The layout edge can be represented by the borders of the polygons. The detection of borders is to determine whether the pixel is the edges for the room layout. And it can be viewed as a binary classification problem.

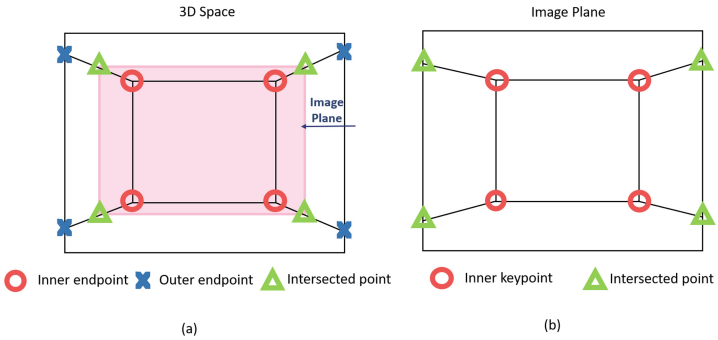


Fig. 4. (a) Showing each endpoint of the room cuboid and the projective image plane. (b) Indicating the projected points of inner and intersected ones.

3.2 Layout-Specific Objective Criterion

As proposed in [15], they find that if directly apply the vanilla semantic segmentation criterion on planar layout estimation, the result often suffers from distortion or tears apart from the center of planes and also "wavy curves" (rather than straight lines) mentioned in DeLay [4]. Hence, imposing extra smoothness criterion is necessary to alleviate the artifacts. The proposed loss function is given by,

$$\mathcal{L}_{seg}(x, target) = CE(x, target). \tag{1}$$

where x is the output of network for each single estimate, a five classes representation for semantic planar segmentation. And the smoothness term is given by,

$$\mathcal{L}_{smooth} = \ell(x, target) = |x - target|_1. \tag{2}$$

Loss of Corner and Edge Detection. The tasks for corner and edge detection can be viewed as binary classification on pixel-level, and thus the loss function can be given by the binary cross-entropy to determine whether one pixel belongs to a layout structure edge,

$$\mathcal{L}_{edge}(x, target) = BCE(x, target). \tag{3}$$

$$\mathcal{L}_{corner}(x, target) = \sum_i BCE(x_i, target_i). \tag{4}$$

For the corners, we categorize them into the inner and outer ones, the criterion would be the summed loss across the two-category corner maps.

Overall Loss. The criterion for the planar layout task is,

$$\mathcal{L}_{plane} = \mathcal{L}_{seg} + \lambda_s \mathcal{L}_{smooth}, \tag{5}$$

And the overall objective loss criterion for our network is the summation for these three branches. The overall loss function for model training is given by,

$$Loss_{Net2D} = \mathcal{L}_{plane} + \mathcal{L}_{edge} + \mathcal{L}_{corner}. \tag{6}$$

4 Layout Beyond Pixels

Most important of all, we further propose the second stage for 3D layout estimation in neural network, the novel approach compared to the existed works, by making use of those 2D intermediates from previous stage. The common rooms in the daily scenes are formed by the cuboid models. In the traditional works [6,14], they considered to model the 3D layout of various room scenes to be composed by boxes, and generated layout proposals based on 2D hand-craft cues and also optimization-based pipeline. We found that, however, in computational geometry, the 2D corners can be considered as the projection of 3D layout when depth information is reduced to the 2D space. Consequently, the task can then be converted to reconstruct the layout structure by estimating the projection of a cuboid, and we can then formulate the parameters for the transformation and projection. We thus parameterize the 3D layout in Sect. 4.1 with transformations and corresponding camera pose in the canonical 3D coordinate.

We make one neural network to predict the cuboid representation for the 3D layout in Sect. 4.2. However, we have no annotated 3D information for the supervised network, thus we resort to make use of the synthesized data with the strategy *abstract layout generation*, and then deliver the knowledge to the real case through transfer learning detailed in Sect. 4.3. With such formulation, we can estimates the 3D room layout with the representation of *projective parameters* from the 2D intermediate representation of layout estimated from stage one. And we now can make the 3D layout estimation framework end-to-end via deep networks.

4.1 Cuboid Model Parameterization

There are two components for the representation in our cuboid model, the scale of cuboid and camera pose. The parameters for the camera pose are decomposed into translation vector $\mathcal{T}$ and rotation matrix $\mathcal{R}$, in which we need three parameters for the position of camera and three parameters for the rotation angles along three coordinate axes, represented in quaternion. And three more parameters for the scaling along three axes of the unit box, template cube placing at the origin of the canonical space. Let $\mathcal{X}_{3D} \in \mathbb{R}^{3\times N}$ denote the 3D coordinates of eight keypoints ($N = 8$) belonging to the unit box, and the locations of box keypoints viewed by a specific camera pose, and $\mathcal{X}_{2D} \in \mathbb{R}^{2\times N}$ denotes the corresponding 2D coordinates in the image space. Thus the relationship between two coordinates is given by (Fig. 5)

$$\mathcal{X}_{2D} \equiv \pi \left(\mathcal{X}_{3D} | \mathcal{K}, P\right). \tag{7}$$

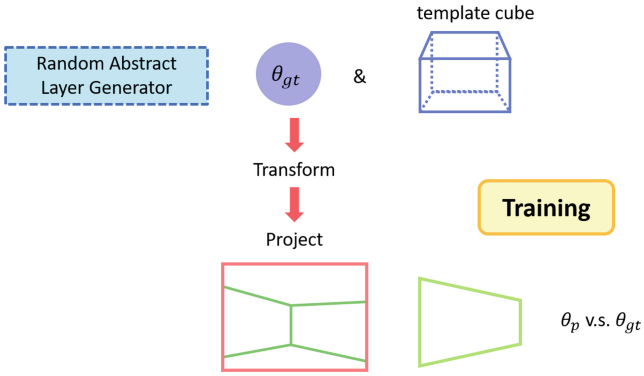


Fig. 5. The training of the regression model with the strategy of random generation of abstract layout. We can synthesize the paired samples, confident layout edge and the corresponding ground truth parameters θ_{gt}.

where $\mathcal{K}$ is the camera intrinsic matrix assumed to be given in the camera calibration procedure and $\mathcal{P}$ is the projection matrix given by

$$\mathcal{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{8}$$

$$P = [\mathcal{R}|\mathcal{T}] \in \mathbb{R}^{3\times 4}. \tag{9}$$

Note that the rotation matrix $\mathcal{R} \in \mathbb{R}^{3\times 3}$ and the translation vector $\mathcal{T} \in \mathbb{R}^3$ contain the extrinsic parameters for camera pose, and the rotation matrix is represented by a quaternion vector $\hat{q}$ as follows:

$$R = quat2mat(\hat{q}) \in \mathcal{SO}(3), \tag{10}$$

Hence, we can extract the 3D cuboid layout from 2D space by estimating the *projective parameters* for the cuboid model.

4.2 Regression Forwarding Network

We formulate the task as one regression task by applying the CNN to learn these projective transformation parameters θ_t. Nevertheless, it is a challenge to train such a regression model since most of the datasets for the layout estimation do not provide any 3D annotations. The datasets for spatial layout are often annotated with 2D information such as the shapes of polygon of layout and the coordinates of the corners. And thus there is no easy way to retrieve 3D signals for supervised learning as one regression task.

So, we reformulate the problem as follows. The original task is to regress the model for the target parameters $\theta_t \in \mathbb{R}^9$ from the input $\in \mathbb{R}^{H\times W}$. Under this configuration, we can resort to the intermediate 2D layout representation

$\mathcal{E} \in \mathbb{R}^{H \times W}$, the estimate of 2D layout network in Sect. 3.1. The key value for the task decoupling is that the intermediate layout representation is easy to be synthesized by projecting the deformed cuboid onto the image plane, and we call it *abstract layout generation*. As a result we can acquire lots of reasonable samples through random generating target parameters θ_g as well as the corresponding 2D layout representation input $\equiv \mathcal{E}_g$ for the regression task, by using the transformation and projection modules described in Eq. 7.

With such a strategy, we can reform the ill-posed regression task and overcome the challenge of lacking 3D annotations in the existing datasets. The design of our regression network is composed of nine compounded layers of strided-convolutional layer with ReLU non-linearity activation and 1×1 convolutional layer acted as fully connected layers at the end of the network for the target projective parameters $\Theta \in \mathbb{R}^9$ of the cuboid layout representation.

4.3 End-to-End Learning Network

Besides the synthesized data, we need to make the trained regression model work on the real signals. under the configuration of our framework, the input of regression model is generic to the intermediate of previous stage, in the same space—2D layout representation. Though the estimated layout edge from 2D layout network is not as perfect as the one generated from ground truth, we can still make use of transfer learning strategy as Fig. 6 to make an extra network to learn and fit as close as to the model training on synthesized samples. Finally, we make an end-to-end framework for the 3D layout estimation via pure deep networks instead of any optimization or post-processing.

5 Experimental Results

We utilize LSUN Room Layout dataset [23], containing 4,000 training images, 394 validation, and 1,000 testing images, for evaluating 2D semantic planar segmentation and corner estimation results. Since there are no public labels for the testing set, we evaluate our method on the validation set with LSUN Room Layout official toolkit like the previous works. In addition, we evaluate the generalization capability of our model on the Hedau dataset [6], which is a challenging dataset due to its strict labeling. We can not evaluate any 3D accuracy metrics for our 3D layout estimation, for these two commonly used datasets do not contain any 3D annotations for the layout estimation. Instead, we also evaluate the 3D layout estimation results with 2D metrics on the re-projection of 3D layout.

Note that we only train our model on the training split of LSUN Room Layout and directly test on the testing split of Hedau dataset without fine-tuning on its training data. During the training, we apply random color jittering for slightly changing in the lightness and contrast of color images to increase the diversity of scenes. In addition, the time efficiency of our approach and other methods is also reported in our experiment.

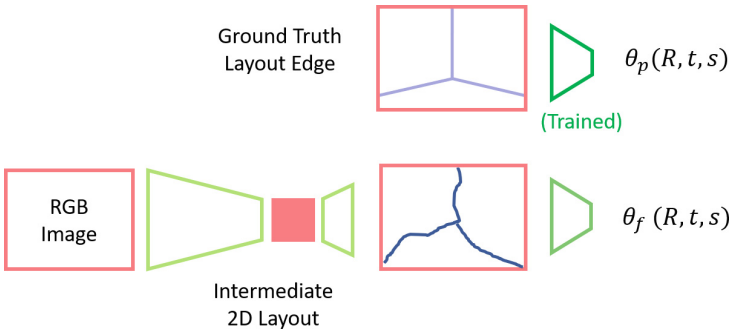


Fig. 6. The transfer learning pipeline for the 3D cuboid parameters estimate on real outputs of network.

5.1 Quantitative Results

We measure the performance of the proposed approach in 2D and 3D layout estimation through the following experimental evaluations: 2D pixel-wise accuracy for semantic planar segmentation in the single task and multi-task networks, 2D corner prediction accuracy for the keypoint corner detection, re-projected accuracy on 2D metrics on the estimated 3D projective parameters, and the visualization for the 3D cuboid rooms of the estimated parameters.

Pixel-Wise Accuracy of Layout Estimation. The performance of our layout estimation results are shown in Table 1. First, we take *DeepRoom 2D* for planar segmentation without any training strategies as our baseline model, and it can already achieve 9.75% error rate. And, the extended model *DeepRoom 2D multi-task* can reduce the error rate to 7.04%, which is 2.71% better than the baseline. Moreover, the performance of the ones trained with *Layout Degeneration* are comparable to the state-of-the-art method in the LSUN Challenge and we achieve 6.73% and 6.25% pixel-wise error rate for the single and multi-task networks, respectively. Furthermore, if we compare under a more fair condition, our proposed model can even beat the best performing method ST-PIO [26] (ST-PIO (2017) w/o optim.) without the extremely high-cost physical-inspired optimization but with the post-processing for proposal ranking.

We list the results from the direct 2D estimation networks and also the re-projected performance from the 3D parameter estimation network in Table 1. For the 3D projective parameters, *DeepRoom 3D*, which takes the ground truth generated edge map as input, can achieve similar performance as the 2D network in the metric of pixel-wise accuracy. Furthermore, the end-to-end approach of our *DeepRoom 2D/3D* achieves about 10% error rate, which is about similar level of the other state-of-the-art methods, LayoutNet [27], without post-processing.

Table 1. The pixel-wise accuracy performance benchmarking on LSUN Room Layout dataset for different approaches. Note that the data in the table is extracted from their papers.

Method	Pixel error (%)
Hedau [6]	24.23
DeLay [4]	10.63
CFILE [19]	7.95
RoomNet [13]	9.86
Zhang [25]	12.49
ST-PIO [26]	**5.48**
ST-PIO w/o optim. [26]	11.28
LayoutNet [27]	11.96
Ours 2D baseline, planar seg	9.75
Ours 2D multi-task	**6.73**
Ours 3D re-projected	6.89

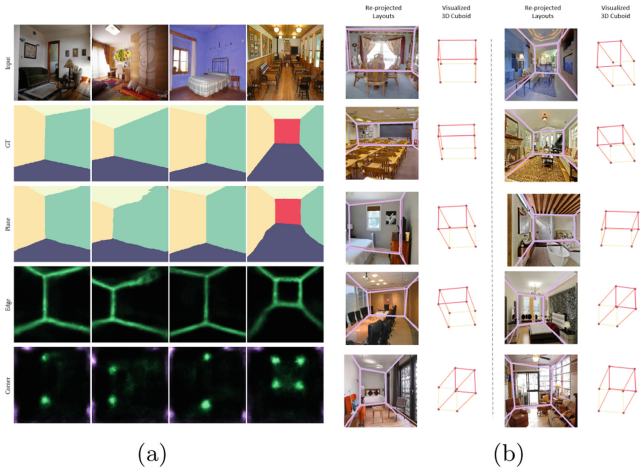
5.2 Qualitative Results

First, we want to demonstrate the effects of our proposed layout objective criteria for layout segmentation. We show the visual outputs for our multi-stream networks with the full training strategies in Fig. 7. They mostly contain sharp but straight edges and strong consistencies in each predicted planar surface; and the inner-outer corner representation can successfully give the detection for the two kinds of keypoints in the layout. The detected layout edges are as impressive as the planar segmentation as they all produced by the same multi-task network.

For the evaluation on the estimated 3D room layout, we visualize the transformed cuboids along with the re-projected results in Fig. 7, in which we can see that the 3D layout estimation results are quite good only from a single image.

Table 2. The performance benchmarking on Hedau testing set.

Method	Pixel error (%)
Hedau [6]	21.20
Mallya [17]	12.83
DeLay [4]	9.73
CFILE [19]	8.67
RoomNet [13] recurrent 3-tier	8.36
Zhang [25]	12.70
ST-PIO [26]	**6.60**
DeepRoom (*ours*) 2D	**7.41**
DeepRoom (*ours*) 3D re-projected	9.97



(a) (b)

Fig. 7. (a) Some layout estimation results of the proposed multi-task network. (b) The representations for 3D cuboid and re-projected layout on LSUN Room.

In addition, we can observe that our model can be applied to different indoor datasets even without re-training. Table 2 shows that the accuracy of our model can almost achieve the state-of-the-art result.

5.3 Time Efficiency

Though our result is not overally the best for the 2D layout estimation metrics in the aforementioned two datasets, however, the most competitive advantage of our work is its computational efficiency since it is an end-to-end system without any optimization process or post-processing.

We implement our approach with PyTorch and perform all the experiments on the machine with single NVIDIA GeForce 1080 GPU and Intel i7-7700K 4.20 GHz CPU. For the analysis of time efficiency, Table 3 shows the consuming time for both network forwarding and post-processing time of the layout estimation methods. Although we cannot find fully released implementations of these papers, the listed entries in the column of the post-processing come from the official papers and cited ones, or the information from their released demo video. For the time consuming in the network forwarding column, several methods released their network configuration file for Caffe, and thus we can measure the time with official Caffe profiling tool and evaluate on our own machine under a fair competition.

Table 3. Comparison of time efficiency of the layout estimation methods in forwarding time and post-processing time (unit: seconds).

Method	Forward	Post-process	FPS
DeLay [4]	0.125	About 30	0.01
CFILE [19]	0.060	–	–
RoomNet [13]	0.168	–	5.96
Zhang [25]	–	About 180	–
ST-PIO [26]	0.067	About 10	0.1
LayoutNet [27]	0.039	0	25.64
DeepRoom (2D)	0.027	0	**36.95**
DeepRoom (2D/3D)	0.032	0	**31.25**

6 Conclusions

We proposed an end-to-end framework that is composed of two explainable networks for decoupling the 3D layout estimation task into two sub-tasks. They can also be jointly used to estimate the 3D cuboid representation of the spatial layout for the indoor scene. To the best of our knowledge, this is the first work that models the layout estimation as a two-stage deep learning forwarding pipeline instead of the conventional systems with an additional post-processing or optimization step. Furthermore, the combination of the two networks relies on the intermediate representation and it makes our framework pipeline open to the extensibility with using extra datasets for training and fine-tuning to achieve better outcomes.

References

1. Arth, C., Reitmayr, G., Schmalstieg, D.: Full 6DOF pose estimation from geolocated images. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7726, pp. 705–717. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37431-9_54
2. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol. Rev. **94**(2), 115 (1987)
3. Coughlan, J., Yuille, A.: Manhattan world: compass direction from a single image by bayesian inference. In: Proceedings of the Seventh IEEE International Conference on Computer Vision (1999). https://doi.org/10.1109/ICCV.1999.790349
4. Dasgupta, S., Fang, K., Chen, K., Savarese, S.: DeLay: robust spatial layout estimation for cluttered indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 616–624 (2016)
5. Gupta, A., Hebert, M., Kanade, T., Blei, D.M.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: Advances in Neural Information Processing Systems, pp. 1288–1296 (2010)
6. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: 2009 IEEE 12th International Conference on Computer vision, pp. 1849–1856. IEEE (2009)

7. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: 2005 Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 654–661. IEEE (2005)
8. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. Int. J. Comput. Vis. **75**(1), 151–172 (2007)
9. Hoiem, D., Efros, A.A., Kanade, T.: Seeing the world behind the image: spatial layout for 3D scene understanding (2007)
10. Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocalization. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 4762–4769. IEEE (2016)
11. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: Proceedings of the CVPR, vol. 3, p. 8 (2017)
12. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2938–2946. IEEE (2015)
13. Lee, C.Y., Badrinarayanan, V., Malisiewicz, T., Rabinovich, A.: RoomNet: end-to-end room layout estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4875–4884. IEEE (2017)
14. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2136–2143. IEEE (2009)
15. Lin, H.J., Huang, S.W., Lai, S.H., Chiang, C.K.: Indoor scene layout estimation from a single image. In: 2018 24th International Conference on Pattern Recognition (ICPR) (2018)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. Mallya, A., Lazebnik, S.: Learning informative edge maps for indoor scene layout prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 936–944 (2015)
18. Nowozin, S., Lampert, C.H., et al.: Structured learning and prediction in computer vision. Found. Trends® Comput. Graph. Vis. **6**(3–4), 185–365 (2011)
19. Ren, Y., Li, S., Chen, C., Kuo, C.-C.J.: A coarse-to-fine indoor layout estimation (CFILE) method. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10115, pp. 36–51. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54193-8_3
20. Roberts, L.G.: Machine perception of three-dimensional solids. Ph.D. thesis, Massachusetts Institute of Technology (1963)
21. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems, pp. 1161–1168 (2006)
22. Schwing, A.G., Urtasun, R.: Efficient exact inference for 3D indoor scene understanding. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 299–313. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_22
23. Princeton University: LSUN room layout estimation dataset (2015). http://lsun.cs.princeton.edu/. Accessed 30 Nov 2017
24. Waltz, D.: Understanding line drawings of scenes with shadows. In: Winston, P.H. (ed.) The Psychology of Computer Vision (1975)
25. Zhang, W., Zhang, W., Liu, K., Gu, J.: Learning to predict high-quality edge maps for room layout estimation. IEEE Trans. Multimed. **19**(5), 935–943 (2017)

26. Zhao, H., Lu, M., Yao, A., Guo, Y., Chen, Y., Zhang, L.: Physics inspired optimization on semantic transfer features: an alternative method for room layout estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
27. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: LayoutNet: reconstructing the 3D room layout from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2051–2059 (2018)