2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Indoor Scene Layout Estimation from a Single Image

Hung Jin Lin*, Sheng-Wei Huang*, Shang-Hong Lai* and Chen-Kuo Chiang†
*Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
†Dept. of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan
Email: vtsh.jn@gmail.com, mlm4590027@gmail.com, lai@cs.nthu.edu.tw, ckchiang@cs.ccu.edu.tw

***Abstract*—With the popularity of the hand devices and intelligent agents, many aimed to explore machine's potential in interacting with reality. Scene understanding, among the many facets of reality interaction, has gained much attention for its relevance in applications such as augmented reality (AR). Scene understanding can be partitioned into several subtasks (i.e., layout estimation, scene classification, saliency prediction, etc). In this paper, we propose a deep learning-based approach for estimating the layout of a given indoor image in real-time. Our method consists of a deep fully convolutional network, a novel layout-degeneration augmentation method, and a new training pipeline which integrate an adaptive edge penalty and smoothness terms into the training process. Unlike previous deep learning-based methods that depend on post-processing refinement (e.g., proposal ranking and optimization), our method motivates the generalization ability of the network and the smoothness of estimated layout edges without deploying post-processing techniques. Moreover, the proposed approach is time-efficient since it only takes the model one forward pass to render accurate layouts. We evaluate our method on LSUN Room Layout and Hedau dataset and obtain estimation results comparable with the state-of-the-art methods.**

I. Introduction

Recent demands for reality interaction originated from associated applications have brought researchers' recognition for scene understanding to new height. Through recognizing scene structure with algorithms, we can easily interact with the environment and provide crucial information for tasks like intelligent home, augmented reality, and robot navigation. Although the research on 3D scene understanding dates back to 1960s' simple Block World assumption [1], with the vision of reconstructing the global indoor layout with local evidences, it has become one of the most pivotal research area in the era of artificial intelligent and deep learning.

In early computer vision methods, a variety of indoor scene estimation methods have been proposed. Line segment extraction, which was the basis for layout estimation, [1], [2] was achieved through volume form like geons [3], edge feature extraction [4], superpixels [5] and segments [6]. However, these methods mostly failed in cases where room structures are occluded by objects. Since it is very difficult to estimate the layout with only local evidences, some researched on inferring estimations through hyper volumetric reasoning [7], [8], [9]. Later on, with the rise of machine learning, structured learning [10] has been developed for layout estimation, and its goal is

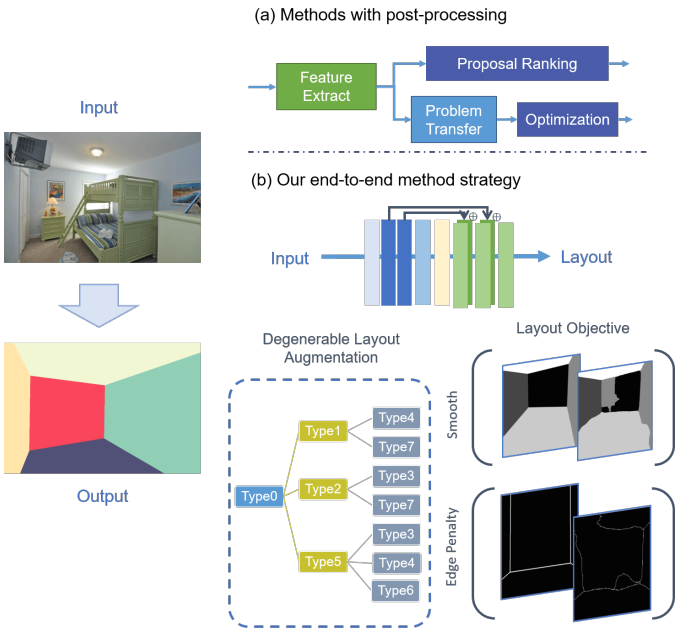

Fig. 1. Overview of the proposed planar semantic layout method and training strategy. Previous methods shown in (a); our end-to-end approach in (b).

to model the environment structure by generating hypotheses with incomplete low-level local features [7], [11], [8].

The said traditional structure inference methods heavily relied on hand-craft feature extraction and were developed under numerous assumptions. Like previous works, we develop our method under the Manhattan world assumption [12], in which we consider a room to be composed of orthogonal planes and taking room layout as the cuboid structure. From a different perspective, layout estimation can also be regarded as a region segmentation problem on each surface of cuboid.

Recently, deep learning approaches have outperformed traditional methods in several computer vision tasks, including semantic segmentation. Long *et al.* [13] proposed the first end-to-end supervised FCN (fully convolutional network) model for object semantic segmentation and reached state-of-the-art performance. Some thus adopted the perspective of seeing layout estimation as a task of critical line detection, for instance the estimation of informative edges in [11] and coarse layout prediction in [14]. Mallya et al.[11] was the first to apply

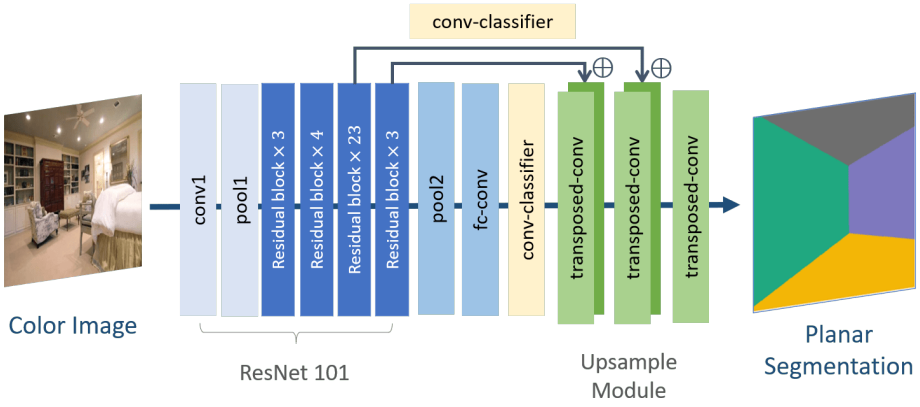

Fig. 2. Full view of our end-to-end layout estimation pipeline of ResNet101-FCN network.

the segmentation network to the layout estimation task by detecting informative edges. Dasgupta *et al.* [15], the winner of LSUN Room Layout Challenge 2015, tackled the task with a two-stage framework: 1) segment the planes and walls of the input image with a deep neural network, 2) optimize the output with vanishing point estimation. The promising result of Dasqupta *et, al* inspired several subsequent works [16], [17] to follow their two-stage pipeline, which consists of a FCN-like network for semantic segmentation and a layout optimization technique (e.g., layout hypotheses proposal-ranking pipeline in [11], [15], [14], [16], and special optimization modules in [17]) for post-processing.

This is not to deny the astonishing results achieved by the aforementioned methods, however, the extra time consumption of post-process techniques rendered these methods unsuitable for applications where time efficient plays a crucial role. We thus propose a single-stage pipeline as a solution to the time consumption-based problem. The main contributions of this paper are listed as follows,

- We propose a single-stage pipeline to train an end-to-end neural network for indoor layout estimation.
- We propose a novel layout structure degeneration method to augment data and compensate the imbalanced distribution problem in the existing dataset.
- Our method can infer the spatial layout with only one network and provide the state-of-the-art results in real-time without any post-processing.

II. Proposed Approach

A. Overview

In this section we first provide detail description of the network design and the tailored training criterion in which edge/surface information are integrated in dense pixel-wise prediction process. An additional training strategy (i.e., layout structure degeneration), designed to alleviate the data imbalance problem, will be covered in the second segment.

B. Planar Semantic Segmentation

Under the Manhattan world assumption [12], we can consider each scene is composed of multiple planar segments and hence there are limited types of layouts captured at different viewpoints. We refer these layouts to planar semantic representations as described in [15], and the planes are labeled as frontal-wall, right-wall, left-wall, ceiling, and floor. We can thus model the spatial layout estimation problem as surface labeling or planar semantic segmentation.

Our model design is inspired by the recent works in layout estimation with FCN. Ren et al.[14] use the similar configuration of original FCN, VGG-16 as the base network, to predict coarse layout and semantic surface; while[15] apply dilated-convolutional version of FCN with CRF refinement, and[17] resort to deeper network with dilated-convolutional version of ResNet101 that pre-trained on large datasets for semantic segmentation and then fine-tuned to transfer the semantic features for layout estimation for much better results.

Thus, we choose the deeper network, the vanilla ResNet101, as our feature extractor and adopt the layout representation proposed in DeLay *et al.* [15] in which the layout estimation can be regarded as a five-class planar semantic segmentation problem. The full view of our network is shown in Figure 2. We replace the last average-pooling layer in original ResNet101 with max-pooling and replace the fully-connected layer with 1×1 convolutional layer, and append three following transposed convolutional layers to upsample the feature maps with skipped connection from previous layer. To make the consistent in dimension of feature maps, the additional convolutional layers are inserted before forward to upsample module. Moreover, for the trunk in network, there are two extra dropout and batch-norm layers before conv-classifier block to prevent over-fitting on such specific task.

C. Layout Structure Degeneration

We mainly train and evaluate our approach on LSUN Room Layout dataset. Although the performance of our model can compete with several existing methods at 9.75% error rate (the entry *Ours w/o degeneration* in Table I). The room type distribution in the existing dataset (Figure 4, shown in green bars and the third column of Table II) is very imbalanced, and we observe from the results that the room types containing

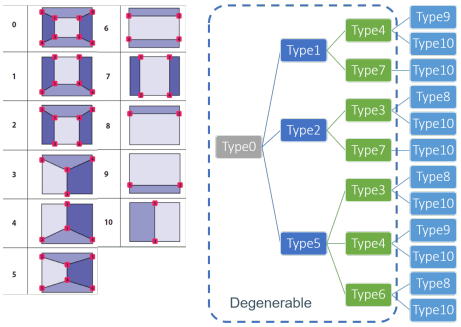

Fig. 3. The definitions of each room type and the degenerative relations graph in LSUN Room Layout dataset.

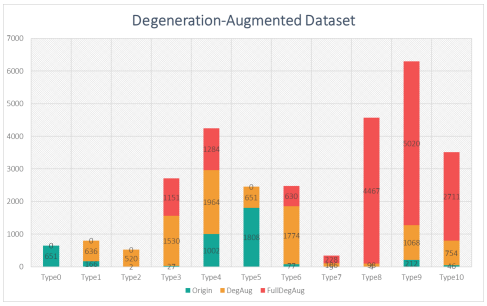

Fig. 4. The distribution of all room types in LSUN Room Layout dataset (green) and the distribution of single degeneration augmentation (orange), only degenerating one step from each node in the relation graph, and the full version of degeneration augmentation (red).

fewer training images result in higher error rates, especially for the room types composed of fewer surfaces (Table I). In other words, it means the model cannot handle these rare cases well, and the most common solution is to augment the original dataset in training phase. However, general augmentation techniques such as random cropping and random rotation are not suitable for our case in which we want to reserve the semantic meaning of surfaces, cropping may corrupt one side of the scene and corrupt the semantic relationship among left-wall, frontal-wall, right-wall. Thus, we propose the **layout structure degeneration** to generate and compensate the imbalanced distribution of different room types.

We observe that the room type with more surfaces, higher degree of surfaces, can be degenerated into lower degree of surfaces by appropriate transformations. Take the room *type0* in LSUN Room Layout for example, it can be degenerated into *type1* by removing ceiling, into *type2* by removing floor, and into *type5* by removing one of left or right wall (need relabel to reserve the left-frontal-right semantic meanings). We can accordingly build the relations among 11 room types in LSUN Room Layout as a DAG (directed acyclic graph) shown in the Figure 3, and all of the non-leaf nodes are degenerative into lower degree of surfaces.

With depth-first searching, we can enumerate the degenerative paths and thus augment the image of specific room type into other types to compensate the insufficient sample of some types. Applying layout degeneration on those room types containing fewer images, such as *type2*, *type3*, *type7* and *type8*, can make the distribution of each type much balanced. In Figure 4, yellow bars mean the number of newly generated data with only one step degeneration from degenerative nodes and the red ones mean more aggressive degeneration with more steps, and it means there will be more data for those types constructed by lower degree of surfaces with more degeneration steps.

Our proposed augmentation strategy can effectively improve the performance and generalization ability of the model by extending the scene variations of the existed dataset with more complex transformation rather than cropping or rotation. Furthermore, it can successfully reduce the error rate from 9.75% to 6.25% (the entry *Ours* in Table II), which can

TABLE I
THE PERFORMANCE ANALYSIS ON EACH ROOM TYPE OF LSUN VALIDATION SET AND ITS CORRESPONDING NUMBER OF TRAINING DATA.

Room Type	Pixel Error (%)	Training samples
type0	9.34	651
type1	17.95	166
type2	-	2
type3	16.47	27
type4	9.01	1002
type5	6.17	1808
type6	25.26	77
type7	-	5
type8	-	4
type9	24.85	212
type10	31.13	46

compete with the state-of-the-art methods, by just using twice or fourth times extra augmented data samples (yellow bars and red bars in Figure 4, respectively) in training. In contrast to plainly applying random augmentation for tens of times training data in previous methods, we can achieve better performance with much fewer data in training.

D. Layout Criterion

Semantic segmentation task is a pixel-wise classification problem, and the original objective function is the cross-entropy loss $\mathcal{L}_{seg}$ on every pixel. From the results of planar semantic segmentation, we find that it often suffers from distortion or tears apart from the center of planes and also "wavy curves" (rather than straight lines) mentioned in [15]. Figure 6 (a) depicts an example which is far away from boxy projection onto 2D images. Hence, we introduce the adaptive edge penalty and smoothness terms to alleviate these artifacts. With these tailored criterion, we can enforce the layout prediction results more smooth and much straight on the edges, and thus get the better qualitative layout estimation.

Smoothness Term: It minimizes the pixel-wise L2 distance $\mathcal{L}_{smooth}$ between ground truth label and segmented layout to enhance the consistency inside each planar plane and impose smoothness constraint on the prediction.

Adaptive Edge Penalty: The distribution of edges is often distorted and not straight enough as the cues of the layout.

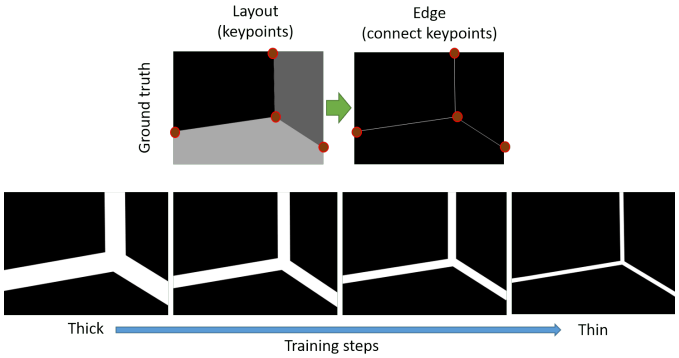

Fig. 5. Edge maps: the upper shows the generation of ground truth edge map, and the lower illustrates the maps with the adaptive edge constraint.

Thus, we calculate the edge map of the predicted layout and use the binary cross-entropy loss $\mathcal{L}_{edge}$ to minimize the difference from the ground truth. Furthermore, the criterion is adaptive by setting it loose at the beginning and tighter as training iterations increase (Figure 5). With this adaptive strategy, the results converge better than the fixed-width edge constraint.

The overall loss function for training the deep neural network model is given by

$$Loss = \mathcal{L}_{seg} + \lambda_e \mathcal{L}_{edge} + \lambda_l \mathcal{L}_{smooth},$$

and

$$\mathcal{L}_{smooth} = \|\mathcal{M}_{gt} - \mathcal{M}_{pred}\|_2,$$

$$\mathcal{L}_{edge} = BCE(\mathcal{E}_{gt}, 1 - exp(\frac{-\|grad(\mathcal{M}^*_{pred})\|}{\sigma})),$$

where M denotes the output heatmap of the network and $\mathcal{E}$ denotes the edge map, the edge map for the predicted one is generated by calculating the gradient on $\mathcal{M}^*$ denoting the final segmented layout prediction, and the lower annotations gt and $pred$ denote the ground truth and prediction output, respectively.

The results in Figure 6 (b) shows the visual effect of applying these two constraints. Although there is not much improvement in quantitative measure (decreasing error rate performance at about 1%), the overall visualization (see Figure 7) demonstrates the constraints can effectively smooth out and suppress the noisy and artificial prediction in surface of cluttered scene, as well as straighten the contours of layout from distortion.

III. Experimental Results

For the evaluation of planar semantic segmentation, we use LSUN Room Layout Estimation dataset containing 4,000 images for training, 394 images for validation, and 1,000 images for testing. During training, we apply random color jittering such as slightly change the lightness and contrast of input images to increase the diversity of scenes. Besides the layout Degeneration augmentation, we further introduce a semantic

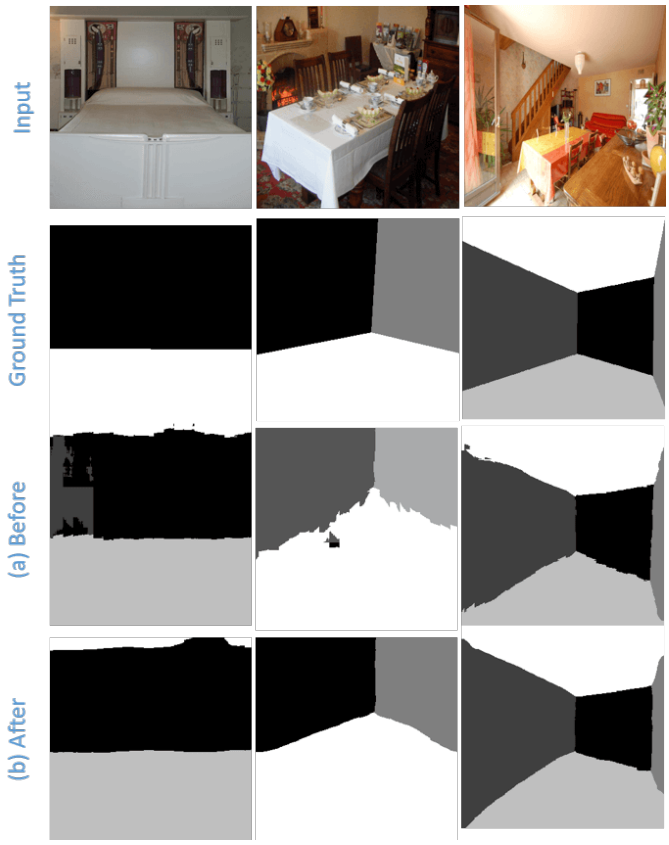

Fig. 6. Comparison of the layout estimation results for model training (a, before) without and (b, after) with edge constraint and smoothness terms.

Fig. 7. Validation results under edge constraints and smoothness criterion in LSUN Room Layout. (Left) input image, (Middle) ground truth label, and (right) predicted layout. Note that the colors between ground truth and prediction are the same, labeling range is 1 to 5 and 0 to 4 respectively.

TABLE II
THE PERFORMANCE BENCHMARKING ON LSUN ROOM LAYOUT OFFICIAL LEADERBOARD FOR DIFFERENT TECHNIQUES, MOSTLY WITH POST-PROCESSING.

Method	Pixel Error (%)	Post-process
Hedau *et al.* (2009)[7]	24.23	(*)
Mallya *et al.* (2015)[11]	16.71	proposal-ranking regression
DeLay (2016)[15]	10.63	layout optimization
CFILE (2016)[14]	7.57	proposal-ranking
Zhang (2017) *et al.* [16]	6.58	proposal-ranking
ST-PIO (2017) [17]	5.29	physical-inspired opt.
Ours w/o degeneration	9.75	*No*
Ours with degeneration	6.25	*No*

TABLE III
THE TIME EFFICIENCY OF EACH METHOD IN FORWARDING TIME AND POST-PROCESSING TIME.

Method	Forward (sec)	Post-process (sec)
DeLay (2016)[15]	0.125	about 30
CFILE (2016)[14]	0.060	(not reported)
Zhang (2017) *et al.* [16]	(not reported)	about 179
ST-PIO (2017) [17]	0.067	about 10
Ours	0.023	0

TABLE IV
THE PERFORMANCE BENCHMARKING ON HEDAU TESTING SET.

Method	Pixel Error (%)
Hedau *et al.* (2009)[7]	21.20
Mallya *et al.* (2015)[11]	12.83
Zhang (2017) *et al.* [16]	12.70
DeLay (2016)[15]	9.73
CFILE (2016)[14]	8.67
ST-PIO (2017) [17]	6.60
Ours	7.41

random horizontal flipping by exchanging the semantic labels in left and right side for more effective augmentation.

Since there are no public ground truth label for the testing set, we evaluate our method on validation set with LSUN Room Layout official toolkit like that in previous work [14]. First, we want to demonstrate the effects of our proposed criterion terms. Under the supervision of edge penalty and additional smooth term, the predictions become smoother and alleviate the artifacts, as shown in Figure 6 (b). Figure 7 depicts some layout estimation results in the validation set after adding the additional smoothness term and adaptive edge constraint into the training criterion.

We further demonstrate the improved results both quantitatively and qualitatively for the model trained under the proposed layout degeneration augmentation strategy. Figure 8 are the visual outputs under our full strategy, and they all have very sharp but straight edges and strong consistencies in each predicted surface. The accuracy of our approach is comparable to the best-performing methods on the official learderboard of LSUN Challenge and it achieves 6.25% pixel-wise error rate.

Besides the LSUN Room Layout dataset, we also evaluate the generalization capability of our model by applying directly on testing set of Hedau dataset without using its training data to fine-tune. From the performance on accuracy and visualization results, we can observe that our model can be applied to different indoor datasets even without re-training. Figure 9 depicts some examples of the high-quality layout estimation results in Hedau testing dataset, and Table IV shows that the accuracy of our model can almost achieve the state-of-the-art result.

We implement our approach with PyTorch and perform all the experiments on the machine with single NVIDIA GeForce 1080 GPU and Intel i7-7700K 4.20GHz CPU. For the analysis of time efficiency, the Table III shows the consuming time for both network forwarding and post-processing time of exited methods. Because we can't find any released full implementation of these papers, we list the official reported record from their paper or the information from their released demo video for the statistics of the post-processing column. For the time consuming in network forwarding column, several methods are implemented with Caffe and also release their network configuration file, so we can measure with Caffe official profiling tool and evaluate on our own machine in a fair competition.

Consequently, we propose a single network for layout estimation, and the model can give much impressive visual results with the proposed structure criterion constraint. Besides, the layout structure degeneration augmentation can effectively alleviate the data imbalance problem and further improve the layout estimation accuracy. Furthermore, our end-to-end network method predicts the room layout directly without any post-processing, hence it can efficiently predict the layout from an image in real-time.

IV. CONCLUSION

We propose an end-to-end deep neural network model that can estimate the room layout by enforcing the smooth constraint on edge information as well as overall smoothness. Also, we introduce a novel augmentation approach to further improve the layout estimation accuracy and the generalization capability of the model. Our real-time approach can reach the state-of-the-art without using any post-processing, we can also port the layout estimation model onto a mobile device that captures temporal images from the real world. However, we find that the temporal inconsistency of the layout estimation results is an issue to be resolved since many real-world applications require robust temporal layout estimation on a video. Hence, our future work will focus on improving the robustness of the layout estimation on a video.

ACKNOWLEDGMENTS

This research work was partly supported by Institute for Information Industry, Taiwan, and Ministry of Science and Technology, Taiwan, under the grant 107-2634-F-007 -003.

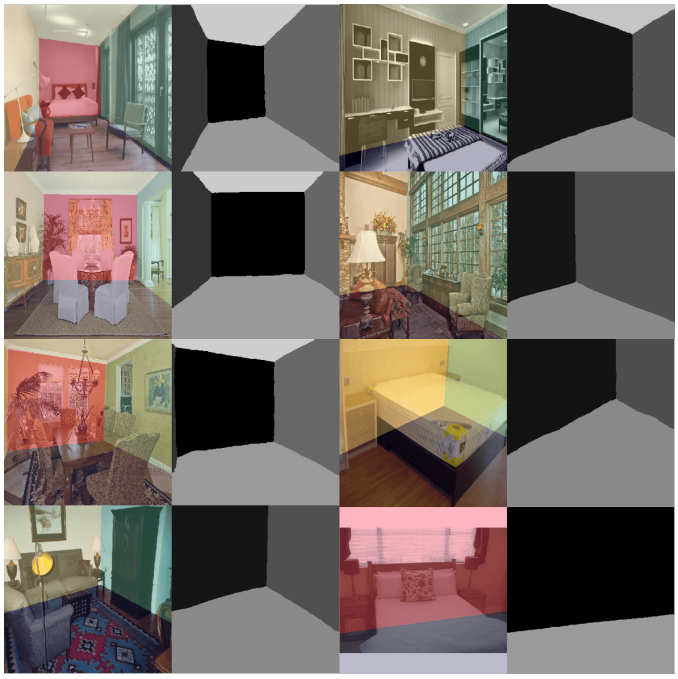

Fig. 8. The visualization results of layout estimation in LSUN Room Layout validation set. The first and the third columns are ground truth images with ground truth label masks; the second and the fourth columns are predicted layouts shown in grayscale label.

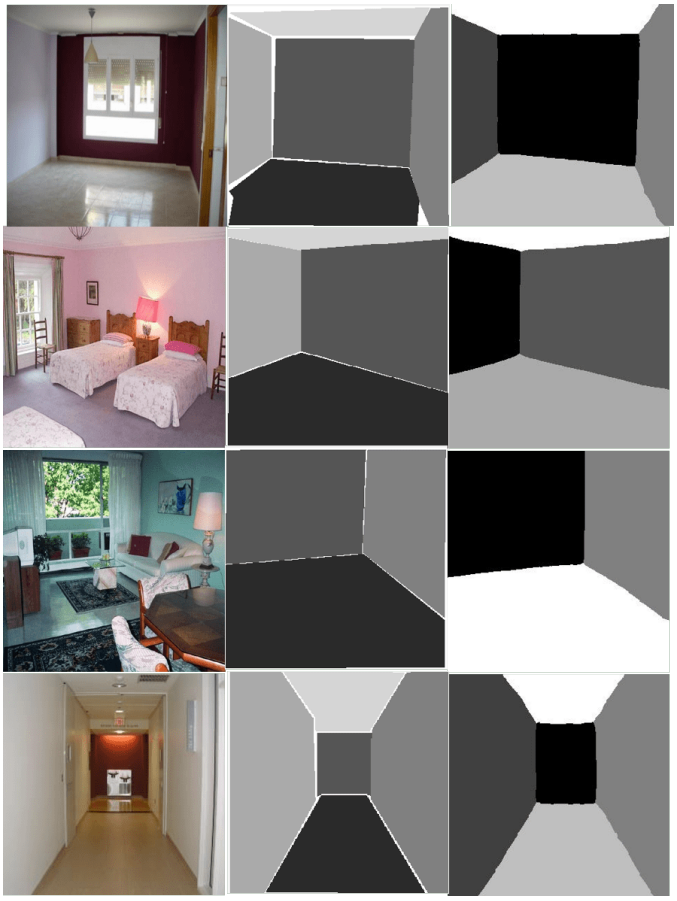

Fig. 9. The visualization results of layout estimation in Hedau testing set. (Left column) input images, (middle column) ground truth labels (white edges are borders), and (right column) prediction results. Note that the visualization labels between ground truth and prediction are not matched, one in Hedau labeling order while another is our semantic label order.

REFERENCES

[1] L. G. Roberts, "Machine perception of three-dimensional solids," Ph.D. dissertation, Massachusetts Institute of Technology, 1963.

[2] D. Waltz, "Understanding line drawings of scenes with shadows." the psychology of computer vision. patrick henry winston, ed," 1975.

[3] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, p. 115, 1987.

[4] D. C. Lee, M. Hebert, and T. 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.

[5] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Advances in neural information processing systems*, 2006, pp. 1161–1168.

[6] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 654–661.

[7] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *Computer vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 1849–1856.

[8] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei, "Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces," in *Advances in neural information processing systems*, 2010, pp. 1288–1296.

[9] A. G. Schwing and R. Urtasun, "Efficient exact inference for 3d indoor scene understanding," in *European Conference on Computer Vision*. Springer, 2012, pp. 299–313.

[10] S. Nowozin, C. H. Lampert *et al.*, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.

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

[12] J. M. Coughlan and A. L. Yuille, "The manhattan world assumption: Regularities in scene statistics which enable bayesian inference," in *Advances in Neural Information Processing Systems*, 2001, pp. 845–851.

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

[14] Y. Ren, S. Li, C. Chen, and C.-C. J. Kuo, "A coarse-to-fine indoor layout estimation (cfile) method," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 36–51.

[15] S. Dasgupta, K. Fang, K. Chen, and S. 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.

[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] H. Zhao, M. Lu, A. Yao, Y. Guo, Y. Chen, and L. Zhang, "Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.