ROBUST ROOM LAYOUT ESTIMATION FROM A SINGLE IMAGE WITH GEOMETRIC HINTS

Ruifeng Deng, Xuejin Chen

University of Science and Technology of China

ABSTRACT

Estimation of room layout suffers from heavy occlusions and clutters in indoor scenes. In this paper, we propose a deep network that combines textures and geometric hints to predict the surface layout from a single image. Our method consists of three steps. First, depths and normals are extracted from the input RGB image. Secondly, a multi-channel FCN (MC-FCN) is presented to integrate these geometric hints for semantic surface segmentation. Thirdly, an optimization framework is adopted to refine the layout estimation. The results on two commonly used benchmark datasets demonstrate the robustness of our method on complex scenes.

***Index Terms*—** Layout estimation, scene understanding, geometric hints

1. INTRODUCTION

The main purpose of room layout estimation is to extract semantic boundaries among walls, ceiling and floor from a single image. It can be realized by obtaining corresponding semantic planes, as shown in Fig. 1. The task is challenging because of complex environmental factors such as illumination variation, viewpoint shift, object diversity and clutters. Furthermore, distinctive clues like room corners and layout boundaries are often occluded by objects.

In recent years, massive researches have been carried out on room layout estimation. Traditional methods usually follow a proposing-ranking framework [1, 2, 3, 4]. Typically, they first generate numbers of proposals through vanishing point detection and ray sampling. Then, a ranking step using hand-crafted rules is adopted to select the best hypothesis. Recent methods built on FCNs [5] or encoder-decoder networks achieve impressive performance [6, 7, 8, 9, 10, 11]. Specifically, Mallya et al. [6] presented an FCN for learning informative edge maps from images; then the edge maps are integrated into the conventional framework as additional features. Dasgupta et al. [9] used an FCN to learn semantic surface labels including {Left wall, Front wall, Right wall, Ceiling, Ground} for each pixel. They obtained a segmentation based on heat maps that are generated from the FCN, and further refined it with geometric projection constraints. In a subsequent technique [7], Ren et al. adopted a multi-task fully convolutional neural network to jointly predict the surface labels and boundaries. Zhang et al. [8] proposed a deconvolution network which has multi-layer deconvolution to attain highly reliable edge maps. Lee et al. [10] explored an end-to-end approach for room layout estimation. They employed an encoder-decoder network to delineate the room layout structure using 2D keypoints. Zhao et al. [11] introduced a semantic transfer FCN to extract reliable edge features and designed a physics inspired inference scheme for optimization.

Different from previous FCN based methods, we propose to improve the performance of room layout estimation from another perspective. Inspired by the work on RGBD semantic segmentation [12, 13, 14, 15, 16], we consider global geometric information as important as textures in room layout estimation. To implicitly employ the geometric hints, depths and normals are estimated from the input RGB image and fused into a multi-channel FCN. We use the same layout formulation defined in [9], and combine their optimization step with a proposing-ranking framework to attain precise layouts. Experimental results show that our method is robust and effective for layout estimation even facing clutters and occlusions.

2. OUR METHOD

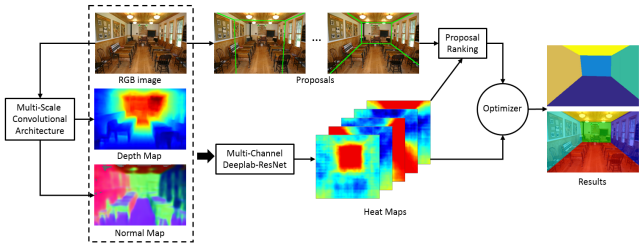

Fig. 1. The pipeline of our layout estimation algorithm. First, we adopt a multi-scale CNN architecture [17] to extract the geometric information, including depths and normals, from the RGB image. Then, we combine all the above-mentioned information into a multi-channel FCN, which helps to accurately estimate the layout. Finally, an optimization framework is employed to obtain the final layout by optimizing the best proposals generated from the input RGB image.

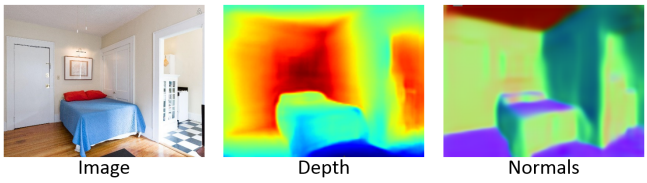

Fig. 2. Depths and normals estimated from an RGB image using the multi-scale FCN [17]. From the global perspective, the extracted geometric information captures the overall spatial relationships between different surfaces.

Under the Manhattan world assumption [18], a room layout can be represented as a cube, of which at most five surfaces {Left, Front, Right, Ceiling, Ground} are visible in an image. Given an RGB image $\mathbf{I}$ with an arbitrary size, our algorithm generates a room layout $\mathbf{L}$ that has a surface label for each pixel L_{ij} in such 5-class set. Fig. 1 shows the algorithm pipeline. We first estimate a depth map D_I and a normal map N_I from the input image to generate geometric hints, using a multi-scale convolutional network [17], as described in Sec. 2.1. After that, to integrate the information from the input RGB image together with the geometric hints, a multi-channel fully convolutional network (MC-FCN) is trained. The MC-FCN is applied to predict five probability maps, each of which describes the probability of a pixel belonging to a specific layout surface. Details can be found in Sec. 2.2. While we can easily get a layout estimation $\hat{\mathbf{L}}$ by choosing the label with the highest score among the five probability maps for each pixel, the estimated $\hat{\mathbf{L}}$ always has wavy boundaries and multiple disjoint connected regions, due to the characteristics of neural networks. To handle these problems and obtain clearer layout estimation with geometric constraints, an optimization step is adopted, as described in Sec. 2.3.

2.1. Geometric Hints Extraction

We use the multi-scale convolutional network proposed in [17] to estimate the depth and normal maps from RGB images. An example is shown in Fig. 2. Note that we do not finetune their models with our data as we do not have the depth and normal data in our dataset. However, the model seems to generalize well in our case.

We notice in practice that the depths and normals estimated from RGB images are not precise in local details. And the output resolution of the depth and normal maps is limited to half of the input. However, the valuable 3D information they provided is much helpful for high-level structure estimation, especially in messy scenes. As analyzed in Sec. 3.1, depths and normals can serve as hints that tend to merge big planes together under the interference factors, such as clutters and occlusions. Therefore, we apply these mid-level geometric hints to improve the performance of layout estimation.

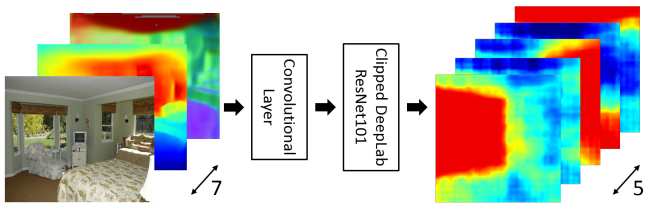

Fig. 3. The multi-channel network architecture build on [19]. We clip the Deeplab-ResNet101 by removing the multi-scale related layers. Then, we combine the RGB image, depths and normals together as input to the network for semantic surface learning.

The depth maps are color coded to 3-channels in [17] for distinction. In order to ensure the essence of depth and to reduce the redundant information, we modify their rendering section to obtain depth maps with a single channel, which is then fed to the following multi-channel FCN.

2.2. Semantic Surface Segmentation Using MC-FCN

In prior techniques, FCNs are widely used for semantic surface segmentation [9, 7, 6]. However, due to clutters, occlusions, complex textures and illumination variations in indoor scenes, the surface segmentation represented as $\hat{\mathbf{L}}$ is sometimes ruined by irregular spurious regions. Fig. 5 demonstrates several typical inferior cases. To get rid of these annoying spurious regions, we attempt to make our network more robust to miscellaneous environmental factors. A multi-channel FCN (MC-FCN) is adopted to incorporate geometric hints that are estimated from the RGB image.

Network Architecture. We build the MC-FCN based on the DeepLab-ResNet101 [19]. While we consider the room layout as a high-level semantic information that relies more on global structure, we modify the network by removing the multi-scale related layers. As illustrated in Fig. 3, the estimated depth and normal maps are treated as additional channels associated with the input RGB image. We simply concatenate different types of data to a seven-channels input (3 channels from RGB, 1 channel from depths and 3 channels from normals), and feed them into the network. Atrous convolution is adopted in our 5-way classifier layer.

Training. We initialize the MC-FCN parameters with a model pre-trained on the SUNRGBD dataset, except for the classifier layer. After that, the MC-FCN is specifically finetuned for semantic surface segmentation. The output of our MC-FCN is a $w \times h \times 5$ probability array $\mathbf{T}$, where w and h stand for the width and length of the input image. Each of the 5 slices can be viewd as a probability map for the corresponding surface label, as shown in Fig. 3. We utilize the probability array $\mathbf{T}$ as the basis for our scoring function in the proposal ranking and optimization step.

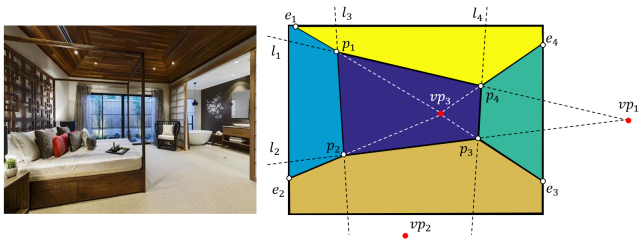

Fig. 4. Illustration of the parameterized model τ. The vanishing points vp_1 and vp_2 are applied for extracting proposals.

2.3. Layout Optimization

We adopt a widely used model [1, 2, 9, 7] to parameterize an indoor layout based on the Manhattan world assumption. A scene layout is modeled as the projection of a cuboid, which can be defined by

$$\tau = (l_1, l_2, l_3, l_4, vp_3), \tag{1}$$

where l_i stands for the i^{th} line and vp_3 stands for the vanishing point near the image center, as illustrated in Fig. 4. Dasgupta et al. [9] proposed an iterative refinement algorithm to obtain a precise layout. They directly take $\hat{\mathbf{L}}$ as initialization and process it to address spurious regions and multiple disjoint components. However, as mentioned above, the initial $\hat{\mathbf{L}}$ generated from $\mathbf{T}$ is often too ambiguous due to the characteristics of neural networks. The preprocessing step can not adapt to all the misleading situations, especially when the sizes of ambiguous walls are similar.

To make the optimization method more general and efficient, we simplify the refinement algorithm in [9] by replacing the initialization and preprocessing step with a proposing-ranking framework. The framework directly generates the initial $\mathbf{L}^*$ consistent with the formulation in Eq. 1. This kind of $\mathbf{L}^*$ is much robust to ambiguous cases and easier for the optimization step to find the best layout.

Our proposing-ranking framework is similar with [1]. The score function is constructed based on the surface prediction $\mathbf{T}$. At first, straight lines are found and grouped to detect three vanishing points. Then, we sample 30 rays from vp_1 and vp_2 respectively to acquire numerous proposals, each of which is specified by two rays from each vanishing point. Finally, for any given proposal $\overline{\mathbf{L}}$ with r semantic surfaces, we define the score function as:

$$S(\overline{\mathbf{L}} \mid \mathbf{T}) = \frac{1}{w \times h} \sum_{r} \mathbf{T}_r^{(\overline{L}_r)}, \tag{2}$$

where r represents the pixels in a certain region for the corresponding semantic surface. Intuitively, the proposal with higher score is more similar to $\hat{\mathbf{L}}$ and also conforms to geometric projection constraints. We select the proposal with the highest score as initial $\mathbf{L}^*$, namely:

$$\mathbf{L}^* = \arg\max_{\overline{\mathbf{L}}} S(\overline{\mathbf{L}} \mid \mathbf{T}). \tag{3}$$

After the proposing-ranking framework, $\mathbf{L}^*$ is further refined to attain precise $\mathbf{L}$ using the optimization step in [9].

3. RESULTS

We prove the validity of the geometric hints in semantic surface segmentation, and evaluate our method on two widely used datasets: LSUN dataset [20] and Hedau dataset [1].

3.1. Analysis of Geometric Hints

To explore the benefits of geometric hints for semantic surface segmentation, we train FCNs with or without geometric hints, respectively. Their performance are evaluated on the LSUN validation set using pixel error of $\hat{\mathbf{L}}$, which can be seen in Table 1. To make fair comparisons with [7, 9], both of which are built on VGG16, we train an MC-FCN based on VGG16 too. For [7], we directly apply their trained model to generate $\hat{\mathbf{L}}$. And for [9], we train an FCN having the same architecture with [9]. As revealed by Table 1, with the help of geometric hints, our MC-FCN obtains lower pixel error of $\hat{\mathbf{L}}$. We further improve the performance by employing ResNet101 [21] in our FCN. Qualitative results are demonstrated in Fig. 5. Fig. 5 (a)-(e) show typical good examples. Intuitively, the traditional FCNs are sometimes confused by different surfaces and generate spurious regions due to occlusions and clutters. Our method with the geometric hints gets rid of these uncertainties, and attains more accurate semantic surface segmentation. Fig. 5 (f) depicts an imperfect result, where FCNs tend to be cheated by large truncated wall-like object surfaces, e.g., the bed looks like the floor. This may be the result of ignoring the semantics of the objects. Fig. 5 (g) is an another challenging case due to clutters, while our MC-FCN produces a relatively clearer result.

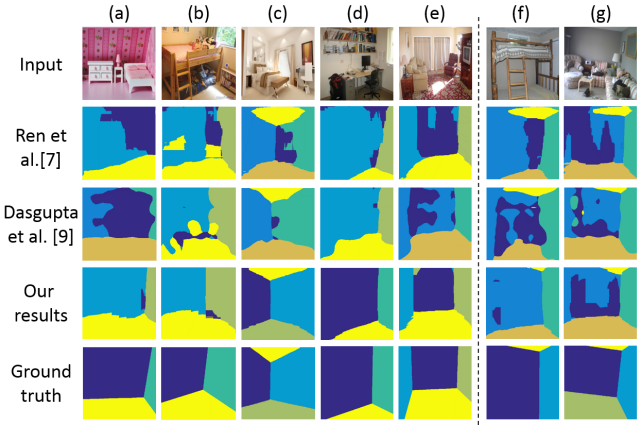

Fig. 5. Surface segmentation results using different methods. All the networks are built on the VGG16 architecture. Our MC-FCN with geometric hints generates more accurate segmentation, especially for complex environments.

Network	ϵ_{pixel} (%)
Ren et al. [7]	21.54
Dasgupta et al. [9]	15.86
MC-FCN (VGG16)	14.05
MC-FCN (ResNet101)	11.45

Table 1. Pixel-wise errors of semantic surface segmentation by different FCNs. By utilizing geometric hints, our proposed MC-FCNs obtain more accurate segmentations.

3.2. Results on LSUN Dataset

We train our multi-channel FCN on the relabeled LSUN dataset released by [7]. The dataset consists of 4000 training images, 394 validation images, and 1000 testing images. We extract geometric hints from original images and resize all the images, depth and normal maps to 321×321 using bicubic interpolation. These three types of data are integrated to train the ResNet101-based multi-channel FCN. We evaluate our results using the official toolkit which provides two standard metrics: pixel-wise error and corner error. The pixel-wise error is computed by counting the percentage of pixels that are mismatched. The corner error is computed by calculating the Euclidean distance between predicted corners and corresponding ground truth corners.

We summarize the performance on the test set of LSUN in Table 2. Qualitative results are displayed in Fig. 6. Our approach outperforms traditional methods [1] and most neural network based methods [6, 8, 9, 7, 10] on both metrics. Besides, when training the FCN, Zhao et al. [11] formulate the room layout using boundaries among semantic surfaces, while we formulate the room layout using the five semantic surfaces, as [9] does. These two representations may have different application prospects in the future. For example, when a home robot wants to locate itself while it moves, the semantic boundaries are not always available. Our method may also be an inspiration for other surface detection problems.

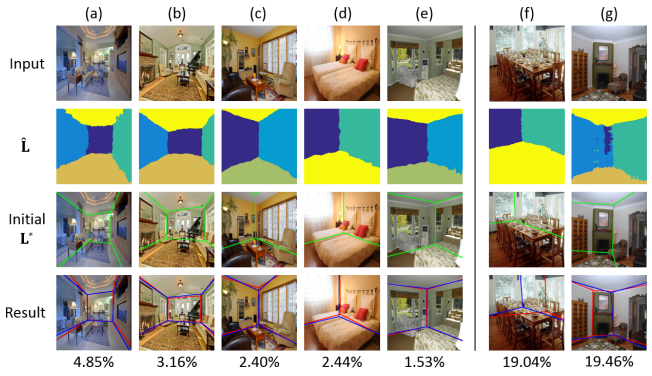

Fig. 6. Qualitative results and their pixel error of our method on LSUN validation set. Ground truth is shown in red. (a)-(e) depict precise results. (f)(g) show failure cases misled by $\hat{\mathbf{L}}$.

Method	ϵ_{corner} (%)	ϵ_{pixel} (%)
Hedau et al. [1]	15.48	24.23
Mallya et al. [6]	11.02	16.71
Zhang et al. [8]	8.70	12.49
Dasgupta et al. [9]	8.20	10.63
Ren et al. [7]	7.95	9.31
Lee et al. [10]	6.30	9.86
Zhao et al. [11]	3.84	5.29
Proposed MC-FCN	4.98	6.91

Table 2. Performance on the LSUN dataset [20].

Method	ϵ_{pixel} (%)
Hedau et al. [1]	21.20
Mallya et al. [6]	12.83
Zhang et al. [8]	12.70
Dasgupta et al. [9]	9.73
Ren et al. [7]	8.67
Lee et al. [10]	8.36
Zhao et al. [11]	6.60
Proposed MC-FCN	**6.07**

Table 3. Performance on the Hedau dataset.

3.3. Results on Hedau Dataset

We also conduct experiments on the Hedau dataset [1], which consists of 209 training images and 104 testing images. Since the semantic boundaries in the ground truth are thick and treated as a new category in this dataset, we perform boundary thinning to single pixel width by assigning them with the closest segmentation label, same as [10]. We directly evaluate our method on the test set of Hedau dataset using the model trained on LSUN. Pixel-wise error is adopted as the metric. As shown by Table 3, our performance is better than [6, 8, 7], all of which are trained on the Hedau's training set. This indicates that our model has a good generalization ability. Our method achieves impressive performance on this dataset.

4. CONCLUSION

In this paper, we propose to utilize geometric hints with an RGB image for room layout estimation. We estimate the depths and normals from the input RGB image, and fuse them in a multi-channel FCN. We demonstrate how the geometric hints improve the performance of surface segmentation. Then, we incorporate a proposing-ranking strategy to optimize the estimated room layout. Our method is robust to complex environmental factors for indoor scenes.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grants. 61472377, 61632006, 61331017.

5. REFERENCES

[1] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 1849–1856.

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

[3] 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.

[4] V. Hedau, D. Hoiem, and D. Forsyth, "Thinking inside the box: Using appearance models and context based on room geometry," in *European Conference on Computer Vision*. Springer, 2010, pp. 224–237.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] C.-Y. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich, "Roomnet: End-to-end room layout estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[11] 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.

[12] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[13] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[14] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1262–1266.

[15] G. Pagnutti, L. Minto, and P. Zanuttigh, "Segmentation and semantic labelling of rgbd data with convolutional neural networks and surface fitting," *IET Computer Vision*, vol. 11, no. 8, pp. 633–642, 2017.

[16] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5199–5208.

[17] 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, pp. 2650–2658.

[18] J. M. Coughlan and A. L. Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *Proceedings of the IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 941–947.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[20] Y. Zhang, F. Yu, S. Song, P. Xu, A. Seff, and J. Xiao, "Large-scale scene understanding challenge: Room layout estimation," *accessed on Sep*, vol. 15, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.