ROBUST ROOM LAYOUT ESTIMATION FROM A SINGLE IMAGE WITH GEOMETRIC HINT

Ruifeng Deng, Xuejin Chen

University of Science and Technology of China

ABSTRACT

Estimation of room layout suffers from heavy occlusion and clutter 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: (1) depth and normals are extracted from the input RGB image; (2) a multi-channel FCN (MC-FCN) is presented to integrate these geometric hints for semantic surface segmentation; (3) an optimization framework is adopted to refine the layout estimation. The proposed method has proven to be more robust to complex environmental factors. We achieve competitive performance on two commonly used benchmark datasets.

Index Terms— Layout estimation, scene understanding, geometric hint

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 clutter. 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. Conventional 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, which were then 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 depended on heat maps from 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 (MFCN) to jointly predict the surface labels and boundaries. Benefit from joint training, the semantic boundaries are more accurate. They also optimized the results later with a refinement framework designed for boundaries. Zhang et al. [8] proposed an another deconvolution network which has multi-layer deconvolution and a receptive field as large as the entire image compared to FCN. As a result, they attained highly reliable edge maps. An end-to-end approach for room layout estimation was explored by Lee et al. [10]. 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. They first trained an FCN for 37-class semantic segmentation and then bridged the gap to 4-class edge labels by adding a fully connected layer. A physics inspired inference scheme was designed for optimization.

Different from previous FCN based methods, we propose to improve the performance of room layout estimation from another perspective. We consider global geometric information as important as textures in room layout estimation. To implicitly employ the geometric hint, depth and normals are estimated from the original image and fused into an multi-channel FCN. We use the same formulation of layout in [9] and combine their optimization step with a proposing-ranking framework to attain precise layout. Experimental results demonstrate that our method is robust and effective for layout estimation even facing clutter and occlusion.

2. OUR METHOD

Under the Manhattan world assumption [13], a room layout can be represented as a cube having at most five surfaces {Left, Front, Right, Ceiling, Ground} visible in an image. Given an RGB image $\mathbf{I}$ with 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 our algorithm pipeline. We first estimate a depth map D_I and a normal map N_I from the input image to generate geometric hint using a multi-scale convolutional architecture [12], as described in Sec. 2.1. After that, to integrate the information from the input RGB image together with the geometric hint, a multi-channel fully convolutional network (MC-FCN) is trained. The MC-FCN is

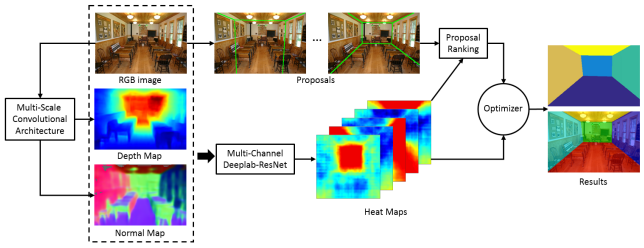

Fig. 1. The pipeline of our layout estimation algorithm. First we adopt a multi-scale CNN achitecture [12] to extract geometric information from RGB images, including depth and normals. Then we combine all the abovementioned information into a multi-channel FCN , which helps to accurately estimate the layout. Finally, an optimization framwork is employed to obtain final precise layout by optimizing the best proposals generated from RGB images.

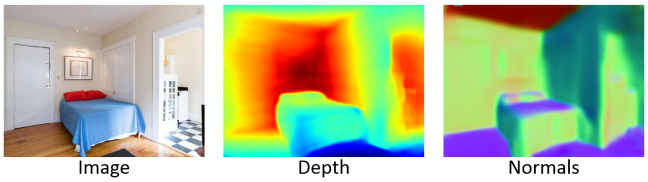

Fig. 2. Depth and normals estimated from RGB images using the multi-scale FCN in [12]. From the global perspective, these geometric information can capture the overall spatial relationships between semantic surfaces.

applied to predict five probability maps, each of which describes the probability of a pixel belonging a specific layout surface, see details in Sec. 2.2. While we can easily get access to a layout estimation $\hat{\mathbf{L}}$ by just choosing the label with the highest score across five probability maps for each pixel. The estimated $\hat{\mathbf{L}}$ always have wavy boundaries and multiple disjoint connected regions due to the characteristics of neural network. To handle these problems and obtain more clear layout estimation with geometric constraints, an optimization step is adopted in Sec. 2.3.

2.1. Geometric Hint Extraction

We use the multi-scale convolutional network proposed in [12] to estimate the depth and normal maps from RGB images, a qualitative result is shown in Figure 2. Note that we do not finetune their models with our data as we do not have the depth and normals of our dataset, however, the model seems to generalize well in our case. Quantitative evaluation of their work can be referred to [12].

We notice in practice that the depth and normals estimated from RGB images are not precise in local details. And the output resolution of the depth and normal maps are limited to half of the input. However, the valuable 3D information

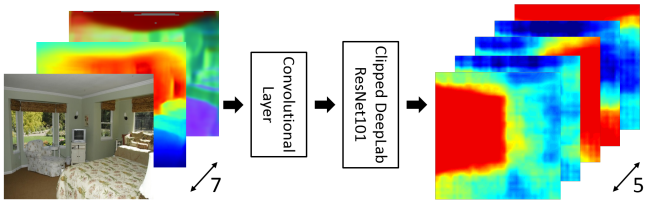

Fig. 3. Illustration of the multi-channel network architecture build on [14]. We clip the Deeplab-ResNet101 by removing the multi-scale related layers. Then we combine the RGB image, depth and normals together as input to the network for semantic surface learning.

they provide is much helpful for high-level structure estimation, especially in messy scenes. As depicted in Fig. 5, depth and normals can serve as hints tending to merge big planes together under the interference factors like clutter and occlusion. Therefore, we can apply these mid-level geometric hints to improve the performance of layout estimation.

The depth maps are color coded to 3-channels in [12] 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.

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 the clutter, occlusion, complex textures and illumination variations in indoor scenes, the surface segmentation represented as $\hat{\mathbf{L}}$ are sometimes ruined by irregular spurious regions. Fig. 5 demonstrates several typical bad cases. To get rid of these annoying spurious regions, we attempt to make our network more robust to miscellaneous environmental factors. And for this purpose, a multi-channel FCN (MC-FCN) is adopted to incorporate geometric hint generated from the original image.

Network Architecture. We build the MC-FCN based on the DeepLab-ResNet101 [14], which is designed to be multi-scale for semantic segmentation. While we consider the room layout as a high-level semantic information that relies more on global information, we modify the network by clipping 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 depth and 3 channels from normals) which is then fed into our network. Atrous convolution is adopted in our 5-way classifier layer.

Training. We initialize the MC-FCN weights with a model pretrained on SUNRGBD dataset except for the classifier layer. Then 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

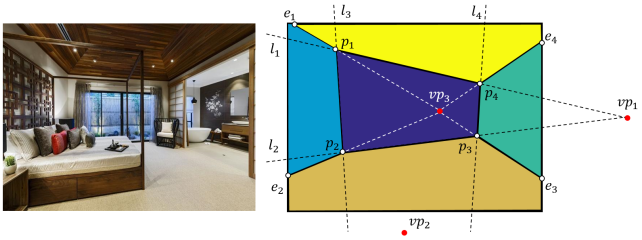

Fig. 4. Illustration of the parameterized model τ. The vp_1 and vp_2 are applied for extracting proposals.

width and length of the input image. Each of the 5 slices can be viewd as a probability map for the corresponding surface label, see Fig. 3. We utilize the probability array $\mathbf{T}$ as the basis for our scoring function in both proposals ranking and optimization step.

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. Indoor 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 (b). 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 usually 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 implemented using the proposing approach in [1] and constructing a score function based on $\mathbf{T}$. First, We sample 30 rays from vp_1 and vp_2 separately to acquire numerous proposals. Then, for any given proposal $\overline{\mathbf{L}}$ with r semantic surfaces, we define the score function:

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

where the subscript r represents a certain region for the corresponding semantic surface. Then 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 hint in semantic surface segmentation and evaluate our method on two widely used dataset: Hedau's dataset [1] and LSUN dataset [15].

3.1. Analysis of Geometric Hint

To explore the benefits of geometric hint for semantic surface segmentation, we train FCNs with or without geometric hint, respectively. Their performance are evaluated on LSUN validation set using pixel error of $\hat{\mathbf{L}}$, see 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 in [9]. As revealed by Table 1, with the help of geometric hint, our MC-FCN obtains lower pixel error of $\hat{\mathbf{L}}$. We further improve the performance by employing a ResNet101 [16] based FCN. Qualitative results are demonstrated in Fig. 5. (a)-(e) show typical good examples. Intuitively, the geometric hint helps remove the spurious regions caused by clutter and generate more accurate semantic surface segmentation. (f) depicts a generally bad case that FCNs tend to be cheated by large truncated wall-like object surface, e.g., the bed looks like the floor. (g) is an another kind of bad case led by clutter while our MC-FCN produces relatively clearer results.

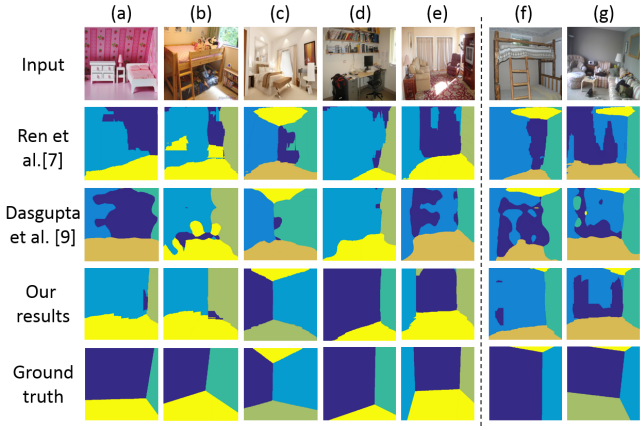

Fig. 5. Surface segmentation results using different methods. All the networks are built on the VGG16 architecture. Our MC-FCN with geometric hint generates more accurate segmentation, especially for complex environments.

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 error of semantic surface segmentation by different FCNs. By utilizing geometric hint, our proposed MC-FCNs acquire more accurate segmentation.

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	5.53	7.84

Table 2. Performance on the LSUN [15] dataset

3.2. LSUN Results

We train our multi-channel FCN on the relabeled LSUN dataset released by [7]. The dataset consists of 4000 training, 394 validation, and 1000 testing images. We extract geometric hint from original images and resize all the images, depth and normal maps to 321×321 using bicubic interpolation. Then 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: pixelwise error and corner error. The pixelwise error is computed by counting the percentage of pixels that are mismatched. The Hungarian algorithm is applied to address the labeling ambiguity problem. The corner error is computed by calculating the Euclidean distance between predicted corners and corresponding ground truth corners.

Our performance on LSUN test set compared with other methods is shown in Table 2. Qualitative results are displayed in Fig. 6. Our approach outperforms conventional methods [1] and most neural network based methods [6, 8, 9, 7, 10]. 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] do. These two representation may have different application prospects in the future. For example, when a home robot wants to locate itself while moving, the semantic boundaries are not always available. Our method may also be an inspiration for other surface detection problems.

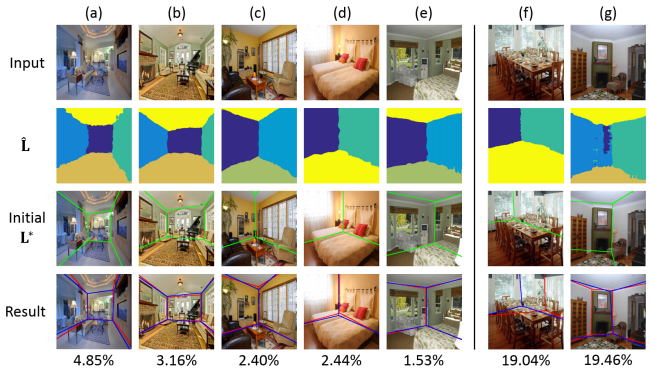

Fig. 6. Qualitative results of our methods on LSUN validation set. (a)-(e) depict precise results. (f)(g) show failure cases misled by $\hat{\mathbf{L}}$.

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	6.63

Table 3. Performance on the Hedau's [1] dataset

3.3. Hedau Results

We also conduct experiment on a relatively smaller dataset published by [1], which consists of 209 training images and 104 testing images. We follow the experimental setup in [10] and directly predict the semantic surface on Hedau's test set using the model trained on LSUN. Pixel error is adopted to evaluate the final results, see Table 3. Our performance is better than [6, 8, 7], all of which have trained their model on Hedau's training set. This indicates that our model has a good generalization ability. We achieve competitive performance on this benchmark dataset.

4. CONCLUSION

In this paper, we propose to fully utilize the geometric information from an RGB image for room layout estimation. We estimate the depth and normal maps from the input color image and then combine them together to train a multi-channel FCN. We demonstrate how these geometric hints improve the performance of semantic surface segmentation. Then we incorporate a proposing-ranking policy to generally optimize the room layout. Benefit from these techniques, our method is robust to complex environmental factors in indoor scenes.

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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.