Training and Post Processing 3D Room Layout Beyond the Manhattan World Assumption

Dongho Choi[1,2]

[1] Mindslab, Inc., Gyeonggi-do, Republic of Korea
[2] Seoul National University, Seoul, Republic of Korea
dongho.choi@snu.ac.kr

Abstract. Predicting 3D room layout from single image is a challenging work with many applications. In this paper, we propose a new training and post processing method for 3D room layout estimation, built on a recent state-of-the-art 3D room layout estimation model. Extensive experiments show that our method boosts performance, and outperforms state-of-the-art methods. Our method has obtained 3^{rd} place in Holistic Scene Structures for 3D Vision Workshop.

Keywords: Deep Learning, 3D Room Layout, Single Panoramic Image

1 Introduction

Last decade saw growing attention for recovering 3D room layout from a single image. 3D room layout can be viewed as a composition of orientation, corner position, and wall boundaries.

Various works [2, 4, 6, 12] have been developed for room layout estimation. Recent methods train deep neural networks to detect room corners and boundaries. Specifically, Xu [9] estimates room layout and pose of objects by detecting surface normal orientations. Yang and Zhang [10] predict the depth from single image to infer 3D room model. Yang [11] performs semantic segmentation of floor plan segmentation. Fernandez-Labrador [3] and Zou [14] predict the probability of boundaries and corners as 2D image, while Sun [8] predicts as 1D vector.

Most full layout estimations work assume that rooms are in Manhattan world [1] and cuboid-shaped. We propose new training and post processing method without those assumptions in the above. Our method perform better in accuracy compared with state-of-the-art methods.

2 Method

Our model is based on the architecture of HorizonNet [8] which predicts 1D layout to recover 3D layout. Given a single image I, the network predicts corner probability y_p, ceiling-wall boundary y_c and floor-wall boundary y_f.

LayoutNet2 [15] showed that HorizonNet's architecture and data augmentation methods boost performance compared to other approaches. Thus, instead of changing network topology, we propose training and post processing method for better result.

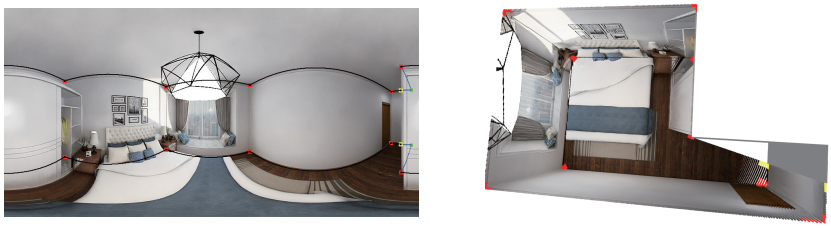

Fig. 1. Network output illustrated on 2D, 3D room image. Continuous visible corners are depicted as red points, while hidden corners as green, and discontinuous visible corners as yellow. The room image is selected from Structured3D [13]. Best viewed in color.

2.1 Loss

HorizonNet uses L1 loss for boundary prediction while other methods [3, 11, 14] use L2 loss with similar approach. L1 loss has constant gradient, making the training process unstable and non-converging. Instead, We adopt L2 loss. The gradient of L2 loss is linear, leading the network converge with many pixels with small losses. Although L2 loss makes the network to predict blurry boundary, boundaries can be predicted when exact corners are given. We try to predict corners first and later predict accurate boundaries.

We give higher weight to those corners during the first half of the training, and reverse the weight at the last half training time. The overall loss function is as follows:

$$L(y_p, y_c, y_f) = w_1(BCE(y_p, y'_p)) + w_2(L2(y_c, y'_c) + L2(y_f, y'_f)) \tag{1}$$

where y'_p, y'_c, and y'_f represent the ground truth of corner probability, ceiling-wall boundary, and floor-wall boundary respectively.

2.2 Post Processing

Without Manhattan layout assumption, hidden corners can not be accurately calculated. We focus on refining visible layouts. In Fig 1, visible boundaries are discontinuous near hidden points since those walls are not adjacent. Thus, we propose an algorithm to detect discontinuity from predicted boundaries.

Our post processing method has two approaches - one directly from 2D panoramic image and the other from generated 3D layout. We first detect candidates of discontinuity from raw y_c and y_f output. In 2D room image, hidden point induces discontinuity in y_c and y_f curves. Since the network's output is given by pixels, high slope can be detected as discontinuity. To distinguish between two candidates, we also search for big change in boundary's slope.

Table 1. Quantitative results on room layout estimation from Structured3D test set. †: evaluated with our full post processing method.

Method	2D IoU(%)	3D IoU(%)	Corner error(%)
HorizonNet[8]	91.72	90.17	0.860
HorizonNet†	91.94	90.55	0.861
Ours	89.95	88.45	0.656
Ours†	**93.50**	**92.20**	**0.639**

Table 2. Quantitative results on visible room layout estimation from Structured3D test. All models are evaluated with our post processing method.

Method	2D IoU(%)	3D IoU(%)	Corner error(%)	Pixel error(%)
HorizonNet	92.77	91.37	0.697	2.126
Ours	**94.31**	**92.99**	**0.468**	**1.340**

For 3D layout, distance from camera to vertical wall can be accurately calculated when floor and ceiling points in 2D panoramic image are given. As shown in Fig 1, hidden point induces jump in distance. We convert 2D panorama image to 3D layout, calculate distance and find distance discontinuity.

We ensemble those candidates and select both highest and lowest point for each floor and ceiling boundary at discontinuous points.

3 Experiments

3.1 Training Details

Dataset We train and evaluate our model using Structured3D [13] dataset. It consists of more than 20k panoramic images of rooms synthesized with rich details including semantic, albedo, depth, surface normal and input. We only use single RGB image and corner labels for training. We follow training, validation, and test set given from the dataset.

Environment Our model is implemented with PyTorch [7] and tested on an single NVIDIA V100 GPU. The training process used total 14 GPU days consisting of 7 days for each training process. During each half of the training, we use Adam optimizer [5] with learning rate 3e-4 and $w_1 : w_2 = 3 : 1$ for the first half, learning rate 1e-4 and $w_1 : w_2 = 1 : 3$ for the last. We train with batch size of 24 and 250 epochs for each half of the training process. Data augmentation is same as the method in HorizonNet.

3.2 Quantitative Results

Evaluations are based on following three standard metrics and three F scores:

Table 3. Ablation study on our post processing method with visible room layout. †: evaluated using 2D candidates of our method. ‡: evaluated using 3D candidates of our method.

Method	Junction	Wireframe	Plane
HorizonNet†[8]	0.8349	0.6655	0.9426
HorizonNet‡	0.8307	0.6669	0.9430
HorizonNet†‡	0.8382	0.6692	0.9430
Ours†	0.8806	0.7380	**0.9543**
Ours‡	0.8730	0.7378	**0.9543**
Ours†‡	**0.8834**	**0.7440**	0.9542

1. IoU, intersection of union between our prediction and ground truth

2. Corner error, which is the distance between predicted corners and ground truth corners, normalized by the length of image diagonal

3. Pixel error, pixel-wise semantic(ceiling, wall, floor) error between prediction and ground truth.

4. Junction, predicted corner is considered correct when prediction and nearest ground truth is within 5, 10, 20 pixels. Final score is average of three scores.

5. Wireframe, predicted boundary is considered correct if prediction and nearest ground truth is within 5, 10, 20 pixels. Final score is average of three scores.

6. Plane, predicted plane is correct when prediction and nearest ground truth plane's intersection over union is over 0.5.

Table 1, 2 show results of layout evaluation on Structured3D dataset. Table 1 shows that even without assuming cuboid layout and predicting hidden points, our post processing slightly boosts on layout including hidden points without pixel error.

We show ablation study of our post processing method in Table 3. We obtain high performance boost when using candidates from 2D image. Results show that candidates from 3D slightly helps to reconstruct room layout.

4 Conclusion

In this paper, we propose a new training and post processing method for predicting 3D room layout. Proposed post processing method predicts corners from 2D and 3D room layout. Experimental results show effectiveness of our method. Our method can be applied to other architectures. Future works include using additional information such as depth, normal, and multi image prediction.

References

1. Coughlan, J.M., Yuille, A.L.: Manhattan world: Compass direction from a single image by bayesian inference. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2, pp. 941–947 (1999)
2. Delage, E., Lee, H., Ng, A.Y.: A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 2418–2428 (2006)
3. Fernandez-Labrador, C., Facil, J.M., Perez-Yus, A., Demonceaux, C., Civera, J., Guerrero, J.J.: Corners for layout: End-to-end layout recovery from 360 images. IEEE Robotics and Automation Letters **5**(2), 1255–1262 (2020)
4. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: IEEE International Conference on Computer Vision. pp. 1849–1856 (2009)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014), https://arxiv.org/pdf/1412.6980.pdf
6. Lee, C.Y., Badrinarayanan, V., Malisiewicz, T., Rabinovich, A.: Roomnet: End-to-end room layout estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4865–4874 (2017)
7. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. pp. 8026–8037 (2019)
8. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1047–1056 (2019)
9. Xu, J., Stenger, B., Kerola, T., Tung, T.: Pano2cad: Room layout from a single panorama image. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 354–362 (2017)
10. Yang, H., Zhang, H.: Efficient 3d room shape recovery from a single panorama. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5422–5430 (2016)
11. Yang, S.T., Wang, F.E., Peng, C.H., Wonka, P., Sun, M., Chu, H.K.: Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3363–3372 (2019)
12. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: A whole-room 3d context model for panoramic scene understanding. In: European Conference on Computer Vision. pp. 668–686 (2014)
13. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. arXiv preprint arXiv:1908.00222 (2019), https://arxiv.org/pdf/1908.00222.pdf
14. 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)
15. Zou, C., Su, J.W., Peng, C.H., Colburn, A., Shan, Q., Wonka, P., Chu, H.K., Hoiem, D.: 3d manhattan room layout reconstruction from a single 360 image. arXiv preprint 1910.04099 (2019), https://arxiv.org/pdf/1910.04099.pdf