Room Layout Estimation with Nested RANSAC and Scale Alignment

Hao Zhao[1], Ming Lu[1], Yangyuxuan Kang[1,2], Anbang Yao[1], Yurong Chen[1],
Enhua Wu[2,3]
{hao.zhao,ming1.lu,yangyuxuan.kang,anbang.yao,yurong.chen}@intel.com
{kyyx@ios.ac.cn,ehwu@umac.mo}

[1]Cognitive Computing Laboratory, Intel Labs China
[2]Institute of Software, Chinese Academy of Sciences
[3]University of Macau

Abstract. In this paper we study the problem of estimating possibly non-cuboid and possibly non-Manhattan room layout from a single panorama. A major challenge is the unknown number of walls. We firstly propose a pool of walls, then randomly select elements from the pool and check the consensus between the predicted layout and pixel-wise features. This RANSAC procedure is **nested** because we randomly choose the number of walls before sampling wall indexes. We treat the number of walls as a discrete latent random variable and improve its likelihood according to the consensus. Within this **nested RANSAC** framework, a **scale alignment** algorithm that simultaneously estimates the relative height of ceilings and fuses the feature maps of ceiling and floor, plays an important role. Many intuitive cases clearly demonstrate how these two features are complementary. This method achieves top results in the ECCV 2020 holistic 3D vision challenge.

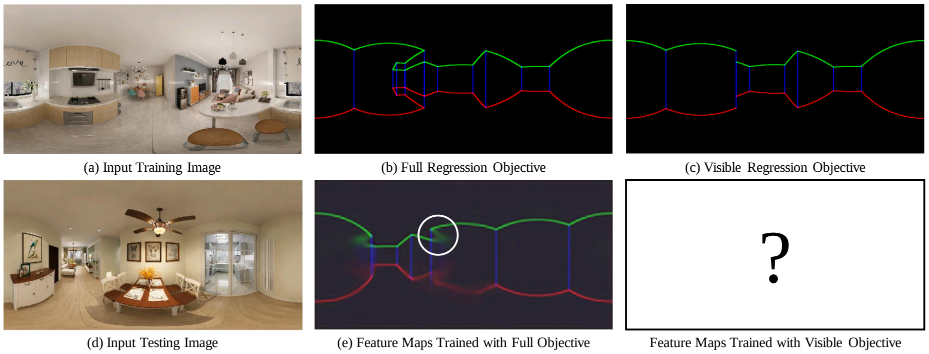
Keywords: 3D Scene Understanding, Semantic Reconstruction, Robust Fitting, Feature Fusion, Room Layout Estimation

1 Introduction

With the increasing popularity of low-cost cleaning robots, room layout estimation is no longer a toy scientific problem but an urgent industrial need. Currently, most of these devices follow a 'mapping by crashing' scheme, which is not intelligent. Room mapping with a single panorama (e.g., captured by iPhones) is an intriguing alternative, yet existing methods are challenged by non-cuboid and non-Manhattan rooms. By rendering enormous photo-realistic images with layout ground truth, a recent dataset Structured3D [6] offers an opportunity to address this challenge. Based upon deep representations learned on this dataset, we develop a robust geometric fitting framework that estimates room layouts with unknown number of walls. It enjoys the benefits of two novel techniques:

- Single-view reconstruction is intrinsically troubled by the scale ambiguity, and the common practice is to assume a physical height for the camera center. However, the relative height of ceilings remains unknown and assuming it leads to the scale mismatch of ceiling/floor features. To this end, a RANSAC-based scale alignment algorithm [1] that simultaneously estimates the relative height of ceilings and fuses ceiling/floor features, is proposed. Scale alignment clearly demonstrates how ceiling/floor features are complementary. Specifically speaking, floor features are usually polluted by occluding furniture while ceiling features are not. Fusing them leads a robust deep feature map that can facilitate robust room layout estimation.

- With the fused feature map at hand, the task asks an algorithm to fit parametric room layouts to it. However, for non-cuboid rooms, the number of walls can wildly vary within a large interval. Our idea is to resort to a nested RANSAC procedure. Among a pool of wall proposals, we first randomly pick the number of walls then randomly sample wall indexes, which accounts for the adjective 'nested'. For each wall, its location and extent are decided by an inner-loop optimization module. In each RANSAC iteration, the final layout output is compared with the feature map and we seek for the one that produces a maximum consensus. However, sampling wall indexes is itself a combinatorial step and the nested algorithm structure makes the search space even larger. Deviating from the random search nature of RANSAC, we consider the number of walls as a latent variable, and improve its likelihood according to the consensus. It turns the random fitting into an optimization algorithm, effectively accelerates the search for a high-consensus output.


(a) Input Training Image

(b) Full Regression Objective

(c) Visible Regression Objective

(d) Input Testing Image

(e) Feature Maps Trained with Full Objective

Feature Maps Trained with Visible Objective

Fig. 1. (a) Input training image. (b) Full regression objective with occluded structure lines. (c) Regression objective with only visible structure lines. (d) Input testing image. (e) Deep feature maps produced by a network trained with full objectives. In the white circle, we highlight how the network hallucinate occluded structure lines.

[1] The scale alignment is also based on RANSAC and don't confuse it with the nested RANSAC fitting framework (see next bullet point).

2 Method

Representation Learning. Following [3], we assume the camera height to be 1.5 meters. As such, the room junctions on the image sphere can be back-projected into 3D points. We uniformly sample 1000 3D points on each structure line and project them onto the image sphere, so that the hard ground truth for structure line regression is obtained. These hard ground truths are softened by gaussian kernels as [4][1] suggest. Then an ordinary fully convolutional network based upon a 105-layer dilated residual network backbone [2] is trained to regress the confidence of three types of structure lines: wall-floor separating lines, wall-ceiling separating lines, and wall-wall separating lines. A visualization of input-objective pair is given in Fig 1-a and Fig 1-b. Since the network is trained with objectives that preserve occluded structure lines [2], it can imagine the existence of them during testing, as shown by Fig 1-d and Fig 1-e. We are investigating the alternative option of training with only visible structure lines as Fig 1-c demonstrates, yet by the time of writing, the network is still in training.

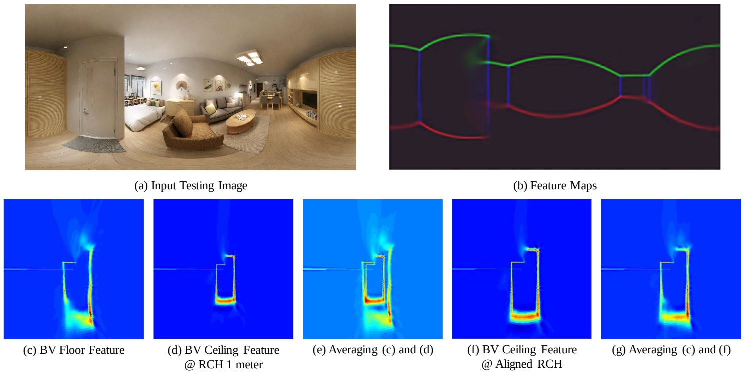

(a) Input Testing Image
(b) Feature Maps
(c) BV Floor Feature
(d) BV Ceiling Feature @ RCH 1 meter
(e) Averaging (c) and (d)
(f) BV Ceiling Feature @ Aligned RCH
(g) Averaging (c) and (f)

Fig. 2. (a) Input testing image. (b) Network output. (c) Bird-view floor feature map by assuming camera height at 1.5 meters. (d) Bird-view ceiling feature map by assuming relative ceiling height (RCH) at 1 meter. (e) Averaging (c) and (d). (f) Bird-view ceiling feature map by using aligned RCH. (g) Averaging (c) and (f).

Scale Alignment. We will refer to the feature map channel for wall-wall separating lines by wall features, and other channels similarly. And we want to use both ceiling and floor features for fitting as they are complementary. To demonstrate this fact, we visualize the bird-view (BV) [3] features in Fig 2-c and Fig 2-d. It is obvious that Fig 2-d reflects the room structure more faithfully yet its scale mismatches the scale of Fig 2-c. Directly averaging them leads to Fig 2-e and this is caused by the unknown relative ceiling height (RCH) [4]. Our

[2] 'Occluded' by other structure lines instead of occluded by furniture.

[3] We (ab)use the term 'bird-view' for a top-down viewpoint.

[4] RCH is the distance between the camera center and the ceiling.

scale alignment algorithm plays the role of estimating RCH so that an aligned BV ceiling feature map (Fig 2-f) is obtained and fused with the BV floor map to get Fig 2-g. It is obvious that Fig 2-g is better than Fig 2-c, in many senses. Specifically speaking, scale alignment is achieved with RANSAC. There are 1024 columns in a panorama and any of them can be used estimate an RCH. We randomly select $n = 1024 \times ratio$ samples from columns and use the variance of n estimates as the consensus. If a RANSAC run produces the lowest variance, the mean value of n estimates in that run is used as the final RCH prediction. This scale alignment algorithm works well both qualitatively and quantitatively.

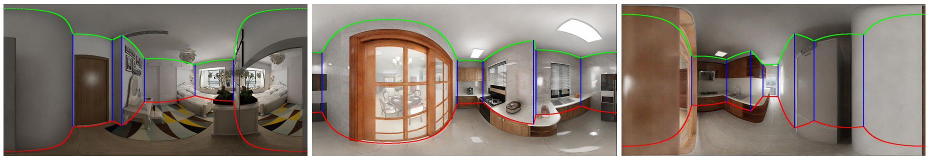


Fig. 3. Our method successful estimates challenging room layouts.

Nested RANSAC Fitting. To clarify, since the number of wall-wall separating lines and the number of walls are equal, we refer to one of these lines as a wall, with a little abuse of terms. We firstly construct a pool of walls by fitting a Gaussian mixture model (GMM) to the wall feature map. Every component of the GMM corresponds to a wall and its mixing coefficient represents its probability to be sampled. Let us assume there are m components and m is determined by finding peaks in the wall feature map before GMM fitting. In every RANSAC iteration, we first sample the wall number o from $[4, m]$ by assuming an uniform distribution then sample o indexes from the pool. For each wall, an inner loop decides the location and extent of this wall. On the fused feature map (e.g., Fig 2-g), each wall corresponds to a fan-shaped region and we seek a point on each radius segment so that the average score on a segment connecting these two points are maximized. In this inner loop, the Manhattan assumption can be enforced or not. Empirically, enforcing this constraint shrinks the solution space and improves final quantitative measures. As such, in every RANSAC run a final room layout prediction is obtained in the aforementioned wall-wise manner. For each layout prediction, we render it into wire-frames and check its consensus with three feature maps (e.g., Fig 1-e), which is similar to the objective in [5]. Finally, the RANSAC run that gives the highest consensus is considered the output. Denoting the feature maps as F, we update the likelihood $P(o|F)$ so that it is no longer an uniform distribution as mentioned above. This is done by normalizing consensus values obtained for different o. This practice turns the random search into an optimization procedure and effectively accelerates the search for a high-consensus layout. Some qualitative results of estimated challenging room layouts are depicted in Fig 3.

References

1. Yan, C., Shao, B., Zhao, H., Ning, R., Zhang, Y., Xu, F.: 3d room layout estimation from a single rgb image. IEEE Transactions on Multimedia (2020)
2. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 472–480 (2017)
3. 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)
4. Zhao, H., Lu, M., Yao, A., Chen, Y., Zhang, L.: Learning to draw sight lines. International Journal of Computer Vision pp. 1–25 (2019)
5. 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. pp. 10–18 (2017)
6. 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)