Joint 3D Layout and Depth Prediction from a Single Indoor Panorama Image

Wei Zeng[1], Sezer Karaoglu[1,2], and Theo Gevers[1,2]

[1] Computer Vision Laboratory, University of Amsterdam, The Netherlands
{w.zeng,th.gevers}@uva.nl
[2] 3DUniversum, Science Park 400, The Netherlands
{s.karaoglu,theo.gevers}@3duniversum.com

Abstract. In this paper, we propose a method which jointly learns the layout prediction and depth estimation from a single indoor panorama image. Previous methods have considered layout prediction and depth estimation from a single panorama image separately. However, these two tasks are tightly intertwined. Leveraging the layout depth map as an intermediate representation, our proposed method outperforms existing methods for both panorama layout prediction and depth estimation. Experiments on the challenging real-world dataset of Stanford 2D-3D demonstrate that our approach obtains superior performance for both the layout prediction tasks (3D IoU: 85.81% v.s. 79.79%) and the depth estimation (Abs Rel: 0.068 v.s. 0.079).

Keywords: Indoor Panorama Image · Layout Prediction · Depth Estimation · Layout Depth Map

1 Introduction

Extracting 3D information from 2D indoor images is an important step towards the enabling of 3D understanding of indoor scenes and is beneficial for many applications such as robotics and virtual/augmented reality. Using the 3D information of indoor scenes, a computer vision system is able to understand the scene geometry, including both the apparent and hidden relationships between scene elements.

Although scene layout and depth can both be used for 3D scene understanding, previous methods focus on solving these two problems separately. For 3D layout prediction, methods mostly use 2D geometrical cues such as edges [20, 25, 35], corners [16, 25, 35], 2D floor-plans [19, 30] or they make assumptions about the 3D scene geometry such that rooms are modelled by cuboids or by a Manhattan World. For depth estimation, different features are used such as normals [17], planar surfaces [21] and semantic cues [22]. Hence, existing methods impose geometric assumptions but ignore to exploit the complementary characteristics of layout and depth information. In this paper, a different approach is taken. We propose a method that, from a single panorama, jointly exploits the 3D layout and depth cues via an intermediate layout depth map, as shown in Fig. 1. The

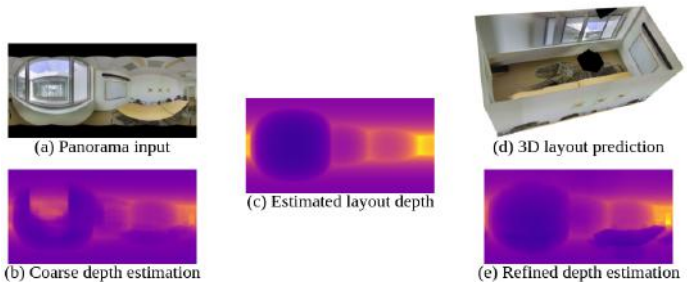

Fig. 1: Given (a) an indoor panorama as input, our proposed method utilizes the (b) coarse depth estimation to compute the (c) layout depth map. Leveraging the estimated layout depth map, our method improves the (d) 3D layout prediction and (e) refines the depth estimation (e.g. the ambiguous window depth is inferred correctly compared to the coarse depth estimation)

intermediate layout depth map represents the distances from the camera to the room layout components (e.g. ceiling, floor and walls) and excludes all objects in the room (e.g. furniture), as illustrated in Fig. 2. Estimating the layout depth as an intermediate representation of the network encompasses the geometric information needed for both tasks. The use of depth information is beneficial to produce room layouts by reducing the complexity of object clutter and occlusion. Likewise, the use of room layout information diminishes the ambiguity of depth estimation and interposes planar information for the room layout parts (e.g. ceiling, floor and walls).

The proposed method estimates the 3D layout and detailed depth information from a single panorama image. To combine the depth and layout information, the proposed method predicts the layout depth map to relate these two tightly intertwined tasks. Previous methods on layout prediction provides proper reconstruction by predicting the layout edges and corners on the input panorama and by post-processing them to match the (Manhattan) 3D layout [16, 25, 35]. However, object clutter in the room poses a challenge to extract occluded edges and corners. In addition, estimating the 3D layout from 2D edge and corner maps is an ill-posed problem. Therefore, extra constraints are essential to perform 2D to 3D conversion in the optimization. In contrast, our method estimates the layout depth map by using more structural information to become less influenced by occlusions. Furthermore, the predicted layout depth map serves as a coarse 3D layout as it can be converted to the 3D point cloud of the scene layout. Thus the proposed method does not require extra constraints for the 2D to 3D conversion. This makes the proposed method more generic for parameterizing a 3D layout. After computing the estimated layout depth maps, the proposed method further enables the refinement of a detailed depth map. Monocular depth estimation methods usually have problems with planar room parts (ceiling, floor and walls) being rugged after the 3D reconstruction process. The layout depth map preserves the planar nature of the room layout components yielding robust-

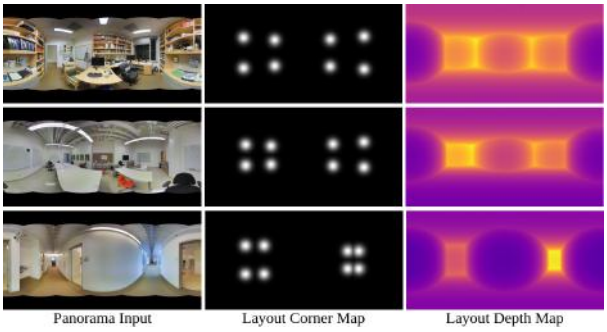

Fig. 2: Illustration of the layout depth maps. From left to right: the panorama input image, the original layout corner map and the layout depth map

ness to these errors. Empirical results on the challenging Stanford 2D-3D indoor dataset show that jointly estimating 3D layout and depth outperforms previous methods for both tasks. The proposed method achieves state-of-the-art performance for both layout prediction and depth estimation from a single panorama image on the Stanford 2D-3D dataset. Our method also obtains state-of-the-art performance for 3D layout prediction on the PanoContext dataset.

In summary, our contributions are as follows:

- We propose a novel neural network pipeline which jointly learns layout prediction and depth estimation from a single indoor panorama image. We show that layout and depth estimation tasks are highly correlated and joint learning improves the performance for both tasks.
- We show that leveraging the layout depth map as an intermediate representation improves the layout prediction performance and refines the depth estimation.
- The proposed method outperforms the state-of-the-art methods for both layout prediction and depth estimation on the challenging real-world dataset Stanford 2D-3D and PanoContext dataset for layout prediction.

2 Related Work

Panorama Images: Operating directly on panorama input images is the primary difference between our method and most of the other layout prediction or depth estimation methods. Instead of perspective images, 360° panorama images are used as input by our proposed method because the field of view (FOV) of panoramas are larger and carry more scene information. However, the equirectangular projections may suffer from strong horizontal distortions. Su et al. [24] propose to learn a spherical convolutional network that translates a planar CNN to process 360° panorama images directly in its equirectangular projection. Tateno et al. [26] proposes a distortion-aware deformable convolution filter. Another approach is to use spherical convolutions as proposed by

Cohen et al. [3]. Other recent papers [4, 8, 13] also focus on spherical CNNs and icosahedron representations for panorama processing. In this paper, standard convolutions with rectangular filter banks are applied on the input layers to account for the different distortion levels.

Layout Prediction: There are numerous papers that address the problem of predicting the 3D room layout from a single image taken from an indoor scene. Traditional methods treat this task as an optimization problem. Delage et al. [5] propose a dynamic Bayesian network model to recover the 3D model of the indoor scene. Hedau [10] models the room with a parametric 3D box by iteratively localizing clutter and refitting the box. Recently, neural network-based methods took stride in tackling this problem. Methods that train deep network to classify pixels into layout surfaces (e.g., walls, floor, ceiling) [12], boundaries [20], corners [16], or a combination [23]. Zou et al. [35] predict the layout boundary and corner map directly from the input panorama. Yang et al. [30] leverage both the equirectangular panorama-view and the perspective ceiling-view to learn different cues about the room layout. Sun et al. [25] encode the room layout as three 1D vectors and propose to recover the 3D room layouts from 1D predictions. Other work aims to leverage depth information for room reconstruction [18, 32, 36], but they all deal with perspective images and use the ground truth depth as input. In contrast, in our paper, we use the predicted depth and semantic content of the scene to predict the layout depth map as our intermediate representation to recover the 3D layout of the input panorama.

Depth Estimation: Single-view depth estimation refers to the problem of estimating depth from a single 2D image. Eigen et al. [9] show that it is possible to produce pixel depth estimations using a two scale deep network which is trained on images with their corresponding depth values. Several methods extend this approach by introducing new components such as CRFs to increase the accuracy [17], changing the loss from regression to classification [2], using other more robust loss functions [15], and by incorporating scene priors [29]. Zioulis et al. [34] propose a learning framework to estimate the depth of a scene from a single 360° panorama image. Eder et al. [7] present a method to train a plane-aware convolutional network for dense depth and surface normal estimation from panoramas. There are some other methods [6, 27] to regress the layered depth image (LDI) to capture the occluded texture and depth. In our work, we demonstrate that the layout prediction and depth estimation are tightly coupled and can benefit from each other. Leveraging the estimated layout depth map, our method refines the depth estimation.

3 Method

The goal of our approach is the joint learning of layout prediction and depth estimation from a single indoor panorama image. The proposed method leverages the layout depth map as an intermediate representation to relate the layout and depth estimation. Fig. 3 shows an overview of our proposed pipeline.

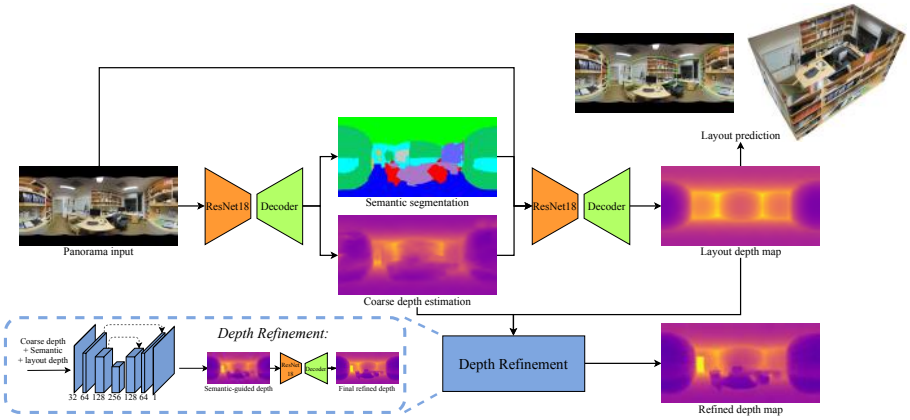

Fig. 3: Overview of the proposed pipeline. Our method first leverages the coarse depth and semantic prediction to enforce the layout depth prediction, and then uses the estimated layout depth map to recover the 3D layout and refine the depth estimation

Inferring high-quality 3D room layout from an indoor panorama image relies on the understanding of both the 3D geometry and the semantics of the indoor scene. Therefore, the proposed method uses the predicted coarse depth map and semantic segmentation of the input panorama to predict the layout depth map. The proposed method enables the refinement of depth estimation by integrating the coarse depth and layout depth with semantic information as a guidance.

3.1 Input and Pre-processing

Following [35], the first step of our method is to align the input panorama image to match the horizontal floor plane. The floor plane direction under equirectangular projection is estimated by first selecting the long line segments using the Line Segment Detector (LSD) [28] in overlapping perspective views and then vote for three mutually orthogonal vanishing directions [33]. This alignment ensures that wall-wall boundaries are vertical lines. The input of our network is the concatenation of the panorama image and the corresponding Manhattan line feature map provided by the alignment.

3.2 Coarse Depth and Semantics

Our approach receives the concatenation of a single *RGB* panorama and the Manhattan line feature map as input. The output of this module is the coarse depth estimation and semantic segmentation of the 2D panorama image.

An encoder-decoder architecture is used for the joint learning of the coarse depth information and semantic segmentation. The input panorama images suffer from horizontal distortions. To reduce the distortion effect, the encoder uses

a modified input block in front of the ResNet-18 architecture. As shown by [34], the input block uses rectangle filters and varies the resolution to account for different distortion levels. The encoder is shared for both the depth estimation and semantic segmentation. The decoders restore the original input resolution by means of up-sampling operators followed by 3×3 convolutions. Skip connections are also added to link to the corresponding resolution in the encoder. The two decoders do not share weights and are trained to minimize the coarse depth estimation loss and semantic segmentation loss, respectively.

Loss Function: For coarse depth estimation, to account for both pixel-wise accuracy and spatially coherent results, this module incorporates the depth gradient and normals with the logarithm of the standard L1 loss, as done by [11]. So the loss function consists of three parts:

$$L_{coarse\_depth} = l_{depth} + \lambda l_{gradient} + \mu l_{normal} \tag{1}$$

where $\lambda, \mu \in R$ are hyper-parameters to balance the contribution of each component loss. The depth loss l_{depth}, the gradient loss $l_{gradient}$ and the surface normal loss l_{normal} are defined by:

$$l_{depth} = \frac{1}{n} \sum_{i=1}^{n} \ln(e_i + 1) \tag{2}$$

where $e_i = \|d_i - g_i\|_1$, d_i and g_i denote the predicted and ground truth depth maps respectively. n is the total number of pixels.

$$l_{gradient} = \frac{1}{n} \sum_{i=1}^{n} (\ln(|\nabla_x(e_i)| + 1) + \ln(|\nabla_y(e_i)| + 1)) \tag{3}$$

where $\nabla_x(e_i)$ is the spatial derivative of e_i computed at the i^{th} pixel with respect to x, and so on.

$$l_{normal} = \frac{1}{n} \sum_{j=1}^{n} \Big(1 - \frac{\langle n_j^d, n_j^g \rangle}{\sqrt{\langle n_j^d, n_j^d \rangle} \sqrt{\langle n_j^g, n_j^g \rangle}}\Big) \tag{4}$$

where $n_i^d \equiv [-\nabla_x(d_i), -\nabla(d_i), 1]^\top$ and $n_i^g \equiv [-\nabla_x(g_i), -\nabla(g_i), 1]^\top$ denote the surface normal of the estimated depth map and the ground truth, respectively.

For semantic segmentation, the loss function is given by the per-pixel softmax cross-entropy between the predicted and ground-truth pixel-wise semantic labels:

$$L_{semantic} = -\sum_{i=1}^{n} p_i \log(\hat{p}_i) \tag{5}$$

where p and $\hat{p}$ are the ground truth and predicted semantic labels, respectively.

3.3 Layout Prediction

To obtain the global geometric structure of the scene, the proposed approach predicts the 3D layout of the scene. Instead of predicting 2D representations, our method directly predicts the layout depth maps of the input panoramas.

The input of this proposed module is a 8-channel feature map: the concatenation of *RGB* panorama, the corresponding Manhattan line feature map, and the predicted depth and semantics obtained by the previous modules of the pipeline. A ResNet-18 is used to build our encoder for the layout depth prediction network. The decoder architecture is similar to the previous ones for depth estimation and semantic segmentation, with nearest neighbor up-sampling operations followed by 3×3 convolutions. The skip connections are also added to prevent shifting of the prediction results during the up-sampling step. The output is the estimated layout depth map with the same resolution as the input panorama.

Loss Function: In addition to the pixel-wise depth supervision as described in Section 3.2, the virtual normal (VN) [31] is used as another geometric constraint to regulate the estimated layout depth map. The point cloud of the scene layout can be reconstructed from the estimated layout depth map based on the panoramic camera model. The virtual normal is the normal vector of a virtual plane formed by three randomly sampled non-colinear points in 3D space, which takes long-range relations into account from a global perspective. By minimizing the direction divergence between the ground-truth and predicted virtual normals, serving as a high-order 3D geometric constraint, the proposed method provides more accurate depth estimation and imposes the planar nature to the prediction of the layout depth map.

N group points are randomly sampled from the point cloud. In each group there are three points: $\Omega = \{P_i = (P_a, P_b, P_c)_i \mid i = 0, ..., N\}$. The three points in a group are restricted to be non-colinear as defined by condition C:

$$C = \{\alpha \geq \angle(\overrightarrow{P_aP_b}, \overrightarrow{P_aP_c}) \leq \beta, \alpha \geq \angle(\overrightarrow{P_bP_c}, \overrightarrow{P_bP_a}) \leq \beta \mid P_i \in \Omega\} \tag{6}$$

where $\alpha = 150°, \beta = 30°$ in our experiments.

Three points in each group establishes a virtual plane. The normal vector of the plane is computed by:

$$N = \{\boldsymbol{n_i} = \frac{\overrightarrow{P_aP_b} \times \overrightarrow{P_aP_c}}{\|\overrightarrow{P_aP_b} \times \overrightarrow{P_aP_c}\|} \mid P_i \in \Omega\} \tag{7}$$

where $\boldsymbol{n_i}$ is the normal vector of virtual plane P_i.

The virtual normal loss is computed by:

$$l_{vn} = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{n_i^{pred}} - \boldsymbol{n_i^{gt}}\|_1 \tag{8}$$

The overall loss for layout depth map estimation is defined by:

$$L_{layout\_depth} = l_{depth} + \lambda l_{gradient} + \mu l_{normal} + l_{vn} \tag{9}$$

The layout depth loss is based on both the local surface normal and the global virtual normal constraint. This ensures that the estimated layout depth map preserves the geometric structure of the scene layout accurately.

3D Layout Optimization: To constrain the layout shape so that the floor and ceiling are planar and the walls are perpendicular to each other (Manhattan world assumption), the proposed method recovers the parameterized 3D layout through optimization in 3D space. Previous methods [16, 35, 25] heavily rely on 2D image features (e.g. edge and corner maps). However, estimating the 3D layout from 2D edge and corner maps is an ill-posed problem and thus requires extra constraints. In contrast, our proposed method directly optimizes on the 3D layout point cloud and does not require extra constraints for the 2D to 3D layout conversion.

Using the point cloud of the scene layout converted from the predicted layout depth map, the floor/ceiling plan map is obtained by projecting the point cloud to the XZ plane. Similar to [30], a regression analysis is applied on the edges of the floor plan map and clustering them into sets of horizontal and vertical lines in 3D space. Then, the floor plan is recovered by using the straight, axis-aligned, wall-floor boundaries. The room height is efficiently computed by using the ceiling-floor distances along the Y axis.

3.4 Depth Refinement

After the coarse depth map and the layout depth map are obtained from the previous modules, a depth refinement step is taken.

A straight-forward way is to concatenate all the data representations as input and use an encoder-decoder network to predict the final depth estimation. This approach is denoted by *direct refinement*. The semantic approach is to use the semantic information as a guidance to dynamically fuse the two depth maps. This approach is denoted by *semantic-guided refinement*. The semantic-guided refinement step produces an attention map incorporating the coarse depth map and the layout depth map. For a structural background representing the scene layout components (ceiling, floor and wall), the network focuses more on the layout depth map. While for objects in the room (furniture), the network switches the attention to the coarse depth estimation. Therefore, in this paper, we combine these two concepts as shown in Fig. 3. First, an encoder-decoder network, taking the concatenation of the coarse depth, layout depth and semantic segmentation prediction as inputs, combines the previous depth maps with the semantic-guided attention map. This semantic-guided depth fusion maximizes the exploitation of the coarse depth and layout depth. Then, the depth refinement module takes the fused depth as input to predict the final refined depth. The encoder-decoder architecture of the depth refinement module is similar to the previous coarse depth estimation network.

Loss Function: The loss function for the depth refinement is the same as the layout depth estimation loss described in Section 3.3.

3.5 Training Details

Following the experimental setting of [35], the proposed method uses horizontal rotations, left-right flippings and luminance changes to augment the training samples. Our network uses the ADAM [14] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to update the network parameters. To train the network, we first train the joint learning of coarse depth estimation and semantic segmentation, and then fix the weights of the depth and semantic network, and train the layout depth map prediction. Then, we set all the trained weights fixed to train the depth refinement module. Finally, we jointly train the whole network end-to-end.

4 Experiments

In this section, the performance of our proposed method is evaluated for both the layout prediction and depth estimation tasks.

Dataset: The dataset used for training is the Stanford 2D-3D dataset [1]. The Stanford 2D-3D dataset contains 1413 *RGB* panoramic images collected from 6 large-scale indoor environments, including offices, classrooms, and other open spaces like corridors, where 571 panoramas have layout annotations. Our experiments follow the official train-val-test split for evaluation. The PanoContext dataset is used to verify the generalizability of our approach for the task of layout prediction. The PanoContext [33] dataset contains 514 *RGB* panoramic images of two indoor environments, i.e., bedrooms and living rooms.

Evaluation Metrics: The following standard metrics are used to evaluate our approach:

3D IoU: $3D\ IoU = \frac{V_{pred} \cap V_{gt}}{V_{pred} \cup V_{gt}}$, where V_{pred} and V_{gt} stand for the volumetric occupancy of the predicted and ground truth 3D layout.

Corner error (CE): $CE = \frac{1}{\sqrt{H^2+W^2}} \sum_{i \in corners} \|c_i^{pred} - c_i^{gt}\|_2^2$, where H and W are the image height and width, c^{pred} and c^{gt} denote the predicted and ground truth corner positions.

Pixel error (PE): $PE = \frac{1}{|N|} \sum_{i=1}^{N} \mathbb{1}(s_i^{pred} \neq s_i^{gt})$, where s^{pred} and s^{gt} denotes the predicted and ground truth pixel-wise semantic (ceiling, floor and wall). $\mathbb{1}(.)$ is an indicator function, setting to 1 when the pixel semantic prediction is incorrect.

Threshold: % of d_i that $max(\frac{d_i}{g_i}, \frac{g_i}{d_i}) = \delta < thr$

Absolute Relative Difference: $Abs\ Rel = \frac{1}{|N|} \sum_{i=1}^{N} \|d_i - g_i\| / g_i$

Squared Relative Difference: $Sq\ Rel = \frac{1}{|N|} \sum_{i=1}^{N} \|d_i - g_i\|^2 / g_i$

RMSE (linear): $RMS = \sqrt{\frac{1}{|N|} \sum_{i=1}^{N} \|d_i - g_i\|^2}$

RMSE (log): $RMS(log) = \sqrt{\frac{1}{|N|} \sum_{i=1}^{N} \|\log d_i - \log g_i\|^2}$

where we use 3D IoU, corner error and pixel error to evaluate the layout prediction and the rest for depth estimation.

Table 1: Quantitative results of layout estimation on the Stanford 2D-3D dataset. Our method outperforms all existing methods

Method	3D IoU(%)	Corner error(%)	Pixel error(%)
LayoutNet [35]	76.33	1.04	2.70
DuLa-Net [30]	79.36	-	-
HorizonNet [25]	79.79	0.71	2.39
Ours	**85.81**	**0.67**	**2.20**

Table 2: Quantitative results on the (a) Stanford 2D-3D and (b) PanoContext for models trained with mixed PanoContext and Stanford 2D-3D training data. Our method outperforms other methods on both datasets

Method	3D IoU(%)	CE(%)	PE(%)
LayoutNet [35]	77.51	0.92	2.42
HorizonNet [25]	83.51	**0.62**	**1.97**
Ours	**86.21**	0.71	2.08

(a) Results for Stanford 2D-3D

Method	3D IoU(%)	CE(%)	PE(%)
LayoutNet [35]	75.12	1.02	3.18
HorizonNet [25]	84.23	0.69	1.90
Ours	**84.40**	**0.61**	**1.74**

(b) Results for PanoContext

4.1 Layout Prediction

A quantitative comparison of different methods on the Stanford 2D-3D dataset is summarized in Table 1. LayoutNet [35] predicts the layout boundary and corner maps directly from the input panorama. DuLa-Net [30] leverages both the equirectangular panorama-view and the perspective ceiling-view to learn different cues for the room layout. HorizonNet [25] encodes the room layout as three 1D vectors and proposes to recover the 3D room layout from 1D predictions by a RNN. The proposed method shows state-of-the-art performance and outperforms other existing methods. By leveraging the layout depth map as an intermediate representation, the proposed network abstracts the geometric structure of the scene from both a local and global perspective. This results in more geometric cues for the scene layout prediction and is less affected by occlusions.

LayoutNet [35] and HorizonNet [25] also combine the Stanford 2D-3D [1] and PanoContext [33] training data to train their methods. Since the PanoContext dataset does not contain any depth or semantic ground truth, our model is first initialized with the Stanford 2D-3D dataset, and then the model is trained on the same mixed dataset with the weight-fixed coarse depth and semantic prediction modules. Table 2 shows the quantitative results trained on this mixed training data. Although the PanoContext dataset has different indoor configurations and no depth or semantic ground truth, our method still obtains competitive performance.

Fig. 4: Qualitative comparison on layout prediction. Results are shown of testing the baseline LayoutNet [35] (blue), our proposed method (green) and the ground truth (orange) on the Stanford 2D-3D dataset and PanoContext dataset

The qualitative results for the layout prediction are shown in Fig. 4. The first two rows demonstrate the results of the LayoutNet and our proposed method on the Stanford 2D-3D dataset. The last two rows are the results obtained for the PanoContext dataset. The proposed method outperforms the other methods on both datasets and shows robustness to occlusion. As presented by the second example for Stanford 2D-3D, since the proposed method explicitly incorporates the depth information, the corners are located more precisely (avoiding locations in the middle of the wall which has continuous depth). The semantic content ensures the detection of the occluded corners, as shown in the third example of Stanford 2D-3D (corners occluded by the door). The last example of the Stanford 2D-3D shows a failure case for both methods. For non-negligible occlusions in the scene, both methods fail to predict the corner positions accurately. Similar improvements are shown for the results obtained for the PanoContext dataset.

Ablation Study: The goal is to evaluate the performance of our layout prediction and layout depth estimation with different configurations: 1) *wo/ depth&semantic*: predicting the layout depth directly from the input; 2) *w/ pred. depth*: only with the predicted depth; 3) *w/ pred. semantic*: only with the predicted semantic; 4) *wo/ VN*: without the VN loss; 5) *edg&cor maps*: predicting the edge and corner maps from the concatenation of input panorama, predicted depth and semantic; 6) *layout depth -> edg&cor maps*: predicting the edge and corner maps from the layout depth map. As shown in Table 3, training with either predicted depth or semantic information increases the accuracy. The VN loss further regulates the estimated layout depth to preserve surface straightness, thus improving the recovered layout. In comparison with the edge and corner maps, the layout depth map contains both local and global information to recover the 3D layout of the scene.

Table 3: Ablation study of layout prediction and layout depth map estimation on the Stanford 2D-3D dataset. We evaluate the influence of different modules and show that our final proposed approach performs the best

				lower is better				higher is better		
	3D IoU(%)	CE(%)	PE(%)	Abs Rel	Sq Rel	RMS	RMS(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
wo/ depth & semantic	77.28	1.21	3.31	0.089	0.044	0.327	0.056	0.914	0.987	0.996
w/ pred. depth	82.65	0.83	2.92	0.069	0.029	0.257	0.045	0.952	0.993	**0.998**
w/ pred. semantic	78.57	1.14	3.18	0.079	0.034	0.311	0.053	0.927	0.990	0.997
wo/ VN	84.22	0.75	2.42	0.065	0.028	0.238	0.043	0.955	0.993	**0.998**
edg & cor maps	82.03	1.05	2.61	-	-	-	-	-	-	-
layout depth –> edg & cor maps	83.67	0.92	2.52	0.067	0.029	0.238	0.044	0.955	0.992	**0.998**
Proposed Final	**85.81**	**0.67**	**2.20**	**0.064**	**0.026**	**0.237**	**0.042**	**0.957**	**0.994**	**0.998**

Fig. 5: Qualitative results of non-cuboid layout prediction. It can be derived that our proposed method also works well for non-cuboid layouts

Non-cuboid Layout: To verify the generalization ability of our proposed method to non-cuboid layout, our model is fine-tuned on the non-cuboid rooms labeled by [25]. As shown in Fig. 5, our proposed method is able to handle non-cuboid layout rooms. Please see more results in the supplemental materials.

4.2 Depth Estimation

Table 4 presents the quantitative results of different methods for depth estimation on the Stanford 2D-3D dataset. FCRN [15] designs a supervised fully convolutional residual network with up-projection blocks. RectNet [34] proposes a specific pipeline for depth estimation using panoramas as input. DistConv [26] trains on perspective images and then regress depth for panorama images by distortion-aware deformable convolution filters. Plane-aware [7] designs the plane-aware loss which leverages principal curvature as an indicator of planar boundaries. The results demonstrate that our proposed method obtains state-of-the-art depth estimation results from a single panorama image. The qualitative comparison is shown in Fig. 6. In the first image, the RectNet [34] is confused by the transparent window, which is a common failure case in depth estimation. The Plane-aware network [7] and our proposed network overcome this issue. Our result for the window region is smoother due to the constraints from the layout depth. In the second image, the distant regions are too ambiguous to predict the corresponding depth. Our proposed method predicts a proper depth map because

Table 4: Quantitative results and ablation study of depth estimation on the Stanford 2D-3D dataset. Our method outperforms all existing methods

	lower is better				higher is better		
	Abs Rel	Sq Rel	RMS	RMS(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
FCRN [15]	0.091	0.057	0.364	0.134	0.913	0.982	0.995
RectNet[34]	0.082	0.046	0.399	0.123	0.928	0.988	0.997
DistConv [26]	0.176	-	0.369	0.083	-	-	-
Plane-aware [7]	0.079	0.029	0.290	0.120	0.934	0.990	**0.998**
Proposed Coarse-depth	0.105	0.045	0.352	0.094	0.934	0.989	0.997
Proposed Direct-refinement	0.089	0.033	0.269	0.095	0.944	0.989	**0.998**
Proposed Semantic-guided	0.086	0.033	0.273	0.096	0.944	0.989	**0.998**
Proposed Final	**0.068**	**0.026**	**0.264**	**0.080**	**0.954**	**0.992**	**0.998**

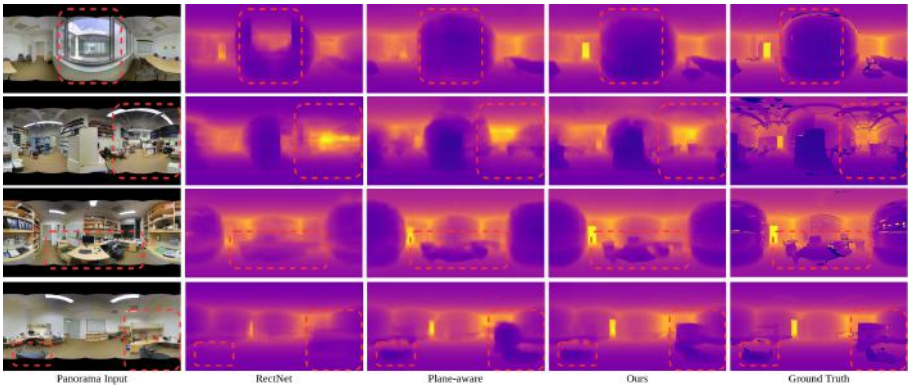

Fig. 6: Qualitative comparison on depth estimation. Results are shown for testing the baseline RectNet [34], Plane-aware network [7] and our proposed method on the Stanford 2D-3D dataset

of the explicit inter-positioning of the layout depth. Because of the proposed semantic-guided refinement, the proposed method also preserves better object details compared to the other two methods, as shown in the third and fourth image. Fig. 7 illustrates the derived surface normals from the estimated depth map. Constrained by the layout depth map, the surface normal results demonstrate that our proposed method preserves the planar property for depth estimation.

Ablation Study: An ablation study is conducted to evaluate the performance of the proposed method for different configurations, as shown in Table 4: 1) *Proposed Coarse-depth*: the depth estimation from the first decoder; 2) *Proposed Direct-refinement*: the depth refinement using all the data representation as input, as stated in Section 3.4; 3) *Proposed Semantic-guided*: the depth fusion using semantic-guided attention map, as state in Section 3.4. It is shown that the direct-refinement performs better than the coarse-depth. This indicates that the joint learning with layout prediction already improves the depth estimation.

Table 5: Quantitative comparison of the proposed method for joint training. It is shown that joint training improves the performance for all the proposed modules

				lower is better				higher is better		
	3D IoU(%)	CE(%)	PE(%)	Abs Rel	Sq Rel	RMS	RMS(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Coarse depth	-	-	-	0.112	0.049	0.379	0.116	0.930	0.988	**0.997**
Coarse depth (joint)	-	-	-	**0.105**	**0.045**	**0.352**	**0.094**	**0.934**	**0.989**	**0.997**
Depth refinement	-	-	-	0.084	0.032	0.273	0.088	0.950	0.989	**0.998**
Depth refinement (joint)	-	-	-	**0.068**	**0.026**	**0.264**	**0.080**	**0.954**	**0.992**	**0.998**
Layout depth	84.69	0.75	2.43	0.069	0.029	0.257	0.046	0.951	0.993	**0.998**
Layout depth (joint)	**85.81**	**0.67**	**2.20**	**0.064**	**0.026**	**0.237**	**0.042**	**0.957**	**0.994**	**0.998**

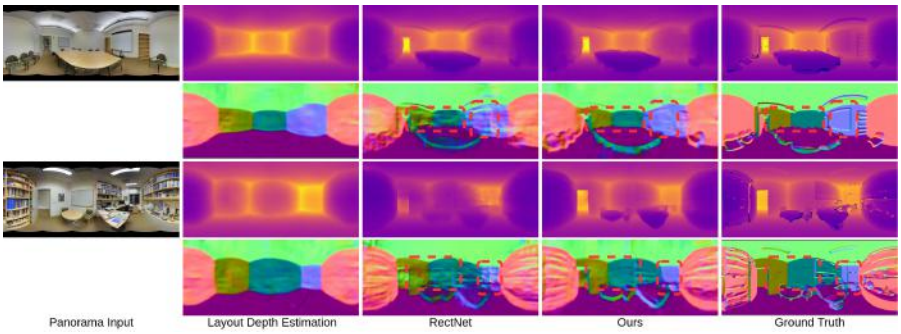

Fig. 7: Comparison of the derived surface normal from the depth estimation. Our proposed method produces smoother surfaces for planar regions

Semantic-guided refinement improves the performance which supports our argument to dynamically fuse the layout depth map and the coarse depth estimation based on background and foreground regions. Our proposed final method obtains the best overall performance for all variations.

Table 5 shows the quantitative comparison for each module of the proposed pipeline before and after joint training. It demonstrates that all the modules benefit from joint training.

5 Conclusion

We proposed a method to jointly learn the layout and depth from a single indoor panorama image. By leveraging the layout depth map as an intermediate representation, the optimization of 3D layout does not require extra constraints and the refined depth estimation preserves the planarity for the layout components. Experiment results on challenging indoor datasets show that, with the proposed method for joint learning, the performance of both the layout prediction and depth estimation from single panorama images is significantly improved and that our method outperforms the state-of-the-art.

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
2. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology (2017)
3. Cohen, T., Geiger, M., Köhler, J., Welling, M.: Convolutional networks for spherical signals. arXiv preprint arXiv:1709.04893 (2017)
4. Cohen, T.S., Weiler, M., Kicanaoglu, B., Welling, M.: Gauge equivariant convolutional networks and the icosahedral cnn. arXiv preprint arXiv:1902.04615 (2019)
5. Delage, E., Lee, H., Ng, A.Y.: A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 2418–2428. IEEE (2006)
6. Dhamo, H., Tateno, K., Laina, I., Navab, N., Tombari, F.: Peeking behind objects: Layered depth prediction from a single image. Pattern Recognition Letters **125**, 333–340 (2019)
7. Eder, M., Moulon, P., Guan, L.: Pano popups: Indoor 3d reconstruction with a plane-aware network. In: 2019 International Conference on 3D Vision (3DV). pp. 76–84. IEEE (2019)
8. Eder, M., Price, T., Vu, T., Bapat, A., Frahm, J.M.: Mapped convolutions. arXiv preprint arXiv:1906.11096 (2019)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
10. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: 2009 IEEE 12th international conference on computer vision. pp. 1849–1856. IEEE (2009)
11. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1043–1051. IEEE (2019)
12. Izadinia, H., Shan, Q., Seitz, S.M.: Im2cad. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5134–5143 (2017)
13. Jiang, C., Huang, J., Kashinath, K., Marcus, P., Niessner, M., et al.: Spherical cnns on unstructured grids. arXiv preprint arXiv:1901.02039 (2019)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 239–248. IEEE (2016)
16. 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)
17. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1119–1127 (2015)

18. Liu, C., Kohli, P., Furukawa, Y.: Layered scene decomposition via the occlusion-crf. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 165–173 (2016)
19. Liu, C., Schwing, A.G., Kundu, K., Urtasun, R., Fidler, S.: Rent3d: Floor-plan priors for monocular layout estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3413–3421 (2015)
20. Mallya, A., Lazebnik, S.: Learning informative edge maps for indoor scene layout prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 936–944 (2015)
21. Micusik, B., Kosecka, J.: Piecewise planar city 3d modeling from street view panoramic sequences. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2906–2912. IEEE (2009)
22. Ramirez, P.Z., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Geometry meets semantics for semi-supervised monocular depth estimation. In: Asian Conference on Computer Vision. pp. 298–313. Springer (2018)
23. Ren, Y., Li, S., Chen, C., Kuo, C.C.J.: A coarse-to-fine indoor layout estimation (cfile) method. In: Asian Conference on Computer Vision. pp. 36–51. Springer (2016)
24. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360 imagery. In: Advances in Neural Information Processing Systems. pp. 529–539 (2017)
25. 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)
26. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 707–722 (2018)
27. Tulsiani, S., Tucker, R., Snavely, N.: Layer-structured 3d scene inference via view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 302–317 (2018)
28. Von Gioi, R.G., Jakubowicz, J., Morel, J.M., Randall, G.: Lsd: A fast line segment detector with a false detection control. IEEE transactions on pattern analysis and machine intelligence **32**(4), 722–732 (2008)
29. Wang, X., Fouhey, D., Gupta, A.: Designing deep networks for surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 539–547 (2015)
30. 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)
31. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5684–5693 (2019)
32. Zhang, J., Kan, C., Schwing, A.G., Urtasun, R.: Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1273–1280 (2013)
33. 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. Springer (2014)

34. Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 448–465 (2018)
35. 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)
36. Zou, C., Guo, R., Li, Z., Hoiem, D.: Complete 3d scene parsing from an rgbd image. International Journal of Computer Vision **127**(2), 143–162 (2019)