Shape-Net: Room Layout Estimation from Panoramic Images Robust to Occlusion using Knowledge Distillation with 3D Shapes as Additional Inputs

Mizuki Tabata Kana Kurata Junichiro Tamamatsu
Nippon Telegraph and Telephone Corporation
{mizuki.tabata.fv, kana.kurata.cb, junichirou.tamamatsu.yb}@hco.ntt.co.jp

Abstract

Estimating the layout of a room from a single-shot panoramic image is important in virtual/augmented reality and furniture layout simulation. This involves identifying three-dimensional (3D) geometry, such as the location of corners and boundaries, and performing 3D reconstruction. However, occlusion is a common issue that can negatively impact room layout estimation, and this has not been thoroughly studied to date. It is possible to obtain 3D shape information of rooms as drawings of buildings and coordinates of corners from image datasets, thus we propose providing both 2D panoramic and 3D information to a model to effectively deal with occlusion. However, simply feeding 3D information to a model is not sufficient to utilize the shape information for an occluded area. Therefore, we improve the model by introducing 3D Intersection over Union (IoU) loss to effectively use 3D information. In some cases, drawings are not available or the construction deviates from a drawing. Considering such practical cases, we propose a method for distilling knowledge from a model trained with both images and 3D information to a model that takes only images as input. The proposed model, which is called Shape-Net, achieves state-of-the-art (SOTA) performance on benchmark datasets. We also confirmed its effectiveness in dealing with occlusion through significantly improved accuracy on images with occlusion compared with existing models.

1. Introduction

Room layout estimation from panoramic images is widely used for 3D room modeling, including applications in virtual reality, augmented reality, and furniture arrangement. The method estimates the positional relationships of the components of a room, e.g., corners, boundaries, and wall surfaces, even without directly estimating the layout, on the basis of the Manhattan World assumption [5] that all walls are orthogonal. Since the geometric information of room components varies among different room types, deep learning has been successful in solving this problem. In recent years, the development of deep neural networks has facilitated remarkable advancements in the estimation of room layouts from a single panorama image [15, 30, 31, 35, 38, 45]. From the perspective of data capacity, the method of estimating layouts using a single panoramic image is a favored approach.

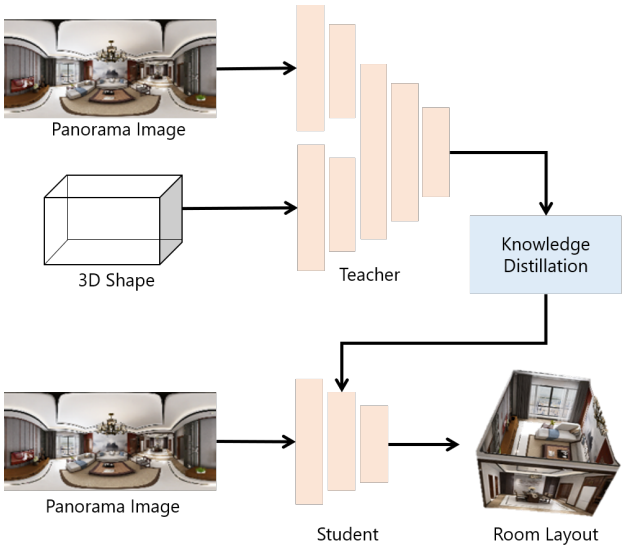

Figure 1. Overall architecture of Shape-Net. Teacher model is trained in advance by providing 3D shape with images, and extracted features of training model are used while student model is trained with only images as input. Trained student model performs inference of room layouts.

Occlusion frequently occurs in panoramic images of rooms since it is difficult to position a camera in such a way that all walls are visible in rooms with intricate shapes. Occlusion degrades the quality of layout estimation as it hides the room components, so it is one of the main issues to be addressed. Conventional approaches that solely rely on a 2D panoramic image face challenges to solve the occlusion

issue, as there is a lack of compensating information behind the occlusion. Therefore, additional information is required to complement occluded areas. 3D shape information, such as drawings contained in layout data, can be used to supplement data surrounding an occlusion. Using 3D information for occluded regions raises two issues: 1) providing 3D shape information to a model does not guarantee its effective use, particularly for areas where occlusion occurs, and 2) although the dataset used for training includes 3D information, in practice, there may be cases where the drawings are unavailable, or reconstruction has resulted in deviations from the drawings. To address these issues, we propose Shape-Net: a Transformer [7]-based knowledge distillation model with a novel 3D IoU loss function (Fig. 1).

To solve the first issue, 3D IoU loss is introduced into the proposed model. The loss calculates the IoU of the ground truth and estimated room shape to take into account the volume of the room and does not decrease that much for areas of occlusion, which works well on occlusion. The knowledge distillation model resolves the second issue, as a student model enables inference from the input of only images while incorporating the training results from a teacher model that takes both images and 3D shapes as inputs. Shape-Net outperforms state-of-the-art (SOTA) models on benchmark datasets. We also tested our model on one of the dataset consisting only of scenes with occlusion. The test results show that the model achieves the highest accuracy, and the difference in accuracy increases compared with that on the dataset without occlusion. This indicates that Shape-Net is robust to occlusion.

The network architecture of Shape-Net allows use in situations where a pair of an image and a drawing of a room does not match, or an image of a room differs from a drawing. In addition, the student model does not require layers to process 3D input, and it gains 3D information by knowledge distillation from the teacher model, which shortens the inference time.

The main contributions of this paper are summarized as follows:

- We introduce a 3D IoU loss function for room layout estimation.
- We propose a knowledge distillation model that infers from a single-shot image while using the training results from the input of both images and 3D shape.
- We evaluate the proposed model on benchmark datasets and demonstrate its effectiveness in dealing with occlusion.

2. Related Work

We review studies on layout estimation from a single-shot panoramic image, knowledge distillation, and cross-modality since our study involves the cross-modality of a room image and its corresponding 3D data using knowledge distillation for room layout estimation.

2.1. Layout Estimation

Most studies have estimated room layouts by detecting room geometry, e.g., boundary probability maps of walls and corners [24,30,45], and wall-surface classes [14,38,42]. Methods that detect wall boundaries and corners show higher accuracy than those that detect wall surfaces, and they have been used more frequently in recent years. While most of these studies use the loss of 2D pixel coordinates, LED^2-Net and LGT-Net consider 3D geometric information through differentiable depth rendering [35] and through depth/height loss [15]. Although these studies consider room geometry in 3D spaces, the losses around occlusion are still small as they account for only the horizontal length or vertical height independently. As a result, their models have difficulty compensating for largely occluded areas. IoU loss can overcome this problem because it calculates the volume of a room. Although IoU loss calculation for 3D bounding boxes was proposed for 3D object detection [44], to our knowledge, no study has proposed 3D IoU loss for complex shapes such as L-shapes. We devised an IoU loss for complex shapes for estimating room layouts in this work.

2.2. Knowledge Distillation

In deep learning, models with a large number of layers and parameters typically show superior performance, albeit at an increased computational cost. To mitigate this issue, the method of knowledge distillation is used. In knowledge distillation, a high-accuracy model referred to as a teacher model is trained, and the knowledge gained is utilized to train a lightweight and easily deployable student model. This approach aims to produce models that are lightweight yet comparable in accuracy to their teacher models [9, 10, 12, 28].

Essentially, knowledge distillation uses the teacher model's output to train the student model. The method uses the teacher's output as a soft target for learning, such that the distribution of the student's output is similar to that of the teacher's output, while the training data labels are used as a hard target. Various methods of distilling knowledge have been proposed, including a method of ensuring that the output distribution of the student model is similar to that of the teacher model, as described above [9, 10, 12], and a method of using features from the middle layer as well as the teacher's output [3, 28]. Notably, the latter method is more effective in training deeper networks [11]. Some approaches use privileged information, such as descriptive text or human posture, which is fed to the teacher model during training, and the trained weights are utilized for the

student model's training [23]. In this study, we used the 3D shape as privileged information in the teacher model, and the student model was trained using the soft target loss of the features from the middle layer, as the proposed model has several feature processing modules.

2.3. Cross-Modality

The fusion of information from different modalities has been actively studied in visual question answering [8, 29, 39, 40], and its applications are increasing beyond image and natural language processing. Various fusion methods have been explored, including concatenation [22], bilinear pooling [8, 40], and co-attention [39, 40].

The fusion of 2D and 3D information has been extensively studied in 3D object detection, which involves the use of both images and point clouds as input [18, 20, 37]. Moreover, the fusion of RGB (red, green, blue) images with depth images obtained through LiDAR (light detection and ranging) sensors has been examined [4, 19]. Given a rough alignment of these RGB and depth images, they are generally concatenated to create fused features [19]. Additionally, integrating shape features with image features through simple concatenation has been demonstrated to improve the accuracy of 3D object pose estimation, even in cases where the features are not aligned. [36].

There have been some studies on cross-modality in indoor spaces, such as 2D-2D cross-modality for matching floor plans used by real estate companies with images of individual rooms [21] or 2D-3D cross-modality for mapping the texture images of manholes onto 3D models [33]. However, no studies have investigated the use of 2D-3D modalities for room layout estimation. Our study addresses this gap by exploring the potential of using these modalities for estimating room layouts.

3. Methodology

The proposed modes is a Transformer-based knowledge distillation model using 3D IoU loss. This section describes its network architecture and loss functions.

3.1. Network Architecture

The network architecture of our model is illustrated in Fig. 2. It consists of a teacher and student model. First, the teacher model is trained with panoramic images and point clouds as inputs. Subsequently, the student model is trained with only panoramic images using the trained results from the teacher model to perform inference. These point clouds can be obtained by interpolating between points generated from ground truth image coordinates of corners. The methods of inputting point clouds are discussed in 4.6.

The teacher model begins with extracting image features using the image encoder proposed in [30] and shape features using a shape encoder based on Point-Net [26]. The input sizes are $512 \times 1024 \times 3$ (height, width, channel) for a panorama image and $n \times 3$ for a point cloud. n is the number of points, and was set to 8000 in this work. Both encoders produce a feature sequence $\mathbb{R}^{N \times D}$, where N is 256, and D is 512 in our implementation. In the shape encoder, the input points are multiplied by a 3×3 affine transformation matrix, which is regressed from a set of input points. They are fed into a multilayer perceptron with output sizes of 64, 128, and 256 and then transformed to 256×1 features using max pooling. To match the size of the image features obtained by the image encoder, the shape features are stacked 512 times in the dimensional direction. The image and shape features extracted from each encoder are processed by Transformer encoder layers [7] since global attention of Transformer [27] can capture the global relationships between spatially distant corners and wall boundaries, which leads to solve the occlusion problem. The Transformer encoder contains six multi-head self-attention layers [34] with eight heads each. The features are concatenated to the output $\mathbb{R}^{N \times 2D}$. The feature sequence is processed by a decoder composed of bidirectional long-short-term-memory layers [13] as in [30]. The student model lacks the shape encoder, and the size of image features is $\mathbb{R}^{N \times 2D}$, otherwise the same network architecture as the teacher model. We used the post-processing method proposed in HorizonNet [30].

3.2. Loss Function

We used the mean absolute error loss (L1 loss) for the image coordinates of the ceiling-wall and wall-floor boundaries, denoted as $\mathcal{L}_b$. To design a system to use a point cloud when occlusion occurs, the weight λ is applied to $\mathcal{L}_b$ only in cases of occlusion. In this study, λ was set to five when occlusion was present and one otherwise. The pixels for occlusion were defined to be vertical coordinates in the image of adjacent pixels that are more than five pixels apart in annotation. For the image coordinates of the corners, we used binary cross entropy loss $\mathcal{L}_c$.

Furthermore, we incorporated 3D IoU loss $\mathcal{L}_{IoU}$ to enable the model to account for the structure of a room in three dimensions for enhancing robustness against occlusion. We devised a calculation method that use the summation of cut-out rectangles from a room (Fig. 3). It can be used even for non-cuboid rooms. Initially, we projected the image coordinates of estimated wall boundaries into 3D space. The number of pixels p_w is that of the 3D boundary coordinates between the ceiling and wall and between the wall and floor. We first calculated the cross-sectional area colored in pink in Fig. 3 from the 3D coordinates (x, y, z) generated by the ground truth and $(\hat{x}, \hat{y}, \hat{z})$ from the estimated boundaries. The cross-sectional area for each pixel is V and $\hat{V}$, respec-

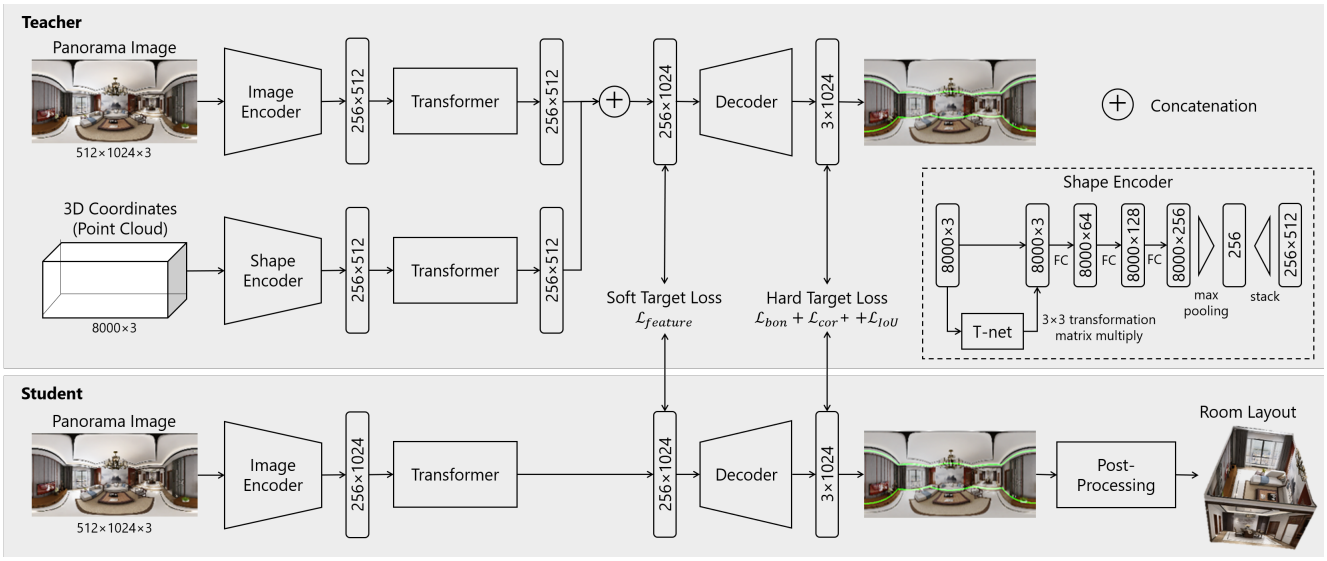

Figure 2. Network architecture of Shape-Net. To train student model, we use both the soft target and hard target loss. Former is computed on the basis of features extracted from middle layer of both teacher and student models, while the latter is determined from labels of datasets. After completion of training, student model is capable of inferring room layouts and reconstructing 3D layouts. Architecture of shape encoder in teacher model is also depicted.

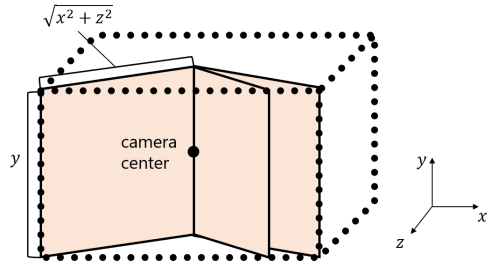

Figure 3. Illustration of IoU calculation for room layout estimation. We calculate IoU by integrating cross sections including camera center colored in pink. 3D coordinates in calculation follow coordinate system in illustration.

tively.

$$V = \sum_{p_w} y\sqrt{(x^2+z^2)} \tag{1}$$

The volume of the intersection is obtained by the following equation where $h \in \{y, \hat{y}\}$ and $w \in \{\sqrt{(x^2+z^2)}, \sqrt{(\hat{x}^2+\hat{z}^2)}\}$.

$$V_{int} = \sum_{p_w} \min hw \tag{2}$$

The volume of the union V_{uni} is $V + \hat{V} - V_{int}$, and IoU is V_{int}/V_{uni}. $\mathcal{L}_{IoU}$ is expressed as follows.

$$\mathcal{L}_{IoU} = 1 - IoU \tag{3}$$

The knowledge distillation uses the teacher's output as a soft target, while it uses the data labels as a hard target. The loss functions: $\mathcal{L}_b$, $\mathcal{L}_c$, and $\mathcal{L}_{IoU}$ are the hard target loss used in both the teacher and the student model. In addition to those loss functions, the student model uses the soft target loss $\mathcal{L}_{soft}$, which measures the L1 loss between the features after concatenation in the teacher model and those after the Transformer layers in the student model (Fig. 2). The total loss function for the teacher model $\mathcal{L}_T$ and that for the student model $\mathcal{L}_S$ are calculated as:

$$\mathcal{L}_T = \lambda\mathcal{L}_b + \mathcal{L}_c + \mathcal{L}_{IoU}. \tag{4}$$

$$\mathcal{L}_S = \lambda\mathcal{L}_b + \mathcal{L}_c + \mathcal{L}_{IoU} + \mathcal{L}_{soft}. \tag{5}$$

4. Experiments

4.1. Implementation Details

Shape-Net was implemented using PyTorch [25], and training was carried out with the Adam optimizer [16] using a batch size of four, learning rate of 0.0001, 1000 epochs on the Pano_S2D3D [1, 41] and Matterport3D [2] datasets, and 50 epochs on the Structured3D dataset [43]. Note that all the results discussed below are calculated by the student model of the best epoch in the validation split. We trained our model on an Nvidia GTX 2080 Ti and an Intel i79700 3.00-GHz CPU.

To augment the input data, we applied horizontal inversion, horizontal rotation, luminescence change, and Pano

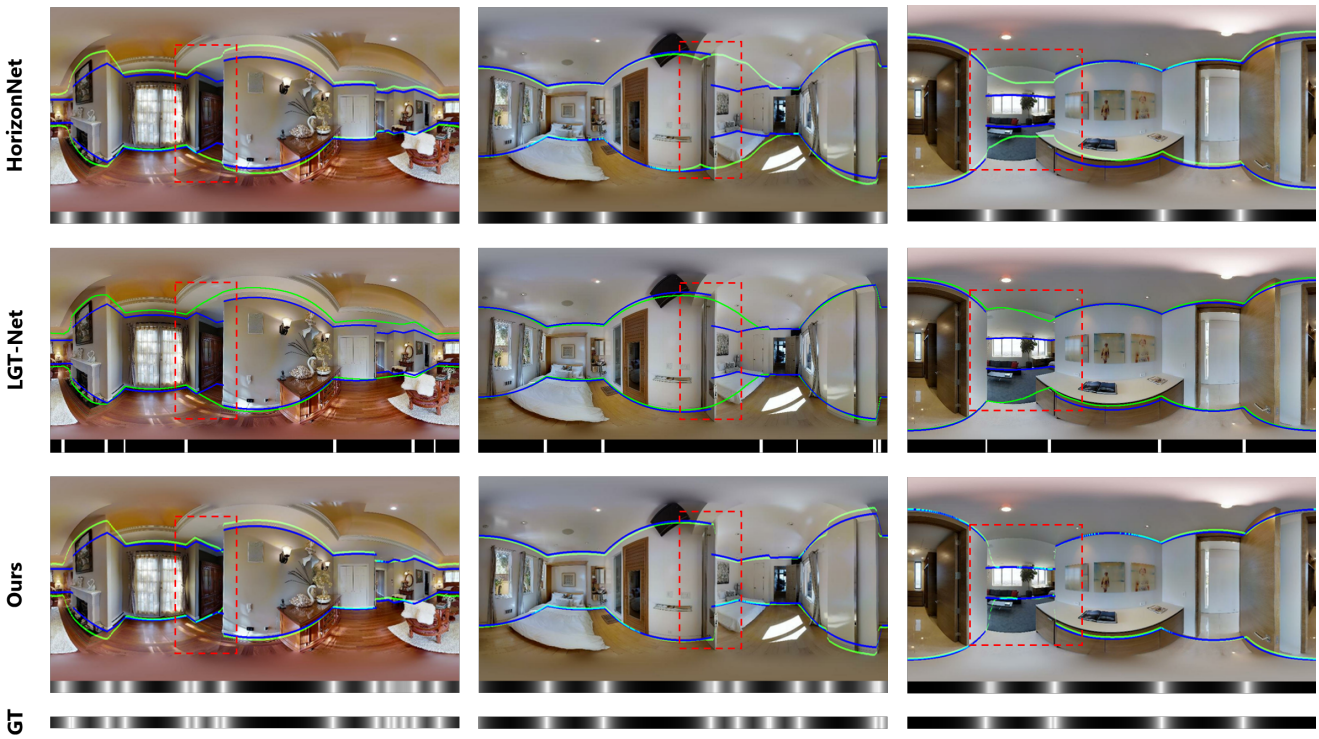

Figure 4. Qualitative results of layout estimation of general rooms without post-processing using HorizonNet [30], LGT-Net [15], and our model on Matterport3D [2]. Blue lines in images show ground truth, while green lines show estimated results of wall boundaries. We also show estimated corner position in white at bottom of image and ground truth in bottom row. Corner locations estimated with LGT-Net were set at point that normal changes for convenience. Areas inside red dashed rectangles indicate where occlusion occurs, and proposed model succeeded in estimating boundaries.

Stretch, which extends the images and annotations in the depth direction [30]. For the input of the point cloud, we applied horizontal inversion and Pano Stretch. To enhance the model's robustness to occlusion, we also trained all models with occlusion-added images by applying the Cutout augmentation [6]. Cutout masks an image with a black square region with fixed side lengths. We chose this augmentation method because it requires fewer model weights than others. To guarantee that critical areas for estimating room layouts were shaved off, we applied masks to images with side lengths of 50 pixels at three arbitrary corner positions in the images. The probability of applying Cutout was set to 50%.

4.2. Datasets

We evaluated the proposed model on three datasets: a dataset that was mix of the PanoContext [41] and Stanford2D-3D [1] datasets (hereafter referred to as Pano_S2D3D), Matterport3D [2], and Structured3D [43]. Evaluation on the Pano_S2D3D datasets has been performed by other models [15, 30, 35], thus we followed the composition of the dataset. While the Pano_S2D3D and Matterport3D datasets [1, 2, 41] consist of real room data, Structured3D [43] is a synthetic dataset.

PanoContext [41] contains 514 room images and Stanford2D-3D [1] contains 552 room images. Pano_S2D3D has only cuboid room layouts. We followed the data split offered by LayoutNet [45]. The data split is composed of 817 pieces of training data, 79 pieces of validation data, and 166 pieces of test data.

Matterport3D [2] contains 2295 room layouts, including non-cuboid rooms. We followed the data split and annotation of LED^2-Net [35]. The data split consists of 1647 pieces of training data, 190 pieces of validation data, and 458 pieces of test data.

We evaluated methods for inputting point clouds using the Structured3D dataset [43]. Structured3D [43] contains 21835 room layouts, and more than 196k photo-realistic 2D renderings of the rooms. We followed the data split and annotation of HorizonNet [30]. The data split consists of 18362 pieces of training data, 1776 pieces of validation data, and 1693 pieces of test data.

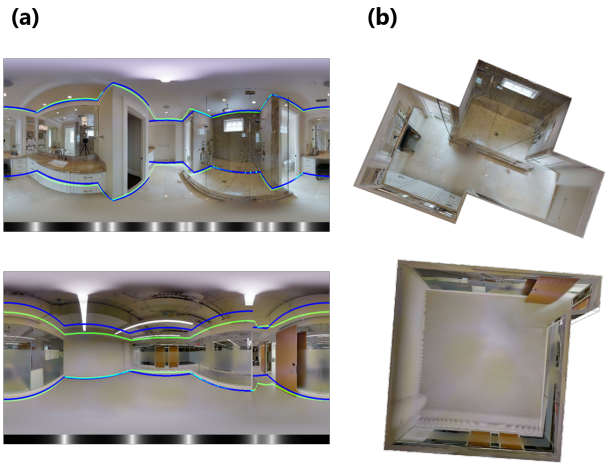

Figure 5. (a) Qualitative results of room layout estimation without post-processing. (b) Visualization of 3D layout corresponding to (a) images.

	3D IoU(%)	CE(%)	PE(%)
HorizonNet [30]	84.61	0.65	**1.89**
LGT-Net [15]	85.29	0.67	2.11
Ours	**86.19**	**0.63**	1.90

Table 1. Quantitative results evaluated on Pano_S2D3D dataset [1, 41].

	2D IoU(%)	3D IoU(%)
HorizonNet [30]	82.51	80.04
LED^2-Net [35]	82.87	80.62
LGT-Net [15]	83.69	81.21
Ours	**83.93**	**81.62**

Table 2. Quantitative results evaluated on Matterport3D dataset [2].

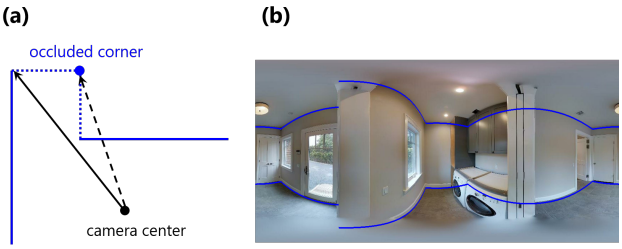

Figure 6. (a) Illustration of occlusion occurring in room. (b) Image when occlusion occurs. Wall boundaries colored in blue line are discontinuous where occlusion occurs in image.

4.3. Overall Performance

We first evaluated our model for cuboid layouts on the Pano_S2D3D dataset [1, 41], as presented in Tab. 1, using the same evaluation metrics proposed in HorizonNet [30]: the IoU of 3D room layouts (3D IoU), corner error (CE), and pixel error (PE). The value of highest accuracy for each metric is shown in bold in all tables. Our model outperformed LGT-Net [15] by 1.06% in 3D IoU and 3.08% in CE. It also achieved almost the same accuracy in PE as HorizonNet [30]. While PE and CE measure the average error between image coordinates of corners and boundaries, the 3D IoU is a metric for evaluating corner errors in 3D space; thus, a higher value in one measure does not necessarily entail a higher value in the other, leading to the above results.

Table 2 shows the performances of our model and those of conventional models for room layouts including non-cuboid rooms on Matterport3D [2]. As CE and PE are evaluation metrics for cuboid rooms, we evaluated the models with the metrics described in [46], i.e., 2D and 3D IoU for non-cuboid rooms. The results in Tab. 2 indicate that our model outperformed all other models. Specifically, our model improved 2D IoU by 0.29% and 3D IoU by 0.50% compared with LGT-Net [15]. The qualitative results in Fig. 4 indicate that our model made improvements in predicting occluded areas with fewer errors compared with other models.

3D layouts reconstructed by our model are illustrated in Fig. 5, exemplifying the precise reconstruction of the room, even in the presence of intricate geometries.

4.4. Results on Occlusion Dataset

To verify the robustness of our model to occlusion, we evaluated the models on 132 images with occlusion from the test split of the Matterport3D dataset [2]. In this experiment, occlusion was defined as a situation where one or more corners of the ground truth are not visible, as illustrated in Fig. 6. The Pano_S2D3D dataset [1, 41] was not included in this experiment because of its cuboid nature, which does not cause occlusion. Quantitative results on the occlusion dataset are presented in Tab. 3. We reported the best performance of our model, a 2.37% improvement in both 2D and 3D IoU compared with LGT-Net [15]. In comparison to Tab. 2, the difference in 2D and 3D IoU between our model and other models was greater on the occlusion dataset. This result suggests that our model is more resistant to occlusion.

To visualize the difference in the results, we used Integrated Gradients: a method for obtaining the contribution of input elements to the output [32]. As shown in Eq. (6), we integrate gradients at all points along a linear path from baseline $x' \in \mathbb{R}^N$ to input $x \in \mathbb{R}^N$. We implemented

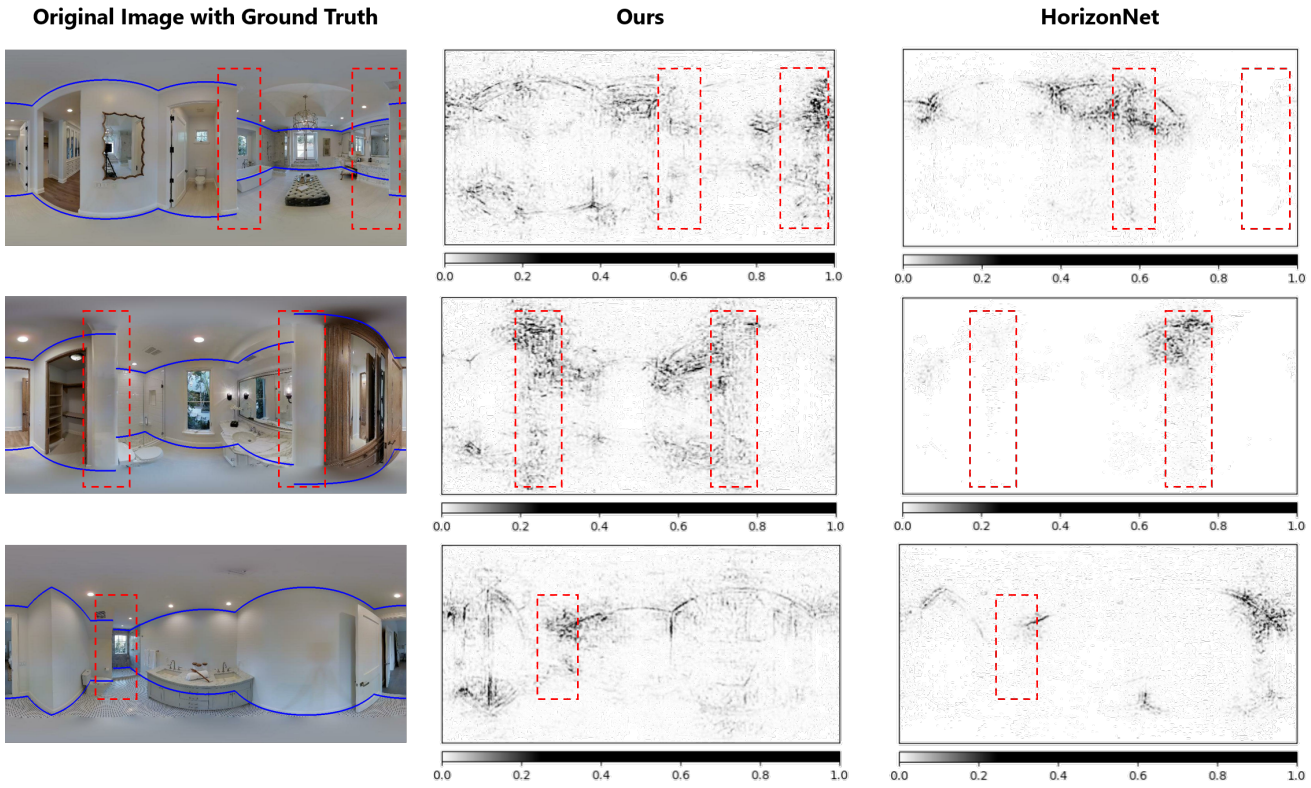

Figure 7. Visualization of Integrated Gradients of each model. Original images are aligned in left column with ground truth of wall boundaries colored in blue. Color bar at bottom of images in two columns on right side shows attribution magnitude; darker color shows higher attribution magnitude. Areas inside red dashed rectangles indicate where occlusion occurs.

the visualization of Integrated Gradients using Captum [17].

$$Integrated\_Gradients_i(x) = (x_i - x'_i) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (6)$$

For visibility, we compared our model with HorizonNet [30], which estimates corners as ours do. The results of Integrated Gradients of each model are depicted in Fig. 7. The black areas had a high magnitude of Integrated Gradients and were more attended by the model. The area enclosed by the red dashed line in our model is darker than in HorizonNet [30], indicating that our model focused more on occluded regions. This suggests that the difference in model attention causes the proposed model to be more resistant to occlusion. Furthermore, our model showed diffused black areas, which indicates that it has a more global attention. This may contribute to increasing in the model's accuracy on overall data.

	2D IoU(%)	3D IoU(%)
HorizonNet [30]	75.92	73.30
LED^2-Net [35]	75.65	73.64
LGT-Net [15]	76.92	75.03
Ours	**78.74**	**76.81**

Table 3. Quantitative results evaluated on occluded data in test split of Matterport3D dataset [2].

4.5. Ablation Studies

We performed ablation studies to assess the effectiveness of Cutout (CUT), knowledge distillation (KD), and IoU loss (IoU) on the Matterport3D dataset [2]. The results are presented in Tab. 4. Model 1 in the index lacks all the mentioned methods, while model 8 in the index incorporates all of them. A comparison of models at indices 2-4 with the model at index 1 demonstrates that each method contributed independently to improving the model accuracy. Cutout shows the smallest effect of the three methods. Notably, the integrating of knowledge distillation and IoU loss (indexed 7) yielded a significant improvement in accuracy.

	Method			Metrics	
Index	CUT	KD	IoU	2D IoU(%)	3D IoU(%)
1				82.53	80.22
2	✓			82.88	80.51
3		✓		83.39	81.13
4			✓	83.41	81.10
5	✓	✓		82.73	80.42
6	✓		✓	83.07	80.69
7		✓	✓	83.84	81.53
8	✓	✓	✓	**83.93**	**81.62**

Table 4. Quantitative results of ablation studies evaluated on Matterport3D dataset [2]. KD, IoU, CUT represents knowledge distillation, IoU loss, Cutout, respectively.

It indicates that the layout information provided by knowledge distillation can be utilized effectively to reduce IoU loss in areas where occlusion occurs. On the other hand, the lower accuracy of models indexed 5 and 6 compared with those indexed 3 and 4 can be attributed to the inability of knowledge distillation and IoU loss alone to handle pseudo-increased occlusion generated by Cutout. Model 8 in the index remarked the highest accuracy of all, which suggests both knowledge distillation and IoU loss can deal with the the pseudo-occlusion image. The difference of models indexed 7 and 8 was relatively small, indicating Cutout plays a supplemental role in layout estimation. These results confirm that the combination knowledge distillation and IoU loss is effective, whereas Cutout augmentation serves as an ancillary method for layout estimation.

4.6. Input Methods

We assessed a method for providing 3D input as a point cloud for layout estimation. Typically, 3D information of rooms is acquired using 3D scanners (e.g., LiDAR) or the room geometry data from drawings. We evaluated three types of point clouds for input to layout estimation: dense pcd, sparse pcd, and layout pcd as shown in Fig. 8. Dense pcd is a point cloud generated from 3D scans, a depth image in this study, and includes all objects observed in the image. Sparse pcd and layout pcd are generated by converting the annotation data on an image to 3D coordinates and interpolating them. Points are generated on planes of walls, ceilings, and floors for sparse pcd, while only on frames for layout pcd.

For evaluation, we used the Structured3D dataset [43] as it contains both RGB and depth panoramic images. Each model was trained with the data augmentation described in the paper, except Cutout [6]. It is because we intended to evaluate the effect of the furniture in the dense pcd on layout estimation. The quantitative results in Tab. 5 demonstrate that the model using layout pcd as input outperformed other

Input	2D IoU(%)	3D IoU(%)
dense pcd	92.59	91.34
sparse pcd	92.74	91.45
layout pcd	**92.83**	**91.61**

Table 5. Quantitative results of models with each input evaluated on Structured3D dataset [43].

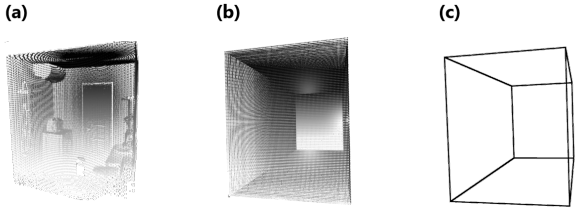

Figure 8. Input point cloud. (a) dense pcd: point cloud including room geometries with furniture generated from depth image. (b) sparse pcd: sparse point cloud including walls generated from annotation on image. (c) layout pcd: point cloud including only wall boundaries generated from annotation on image.

models in 2D and 3D IoU. The furniture in the dense pcd occluded some parts of the room geometries, which may have led to reduced accuracy. The results of sparse pcd and layout pcd indicate that walls in the sparse pcd disturb layout estimation. Thus, we verified the layout pcd is the most effective method of feeding point clouds to our model, and adopted layout pcd for the input to our model in the paper.

5. Conclusion

This paper proposes a novel model that estimates room layouts, taking into consideration the presence of occlusions. To achieve this, we distilled the knowledge from panoramic images and 3D coordinates as inputs, and utilized a 3D IoU loss function. Knowledge distillation allows the model to estimate layouts of rooms even without drawings. Our model outperformed existing models on benchmark datasets, and we demonstrated the robustness of our model to occlusion through evaluation on the dataset of occluded images and visualization of the model's attention. Furthermore, the efficacy of the proposed modules is confirmed through ablation studies.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 4, 5, 6

[2] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from

rgb-d data in indoor environments. *International Conference on 3D Vision*, pages 667–676, 2017. 4, 5, 6, 7, 8

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *International Conference on Neural Information Processing Systems*, pages 742–751, 2017. 2

[4] Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, and Hongdong Li. Noise-aware unsupervised deep lidar-stereo fusion. In *Conference on Computer Vision and Pattern Recognition*, pages 6332–6341, 2019. 3

[5] James M. Coughlan and Alan Loddon Yuille. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *Neural Information Processing Systems*, pages 845–851, 2000. 1

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. 5, 8

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2, 3

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016. 3

[9] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616, 2018. 2

[10] Jiyang Gao, Zhen Li, Ram Nevatia, et al. Knowledge concentration: Learning 100k object classifiers in a single cnn. *CoRR*, abs/1711.07607, 2017. 2

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 2

[12] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*, pages 1–9, 2015. 2

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 3

[14] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Conference on Computer Vision and Pattern Recognition*, pages 5134–5143, 2017. 2

[15] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2022. 1, 2, 5, 6, 7

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2015. 4

[17] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020. 7

[18] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *International Conference on Intelligent Robots and Systems*, pages 1–8, 2018. 3

[19] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022. 3

[20] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *European Conference on Computer Vision*, pages 641–656, 2018. 3

[21] Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. Deep multi-modal image correspondence learning. *CoRR*, abs/1612.01225, 2016. 3

[22] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *CoRR*, abs/1805.11730, 2018. 3

[23] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. 3

[24] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *International Conference on Computer Vision*, pages 936–944, 2015. 2

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 4

[26] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 3

[27] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, pages 12116–12128, 2021. 3

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, pages 1–13, 2015. 2

[29] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Dualnet: Domain-invariant network for visual question answering. In *International Conference on Multimedia and Expo*, pages 829–834, 2017. 3

[30] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 1, 2, 3, 5, 6, 7

[31] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 1

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017. 6

[33] Mizuki Tabata, Yosuke Takeuchi, Yasuhiro Yao, Ryou Tanaka, and Junichirou Tamamatsu. 3d mapping for panoramic inspection images to improve manhole diagnosis efficiency. In *International Symposium on System Integration*, pages 590–595, 2022. 3

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 3

[35] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360° layout estimation via differentiable depth rendering. In *Conference on Computer Vision and Pattern Recognition*, pages 12956–12965, 2021. 1, 2, 5, 6, 7

[36] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3D objects. In *British Machine Vision Conference*, 2019. 3

[37] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018. 3

[38] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Conference on Computer Vision and Pattern Recognition*, pages 3363–3372, 2019. 1, 2

[39] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 3

[40] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *International Conference on Computer Vision*, pages 1839–1848, 2017. 3

[41] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European Conference on Computer Vision*, pages 668–686, 2014. 4, 5, 6

[42] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 10–18, 2017. 2

[43] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pages 519–535, 2020. 4, 5, 8

[44] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2d/3d object detection. In *International Conference on 3D Vision*, pages 85–94, 2019. 2

[45] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *Conference on Computer Vision and Pattern Recognition*, pages 2051–2059, 2018. 1, 2, 5

[46] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360° image: A comparative study of state-of-the-art methods. *International Journal of Computer Vision*, 129:1410–1431, 2021. 6