Geometric Structure Based and Regularized Depth Estimation From 360° Indoor Imagery

Lei Jin[1*] Yanyu Xu[1*] Jia Zheng[1] Junfei Zhang[2] Rui Tang[2] Shugong Xu[3]
Jingyi Yu[1] Shenghua Gao[1†]
[1]ShanghaiTech University [2]KooLab, Kujiale.com
{jinlei,xuyy2,zhengjia,yujingyi,gaoshh}@shanghaitech.edu.cn {ahui,ati}@qunhemail.com
[3]Shanghai University
shugong@shu.edu.cn

Abstract

Motivated by the correlation between the depth and the geometric structure of a 360° indoor image, we propose a novel learning-based depth estimation framework that leverages the geometric structure of a scene to conduct depth estimation. Specifically, we represent the geometric structure of an indoor scene as a collection of corners, boundaries and planes. On the one hand, once a depth map is estimated, this geometric structure can be inferred from the estimated depth map; thus, the geometric structure functions as a regularizer for depth estimation. On the other hand, this estimation also benefits from the geometric structure of a scene estimated from an image where the structure functions as a prior. However, furniture in indoor scenes makes it challenging to infer geometric structure from depth or image data. An attention map is inferred to facilitate both depth estimation from features of the geometric structure and also geometric inferences from the estimated depth map. To validate the effectiveness of each component in our framework under controlled conditions, we render a synthetic dataset, Shanghaitech-Kujiale Indoor 360° dataset with 3550 360° indoor images. Extensive experiments on popular datasets validate the effectiveness of our solution. We also demonstrate that our method can also be applied to counterfactual depth.

1. Introduction

Depth estimation is a fundamental task in vision research, with widespread applications from map reconstruction and navigation in robotics [32] to general scene understanding in the 3D world. With the recent emergence of portable and compact spherical cameras, estimating depth data from omnidirectional [1] content is gaining more attention, as this is a natural solution for many indoor applications.

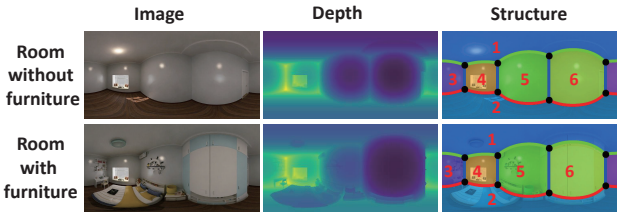


Figure 1. The panorama images and their corresponding depth maps and structures. These two images correspond to the same room. The number indicates the order of the planes.

Some research [22, 31, 39] has been conducted on depth estimation from 360° imagery. These research adapt the solutions of perspective depth estimation to 360° imagery or propose various types of distortion-aware convolution filters. However, few have explored the large FOV nature provided by omnidirectional images — a typical indoor panoramic image, for example, covers global structural information contained within the whole room.

Considering these characteristics of 360° imagery, we present a deep learning framework that leverages geometric structure for indoor depth estimation. Specifically, as shown in Fig. 1, the geometric structure of an indoor scene, which is usually characterized by corners, plane-plane intersection lines and planes, is closely correlated with depth. Depth within the same geometric primitives show strong patterns; related to the location and height of the camera; meanwhile, depth provides a strong clue towards inferring the geometric

*Equal contribution
†Corresponding author

[1]Omnidirectional, spherical and 360° are used interchangeably in this paper.

structure of indoor scenes. Thus we propose the estimating of depth data by facilitating geometric structures with a two-stage solution. In particular, in the first stage, we extract the geometric structure from a 360° image and use features corresponding to the geometric structure for depth estimation. Here the geometric structure functions as a prior; in the second stage, we propose the estimation of this structure by leveraging the depth map from the first stage, thus using the geometric structure as a regularizer. However, indoor scenes usually have furniture, which can affect both the depth estimation from the structure and the structure estimation from depth data. Thus we propose leveraging an attention module to avoid this, and such a module should ideally correspond to the confidence on whether a pixel is occluded by furniture or not.

To facilitate a performance evaluation of our solution, in addition to using publicly available datasets with real scenes, we also build a synthetic dataset. Our dataset contains 3550 images corresponding to 1775 rooms, and each room corresponds to two images whose only difference is whether the furniture is visible or not. This synthetic dataset contains RGB omnidirectional images, their corresponding depth data, corners, plane-plane intersection lines and planes. This synthetic dataset may facilitate the evaluation of geometric structures as both priors and regularizers.

Apart from depth estimation, another interesting line of application is counterfactual depth [15]: estimating non-furniture depth given the object mask. We demonstrate that our proposed representation is also beneficial for such a task.

The contributions of this paper are summarized as follows: i) We propose the representation of an indoor panorama as a collection of geometric structures with points, lines and planes. Such representation is beneficial for depth estimation, 3D reconstruction and counterfactual depth; ii) We propose leveraging geometric structure as both a prior and a regularizer in a novel framework for depth estimation; iii) We build a synthetic dataset for performance evaluation. Extensive experiments on popular datasets validate the effectiveness of our approach. Our dataset: Shanghaitech-Kujiale Indoor 360° dataset is available at https://svip-lab.github.io/dataset/indoor_360.html.[2]

2. Related Work

2.1. Learning in Panorama

CNNs have demonstrated their effectiveness in many vision tasks not only on planar images, but also on panoramic images. Unlike planar images, the convolution operation used on panoramic images must deal with the distortion problem caused by the equirectangular projection. In [29], Su *et al.* propose the use of convolution kernels of different sizes at different locations under equirectangular projections to compensate for this distortion. The computational overhead of the aforementioned method, however, is high. Moreover, in [8], Deng *et al.* apply a deformable convolution [4] and an active convolution [16] to fisheye lenses, another distortion challenge. In the more recent work of [2], spherical CNNs are proposed for classification or single variable regression tasks. Very recently, more efficient reprojection based approaches have been proposed, including distortion-aware convolution [3, 9, 31, 37], spherical convolutions with crown kernels [36], and spherical convolutions operating on the PHD [22] or unstructured grids [17]. These are designed for various tasks including depth estimation [24, 31, 39], saliency prediction [36], image classification and object detection [3, 24, 37].

2.2. Geometric Understanding in Panorama

In this work, information on geometric structures consists of three key components — points, lines and planes, which are related to the so-called "layout" of existing works. Room layouts specify detailed information regarding the walls in a room. In [6], a dynamic Bayesian network is constructed to reconstruct monocular 3D. In [14], vanishing points and other structure features are combined to produce candidate layouts. PanoContext [35] first generates room layout hypotheses with different image-level evidences and then construct the 3D scene with the global context. Im2Pano3d [28] generates a 360° room with a partial perspective observation as input. More recent work, such as [5, 23], solve the problem by viewing it as a segmentation problem in a deep convolutional neural network. In the 360 domain, the layout is represented as a set of corners and boundary lines. Lately, LayoutNet [40, 41] has formulated the problem from a regression and post-optimization approach. HorizonNet [30] has further incorporated an LSTM model. This approach can be applied to non-cuboid Manhattan layouts. In [12], a method is proposed to estimate the layout of indoor scenes from a panoramic image by extracting structural lines and vanishing points and combining them with additional geometric cues. In this work, the main goal is to apply geometric structures for depth estimation, as opposed to estimating them.

2.3. Depth Prediction

Perspective depth estimation has been an active research topic over the past decade. Recently, CNN-based work has typically achieved state-of-the-art performance, with various consecutive up-convolution layers [20], multi-scale networks [11] or conditional random fields [19]. In addition, other researchers have explored the relationship between depth and other tasks, i.e., segmentation [18] and surface norm [26].

[2]Besides depth estimation, our dataset can also be used to empty room synthesis from a furnished room, and layout estimation.

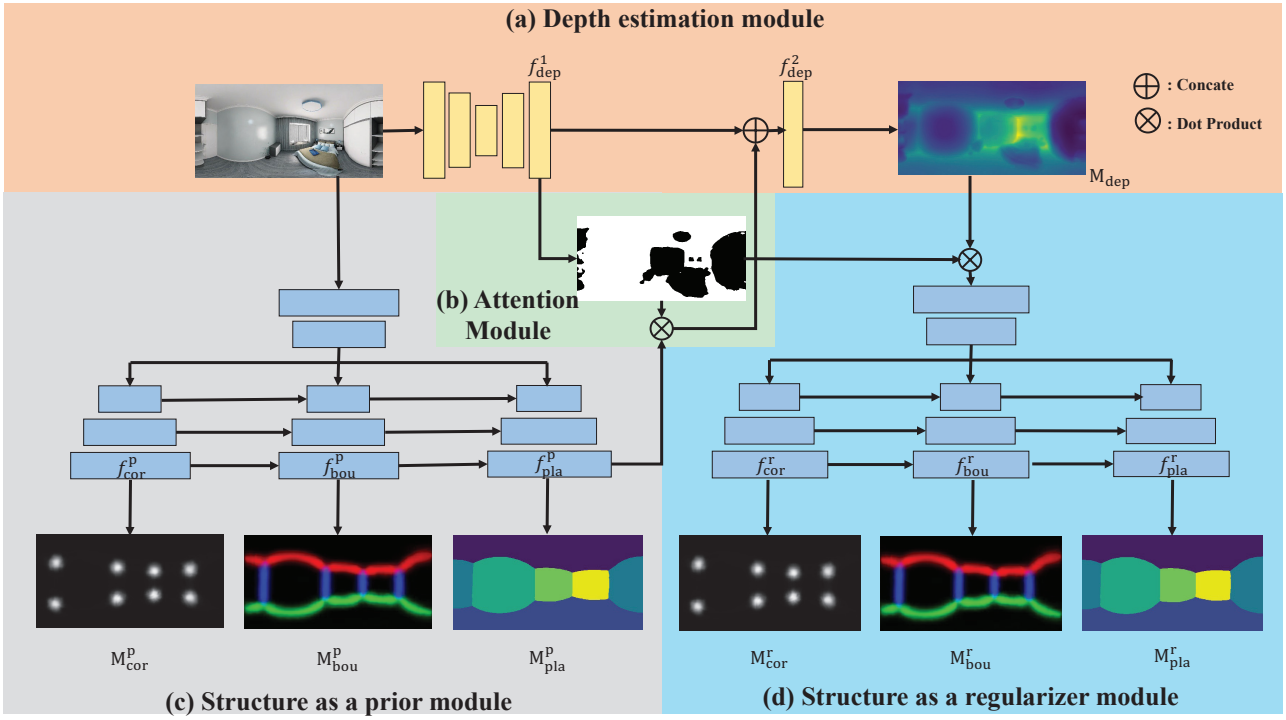


Figure 2. Overview of our architecture. **(a) Depth estimation** predicts depth from a given panorama image. **(c) Structure as a prior** takes panorama as inputs, and estimates the structure of the room. **(b) Attention Module** aims to generate attention map to avoid the inconsistency between structure and depth map caused by furniture. **(d) Structure as a regularizer** is designed to regularize the estimated depth maps by predicting structure from them. We neglect the skip connections in U-Net here for simplicity. Different rectangles with different colors represent convolution blocks.

Recently, depth estimation in panoramic images has been gaining popularity. In [31], Keisuke *et al.* propose the substitution of convolutions with distortion-aware convolutions. In [24], Garanderie *et al.* propose leveraging existing perspective datasets to an omnidirectional domain through a style and projection transfer network in outdoor autonomous driving scenes. A large-scale dataset for indoor depth estimation is proposed in [39], consisting of 22,096 re-rendered images from four existing datasets. This work also proposes two kinds of encoder-decoder networks: UResNet with strided convolutions and RectNet with dilated convolutions. Unlike the above approaches, we use geometric structural information in 360° images to assist and regularize depth estimation.

3. Method

We represent a spherical indoor image as a collection of geometric structures including corners, lines and planes. Due to a lack of annotations, here we only adopt the majority part of the geometric structure in the room, the room layout.

3.1. Overall Architecture

Fig. 2 shows the overall depth estimation network architecture. Given an indoor 360° image $I \in R^{H\times W\times 3}$, our network estimates its depth map M^{gt}_{dep} by leveraging geometric structure as both a prior and a regularizer. Our whole network contains two stages: a geometric structure based depth estimation module and a geometric regularized depth estimation module. Further, to make the network robust to furniture-induced inconsistencies between the depth map and the geometric structure, an attention module is introduced to help both the depth estimation from the geometric structure and the inference of the geometric structure from the depth data. We will introduce these modules in the following sections.

3.2. Geometric Structure Based Depth Estimation

As shown in Fig. 1, for an empty room without furniture, depth is strongly correlated with the geometric structure of the scene: corners are located at local maximum depth, and the depth distribution within the same line or the same plane exhibits a regular pattern. Thus once the geometric structure of a scene is given, it can be used as a prior in depth prediction. We therefore propose leveraging the geometric structure of rooms when conducting depth estimation.

Specifically, we represent the geometric structure of a room as corners, boundaries and planes. Corners are represented by a heat map $\mathrm{M}^{\mathrm{gt}}_{\mathrm{cor}} \in R^{H\times W\times 1}$, where each corner corresponds to a Gaussian centered at a point, with other entries being zeros. Similarly, the boundaries of the room are also represented by a heat map $\mathrm{M}^{\mathrm{gt}}_{\mathrm{bou}} \in R^{H\times W\times 3}$ where boundaries are blurred with a Gaussian. The planar mask is an array of binary maps $\mathrm{M}^{\mathrm{gt}}_{\mathrm{pla}} \in R^{H\times W\times 6}$ where each map corresponds to one plane. To facility the visualization, we currently show it with an index map, whose order is listed in Fig. 1. It is worth noting here that these planes are formulated as semantic segmentation problems rather than surface norm predictions because regression problems are usually more difficult than classification problems.

Given a 360° image, we feed it to U-Net [27] to extract a depth feature f^1_{dep} with the same resolution as the input image. We also feed the image to the LayoutNet [40] to predict geometric structures. Slightly unlike the original LayoutNet, we add another branch to predict planes. Shortcuts are also implemented between the three branches from corner to boundary, and from boundary to plane. In this way, we fuse the representations of our geometric structures in a bottom to top order. We demonstrate that such a representation is more beneficial for general indoor omnidirectional learning in later experiments.

We denote our predicted corner map, boundary map and plane maps at this stage as $\mathrm{M}^{\mathrm{p}}_{\mathrm{pla}}$, $\mathrm{M}^{\mathrm{p}}_{\mathrm{cor}}$ and $\mathrm{M}^{\mathrm{p}}_{\mathrm{bou}}$, respectively, and we denote the features before the last output layers to predict corners, boundaries and planes as $f^{\mathrm{p}}_{\mathrm{cor}}$, $f^{\mathrm{p}}_{\mathrm{bou}}$, and $f^{\mathrm{p}}_{\mathrm{pla}}$, respectively. Following [40], we arrive at the following loss functions for corners, boundaries and planes prediction:

$$\begin{aligned} L^{\mathrm{p}}_{\mathrm{str}} &= L^{\mathrm{p}}_{\mathrm{cor}} + L^{\mathrm{p}}_{\mathrm{bou}} + L^{\mathrm{p}}_{\mathrm{pla}} \\ &= \frac{1}{n} \sum_{c\in \mathrm{M}^{\mathrm{gt}}_{\mathrm{cor}}, \hat{c}\in \mathrm{M}^{\mathrm{p}}_{\mathrm{cor}}} (\hat{c}\log(c) + (1-\hat{c})\log(1-c)) \\ &+ \frac{1}{n} \sum_{b\in \mathrm{M}^{\mathrm{gt}}_{\mathrm{bou}}, \hat{b}\in \mathrm{M}^{\mathrm{p}}_{\mathrm{bou}}} (\hat{b}\log(b) + (1-\hat{b})\log(1-b)) \\ &+ \frac{1}{n} \sum_{p\in \mathrm{M}^{\mathrm{gt}}_{\mathrm{pla}}, \hat{p}\in \mathrm{M}^{\mathrm{p}}_{\mathrm{pla}}} -\hat{p}\log(p) \end{aligned} \tag{1}$$

, where $\hat{c}$ and $\hat{b}$ are the single-pixel probabilities for corner and boundary predictions, $\hat{p}$ is the plane prediction. c, b and p are the ground truths respectively. In addition $n = W \times H$ is the total number of pixels. We can then feed the depth features and the geometric features into another decoder subnetwork for depth refinement. In real scenarios, however, rooms are usually filled with furniture, leading to inconsistencies between the depth data and the geometric structures of the scenes. To tackle this, an attention module is introduced.

3.3. Attention Module

One semantic segmentation branch is added directly after f^1_{dep} to predict a furniture/non-furniture map $\mathrm{M_f}$. The map is used as an attention module to remove the negative influence of the furniture. Given feature maps from $f^{\mathrm{p}}_{\mathrm{pla}}$, f^1_{dep}, we generate a refined feature map as

$$f^2_{\mathrm{dep}} = f^1_{\mathrm{dep}} \oplus (f^{\mathrm{p}}_{\mathrm{pla}} * \mathrm{M}^{\mathrm{p}}_{\mathrm{f}}) \tag{2}$$

where $\oplus$ represents the concatenation operation, and $*$ is the dot product. The reason for using features corresponding to planes rather than corners or boundaries is that planes already contain information on boundaries and corners. Here we use cross entropy loss for furniture map prediction.

$$L_{\mathrm{f}} = \frac{1}{n} \sum_{p\in \mathrm{M}^{\mathrm{gt}}_{\mathrm{f}}, \hat{p}\in \mathrm{M}^{\mathrm{p}}_{\mathrm{f}}} -\hat{p}\log(p) \tag{3}$$

Then we concatenate the attention map weighted structure feature with depth features and feed it to another depth decoder for depth estimation. We denote the predicted depth map before and after the attention map as $\mathrm{M_{dep}}$ and $\hat{\mathrm{M}}_{\mathrm{dep}}$; and by comparing it with a ground truth depth map, we arrive at the following loss function:

$$L_{\mathrm{dep}} = \|\mathrm{M_{dep}} - \mathrm{M^{gt}_{dep}}\|_1 + \|\hat{\mathrm{M}}_{\mathrm{dep}} - \mathrm{M^{gt}_{dep}}\|_1 \tag{4}$$

Note that the generated attention map bridges depth data and structures for real scenes. It can be used for both structure-based depth estimation and structure-regularized depth estimation.

3.4. Structure Regularized Depth Estimation

Depth corresponds to the distance between the camera and the visible regions in a room. For a room without furniture, we can infer the structure of the room based on the estimated depth map because boundaries and corners correspond to local extremes in the depth data. Inspired by this, we propose using structures as regularizers by inferring structures from estimated depth maps. That is, we want our estimated depth maps to also conserve our geometric information.

In practice, however, rooms always contain furniture, and furniture-occluded areas make depth-based structure estimation difficult; thus we propose multiplying the depth map with our inferred attention map $\mathrm{M_f}$. We then feed the output to an auto-encoder in order to infer structural data. The auto-encoder architecture here is almost identical to the auto-encoder in structure-based depth estimation, except that it takes a single-channel attention masked depth map as input.

We denote the predicted corner maps, boundary maps and plane maps during the structure regularized depth estimation stage as $\mathrm{M}^{\mathrm{r}}_{\mathrm{cor}}$, $\mathrm{M}^{\mathrm{r}}_{\mathrm{bou}}$ and $\mathrm{M}^{\mathrm{r}}_{\mathrm{pla}}$ respectively. We arrive

at a similar loss function for corner, boundary and plane predictions as in Section 3.2.

3.5. Training and Inference

We combine the losses corresponding to the structure-based depth estimation stage and the structure-regularized depth estimation stage, arriving at the following objective:

$$L = L_{\text{dep}} + L_{\text{str}}^{\text{p}} + L_{\text{str}}^{\text{r}} + L_{\text{f}} \tag{5}$$

In the training stage, we first train the depth and prior sub-network. Then we use the pretrained parameters as weights to retrain the whole network with prior and regularizer in an end-to-end learning manner. We find that such a pre-training is useful in obtaining performance improvement.

It is worth noting that the geometric structure regularizer helps in learning a more robust depth estimation network. Once the network is trained, in the inference stage, the structure only works as a prior for attention map calculations and depth estimation.

We choose ResNet50 [13] as backbone for depth estimation, and ResNet34 for layout estimation. We implement our solutions under the PyTorch framework and train our network with the SGD optimizer, batch size 8, an initial learning rate 1e-2, weight decay 0.0005 for 30 epochs. We reduce the learning rate by 0.1 whenever we observe plateau following [21]. Finally, we fine-tune the whole network for another 30 epochs. All images are resized to 256*512 with nearest-neighbor during training and testing as in [31].

4. Experiments

In this section, we evaluate our approaches on various datasets. We first demonstrate the effectiveness of our proposed representation on our synthetic dataset. Then we move on to the realistic Stanford 2D-3D-S [1]. We demonstrate quality 3D reconstruction results and quantitative numbers with the standard depth metrics from [39]. Finally, we show that our representation can also be applied to counterfactual depth estimation with some simple modifications.

4.1. Evaluation with Synthetic Dataset

Dataset and experimental setup. Ideally, geometric structural data offers the best assistance in depth estimation for empty rooms. Images collected in real scene datasets [1] always contain furniture. Since there is always furniture in the indoor scenes of existing datasets, it remains a challenge to evaluate structure as both a prior and a regularizer for depth estimation from 360° indoor images. To facilitate the evaluation of the importance of geometric priors and geometric regularizers, we built a synthetic dataset that contains 1775 indoor rooms. Each room has one image with furniture and without furniture. For each image, there is a panoramic RGB image, corresponding depth data, as well as corners, boundaries and planes (as shown in Fig. 3 (a)). With/Without furniture masks are generated by calculating the depth difference between two corresponding depth maps. The images are synthesized with a photo-realistic renderer built upon Embree [33], and we use a well-known path tracing method [25] to achieve realistic rendering, following [38]. Different from [38], our data provides with/without furniture pairs. Fig. 3 (b) and (c) show a comparison of the distributions in terms of the depth distances between our synthetic dataset and existing datasets [1]. We can see that our synthetic dataset is challenging in terms of its depth distributions.

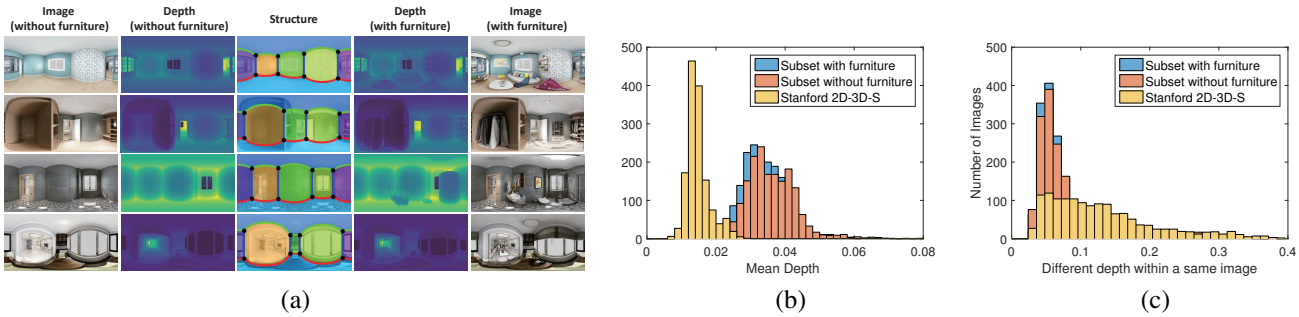
We divide this synthetic dataset into two subsets: a subset with furniture (1775 images) and a subset without furniture (1775 images). The images corresponding to 1500 rooms are used for training, and the remaining 275 rooms for testing. The only difference between the two subsets is whether furniture exists in the room or not. We denote the two subsets as **w.** (with furniture) and **w.o.** (without furniture) in the following context.

Result. In order to evaluate the effect of structure both as a prior and a regularizer, as well as the effect of our attention module for depth estimation, we design two groups for comparison, as shown in Table 1. The first group is designed to evaluate the effect of using geometric structural information. As shown in the first four rows in Table 1, we train our method with and without furniture on two subsets, resulting in four experimental results. On the without furniture subset, we remove the attention branch as it is entirely ones. The refined features here are just a concatenation of depth features and layout features. We can see that the networks trained with structural information perform better than our baselines. In addition, incorporating geometric information brings more improvements over the with furniture subset. On the without furniture subset, the network can itself learn geometric information from input images without further regularization.

In addition, in order to validate the effect of the attention module, we train three networks as shown in the last three rows of Table 1. For the models trained on the with furniture subset and tested on the without furniture subset, we can see that the attention module improves the performance significantly. It also narrows the gap between the models trained on subsets with or without furniture and tested on the subset without furniture. This further demonstrates that our attention module neglects part of the negative impact of furniture on the geometric structure.

4.2. Evaluation on the Stanford 2D-3D-S Dataset

We compare our method with other state-of-the-art approaches on the Stanford 2D-3D-S [1] dataset. This dataset



(a) (b) (c)

Figure 3. Our synthetic dataset: (a) Some images from our synthetic dataset; (b) and (c) The comparison of the distributions in term of the depth distances and the difference of depth within the same image between our synthetic dataset and the Stanford 2D-3D-S dataset.

Train set	Testing set	Structure	RMS ↓	Rel ↓	log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
w.o. fur	w.o. fur		0.668	0.079	0.032	0.948	0.983	0.992
w.o. fur	w.o. fur	√	0.642	0.071	0.029	0.958	0.986	0.992
w. fur	w. fur		0.721	0.114	0.045	0.894	0.973	0.989
w. fur	w. fur	√	0.666	0.103	0.041	0.912	0.978	0.990
Train set	Testing set	Attention	RMS ↓	Rel ↓	log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
w.o. fur	w.o. fur		0.668	0.079	0.032	0.948	0.983	0.992
w. fur	w.o. fur	√	0.730	0.079	0.034	0.943	0.982	0.991
w. fur	w.o. fur		0.784	0.089	0.039	0.927	0.979	0.990

Table 1. Performance comparison under controlled condition on our synthetic dataset. The first block (the first four rows) is to evaluate the effect of the structural information and the second block (the last three rows) aims to validate the effect of the attention module. *w.o. fur* is the subset without furniture and *w. fur* is the subset with furniture. ↑ indicates the higher the better, ↓ the lower the better.

provides a large number of indoor RGB images with corresponding depth data and semantic annotations. In this work, we only use the subset of equirectangular images with layout annotation from [40], which contains 404 images for training and 113 images for testing. Note that during the original layout annotation process, the authors converted the images into an aligned camera pose with the floor. Here we rotate all annotations back to the original view for consistency. With/Without furniture masks are generated from the semantic annotation. We use the ceiling, floor, and wall masks from the original segmentation ground truths as our furniture mask.

Baselines. Following pioneering works on depth estimation from 360° indoor imagery [39], we compare with the following state-of-the-art methods. First, we compare with FCRN [20], one of the state-of-the-art single-model approaches on perspective depth estimation. Then, for a fair comparison, we choose methods designed for dealing with the distortion problem, including UResNet and RectNet [39]. Following [10], we remove the smoothness branch as it may lead to over-smoothed results. This would otherwise lead to a performance decay due to the complexity of the Stanford 2D-3D-S dataset. Following [3], we also replace the planer convolution in FCRN with a spherical convolution and denote this baseline as spherical FCRN. [3]

Result. Table 2 shows the results comparison between our method and other approaches on the Stanford 2D-3D-S dataset. All planar approaches are pretrained on ImageNet [7]. For a fair comparison, spherical approaches are pretrained on our own proposed dataset. RectNet only achieves 0.269 Rel[m] without pretraining, which further validates the effectiveness of our synthetic dataset. Similar results can be observed with UResNet and Spherical FCRN. Interestingly, U-Net [27] pretrained on the ImageNet achieves the best results of all these approaches, and we hence adopt it as our baseline. In general, we believe that some high-level filters trained from perspective images are also useful to omnidirectional images with some simple fine-tunings. That is the reason why an ImageNet pretrained network can achieve the best performance.

From the tables, we can conclude that the results with distortion-aware convolutions are better than those with standard convolutions, which shows the dominating effect of the distortion-aware convolution in panoramic images. In addition, our method outperforms other state-of-the-art

[3] We follow the [40] and use planner convolution rather than spherical convolution due to its good performance for layout estimation from 360° imagery. Further, spherical convolution requires a large memory caused by the bi-linear interpolation and leads to out-of-memory issue.

Methods	RMS[m] ↓	Rel[m] ↓	log10 ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
FCRN [20]	0.534	0.164	0.073	0.749	0.941	0.986
UResNet [39]	0.590	0.187	0.084	0.711	0.921	0.973
RectNet [39]	0.577	0.181	0.081	0.717	0.929	0.979
Spherical FCRN [31]	0.523	0.145	0.067	0.783	0.948	0.986
Ours-baseline	0.472	0.140	0.062	0.803	0.959	0.991
Ours-full	0.421	0.118	0.053	0.851	0.972	0.993

Table 2. Performance comparison on the Stanford 2D-3D-S dataset.

methods, which validates the effectiveness of our architecture.

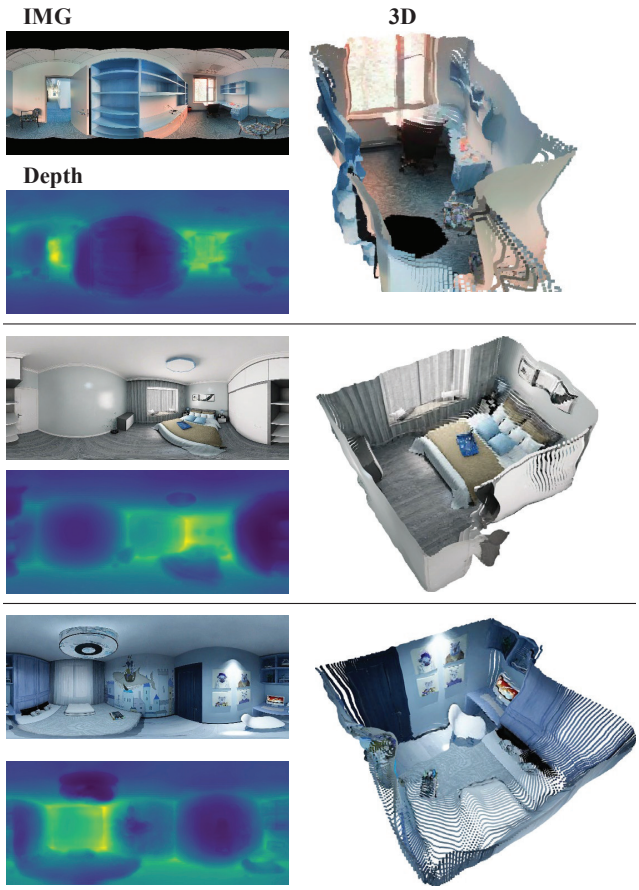


Figure 4. Quality reconstruction results. From left to right: input RGB, predicted depth and reconstruction results. The first image is from the Stanford 2D-3D-S dataset and the last three are from our own with furniture subset.

4.3. Ablation Study

The effect of geometric structures. To show the importance of corners, boundaries, and planes in geometric structure, we remove each component in our method one at a time, and show the results in the first block of Table 3. We also remove the prior module and regularizer module at the same time. This corresponds to our simple U-Net baseline. By comparing these results with our model, we can see that all of the structure elements used in our solution contribute to performance improvement. Overall, boundaries play the most important part, as they are the linkages between points and planes. Without boundaries, it can be difficult to directly infer planes from points.

The effect of structure as a prior and regularizer. To demonstrate the effectiveness of structure as a prior and a regularizer, we remove each module separately. The results are shown in the second block of Table 3. By comparing these baselines with the baseline, we can see that both contribute to performance improvement.

Visualization We also show the predicted depth map corresponding to the direct depth regression without structure (**DR**), the DR with structure but no regularizer module (**DR+structure**) and our full approach (**Ours**) in Fig. 5. In general, our model can extract more of a room's fine details while still preserving the global structure. We further demonstrate this with some reconstruction results in Fig. 4.

5. Application on Counterfactual Depth

Counterfactual depth is first proposed in [15]. It refers to the estimation of non-furniture depth given an image and object mask as input. We demonstrate that our representation can also be beneficial to such a task with some simple modifications. Specifically, in our dataset, the target is to recover the depth of an empty room given the full room and full object masks. We remove the attention branch, and use the object mask directly as input. Object mask is concatenate with f^2_{dep}. We compare our method with the following approaches: (1) **DirectReg**. We train a U-Net with full images as input and non-furniture depth as output. (2) **CounterDepth**. We follow [15] and concatenate the object mask with each upsampling block in U-Net. We also implement LayoutNet [40, 41] by directly appending a plane parameter branch after the network output. We use full images as the input and fit the plane parameter ground truth from empty depth data with RANSAC. The parameter branch does not converge well as is pointed out in the original paper. LayoutNet only aim to recover the corners and lines. It is hard to infer the plane parameters without

Prior	Regularizer	Point	Boundary	Plane	RMS ↓	Rel ↓	log10 ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
×	×	×	×	×	0.472	0.140	0.062	0.803	0.959	0.991
√	√	×	√	√	0.425	0.120	0.055	0.849	0.972	0.993
√	√	√	×	√	0.436	0.122	0.055	0.838	0.970	0.993
√	√	√	√	×	0.429	0.123	0.055	0.845	0.970	0.993
√	√	√	√	√	0.421	0.118	0.053	0.851	0.972	0.993
×	×	×	×	×	0.472	0.140	0.062	0.803	0.959	0.991
√	×	√	√	√	0.446	0.127	0.058	0.823	0.968	0.993
×	√	√	√	√	0.448	0.129	0.057	0.829	0.964	0.991
√	√	√	√	√	0.421	0.118	0.053	0.851	0.972	0.993

Table 3. Ablation study on Stanford 2D-3D-S dataset. The first block (the first five rows) is used to evaluate the effect of each geometric component, where boundaries contributes the performance improvement most. The second block (the last four rows) is used to evaluate the effect of structure as a prior and regularizer, respectively

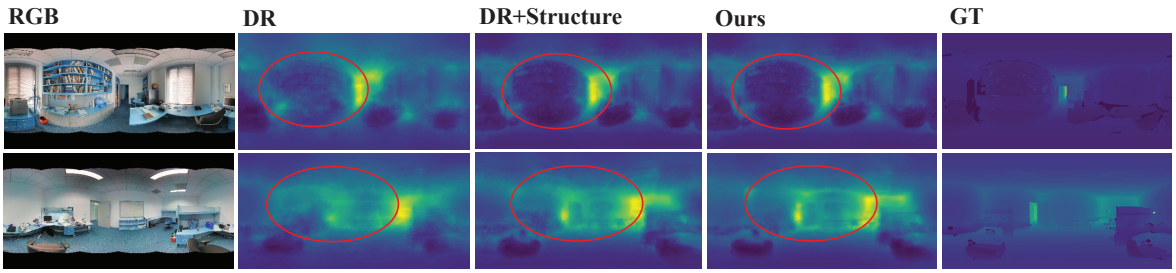


Figure 5. Visualization on the Stanford 2D-3D-S dataset. From left to right: panoramic images, direct regression (DR), DR with structure (DR+Structure), Ours and ground truth depth (GT).

any prior information with planes. On contrast, our representation further includes a plane branch, which makes it a better choice for general indoor omnidirectional learning. We demonstrate quality reconstruction results in Fig. 6.

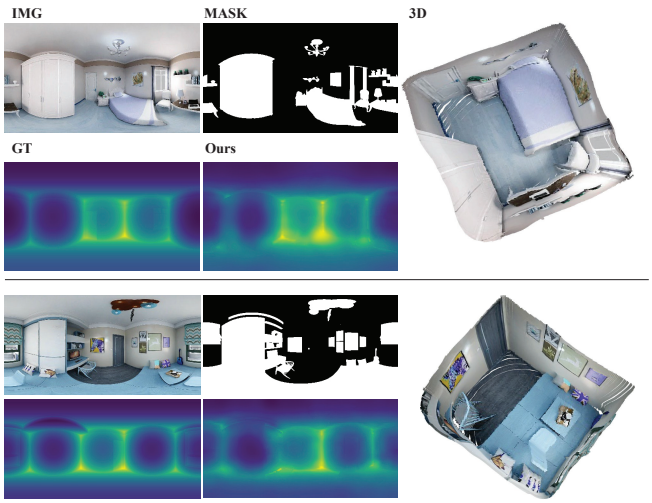


Figure 6. Two quality counterfactual depth estimation results. From top left to bottom right: the input with-furniture image, object mask, ground truth non-furniture depth and our predicted depth respectively. On the right is the reconstructed result in 3D. A rough room shape can be recovered from the input image.

Methods	RMS ↓	Rel ↓	log10 ↓
DirectReg	0.893	0.112	0.046
CounterDepth[15]	0.845	0.104	0.043
Ours	0.823	0.099	0.040

Table 4. Counterfactual depth estimation.

6. Conclusion

We propose a structure based and regularized framework to estimate depth from 360° imagery. In detail, we present geometric structures as corners, boundaries and planes. Then we use this structure information as a prior to help with depth estimation. We build a synthetic dataset to evaluate the effect of structure and the attention module under controlled conditions. In the future, if instance-level object annotation is provided, we can introduce another instance segmentation branch following [34], and this branch may further improve the performance.

Acknowledgements

The work was supported by National Key R&D Program of China (2018AAA0100704), NSFC #61932020, ShanghaiTech-Megavii Joint Lab and partially by NSFC #61871262.

References

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

[2] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

[3] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision*, pages 518–533, 2018.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017.

[5] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–624, 2016.

[6] Erick Delage, Honglak Lee, and Andrew Y Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 2418–2428, 2006.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[9] Marc Eder and Jan-Michael Frahm. Convolutions on spherical images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–5, 2019.

[10] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In *International Conference on 3D Vision*, pages 76–84, 2019.

[11] D Eigen, C Puhrsch, and R Fergus. Prediction from a single image using a multi-scale deep network. In *Proc. Conf. Neural Information Processing Systems*, volume 2, page 4, 2014.

[12] Clara Fernandez-Labrador, Alejandro Perez-Yus, Gonzalo Lopez-Nicolas, and Jose J Guerrero. Layouts from panoramic images with geometry and deep learning. *IEEE Robotics and Automation Letters*, 3(4):3153–3160, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1849–1856, 2009.

[15] Theerasit Issaranon, Chuhang Zou, and David Forsyth. Counterfactual depth from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[16] Yunho Jeon and Junmo Kim. Active convolution: Learning the shape of convolution for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 4201–4209, 2017.

[17] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, and Matthias Niessner. Spherical CNNs on unstructured grids. In *International Conference on Learning Representations*, 2019.

[18] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision*, pages 53–69, 2018.

[19] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *Proceedings of the European Conference on Computer Vision*, pages 143–159, 2016.

[20] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision*, pages 239–248, 2016.

[21] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision*, pages 239–248. IEEE, 2016.

[22] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 9181–9189, 2019.

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

[24] Greire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P Breckon. Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery. In *Proceedings of the European Conference on Computer Vision*, pages 789–807, 2018.

[25] Timothy J. Purcell, Ian Buck, William R. Mark, and Pat Hanrahan. Ray tracing on programmable graphics hardware. *ACM Trans. Graph.*, 21(3):703–712, 2002.

[26] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[28] Shuran Song, Andy Zeng, Angel X Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. In *The IEEE Conference on Computer Vision and Pat-*

tern Recognition, pages 3847–3856, 2018.

[29] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. In *Advances in Neural Information Processing Systems*, pages 529–539, 2017.

[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 *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019.

[31] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision*, pages 707–722, 2018.

[32] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 6243–6252, 2017.

[33] Ingo Wald, Sven Woop, Carsten Benthin, Gregory S. Johnson, and Manfred Ernst. Embree: a kernel framework for efficient CPU ray tracing. *ACM Trans. Graph.*, 33(4):143:1–143:8, 2014.

[34] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1037, 2019.

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

[36] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360 videos. In *Proceedings of the European Conference on Computer Vision*, pages 488–503, 2018.

[37] Qiang Zhao, Chen Zhu, Feng Dai, Yike Ma, Guoqing Jin, and Yongdong Zhang. Distortion-aware cnns for spherical images. In *IJCAI*, pages 1198–1204, 2018.

[38] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling, 2019.

[39] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision*, pages 448–465, 2018.

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

[41] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. 3d manhattan room layout reconstruction from a single 360 image. *CoRR*, abs/1910.04099, 2019.