GPR-NET: MULTI-VIEW LAYOUT ESTIMATION VIA A GEOMETRY-AWARE PANORAMA REGISTRATION NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing 3D layouts from multiple 360° panoramas has received increasing attention recently as estimating a complete layout of a large-scale and complex room from a single panorama is very difficult. A state-of-the-art method, called PSMNet (Wang et al., 2022), introduces the first learning-based framework that jointly estimates the room layout and registration given a pair of panoramas. However, PSMNet relies on an approximate (i.e., "noisy") registration as input. This assumption is inherently a challenging problem in the context of *wide baseline registration*. In this work, we present a complete multi-view panoramic layout estimation framework that jointly learns panorama registration and layout estimation given a pair of panoramas without relying on a pose prior. The major improvement over PSMNet comes from a novel Geometry-aware Panorama Registration Network or GPR-Net that effectively tackles the wide baseline registration problem by exploiting the layout geometry and computing fine-grained correspondences on the layout boundaries, rather on the global pixel-space. Our architecture consists of two parts. First, given two panoramas, we adopt a vision transformer to learn a set of compact 1D horizon features sampled on the panorama. These 1D horizon features encode the depths of individual layouts and the correspondence and covisibility maps between layout boundaries. We then exploit a non-linear registration module to convert these 1D horizon features into a set of corresponding 2D boundary points on the layout. Finally, we estimate the final relative camera pose via RANSAC and obtain the complete layout simply by taking the union of registered layouts. Experimental results indicate that our method achieves state-of-the-art performance in both panorama registration and layout estimation on a large-scale indoor panorama dataset.

1 INTRODUCTION

In this paper, we tackle the problem of room layout estimation from multiple 360° panoramas. Many approaches that can estimate room layouts from a single panorama have been proposed (Zou et al., 2018; Yang et al., 2019; Sun et al., 2019; Pintore et al., 2020). However, these methods did not take advantage of "multi-view" data in which multiple panoramas are taken to better capture a single room. These kinds of data are actually common as evidenced by several indoor datasets such as ZInD (Cruz et al., 2021), Matterport3D (Chang et al., 2017), Gibson (Xia et al., 2018), and Structure3D (Zheng et al., 2020) in which photographers often take multiple panoramas (shoot at different locations) to better capture the views of complex, non-convex rooms that would be occluded from just a single location.

Our work mainly improves upon a recent paper, PSMNet (Wang et al., 2022), that tackled the problem of layout estimation from two panoramas captured in the same room. The idea of PSMNet is to build an architecture that first registers two panoramas in their ceiling view projections and then jointly estimates a 2D layout segmentation. An important aspect of their architecture is that the layout estimation and registration can be trained jointly. However, a major limitation of PSMNet (also mentioned in their paper) is that the architecture relies on an initial approximate registration. The authors argued that such an approximate registration could be given either manually or computed by external methods such as Structure from Motion (SfM) methods or Shabani et al. (2021). While

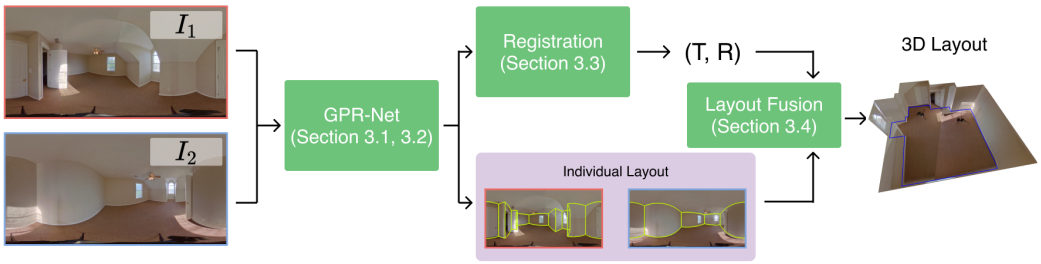

Figure 1: **Proposed multi-view panorama registration and layout estimation framework.** Given two panorama images, a neural network (GPR-Net) jointly predicts layout boundary correspondences and individual layouts for the two panoramas (section 3.1,section 3.2). The correspondences are fed to a registration module to compute the relative camera pose (T, R) (section 3.3). A layout fusion module then computes a unified 3D layout given the camera pose and the individual layouts (section 3.4).

a manual registration may work, it would make the method no longer automatic. As far as using external methods for approximate registration, our observation is that they may make registration errors frequently and even fail to provide a registration in a substantial number of cases. The main reason is that the required registration is mainly the challenging case of wide baseline registration and only two images are given. For example, our results show that the state-of-the-art SfM method OpenMVG (Moulon et al., 2016) fails to register 76% of panorama pairs from our test dataset. It is thus impractical to assume an independent algorithm that can reliably give an approximate solution to the challenging wide baseline registration problem. In addition, relying on such an algorithm moves a critical part of the problem to a pre-process.

Therefore, we set out to develop a complete multi-view panorama registration and layout estimation framework that no longer relies on an approximate registration given as input as shown in Figure 1. To achieve this, we propose a novel Geometry-aware Panorama Registration Network, or GPR-Net, based on the following design ideas. First, our experiments indicate that a global (pixel-space) registration that directly regresses pose parameters (i.e., translation and rotation) is too ambitious. Instead, we propose to compute more fine-grained correspondences in a different space. Specifically, GPR-Net aims at jointly estimating the layout depth of input panoramas as well as the correspondence and covisibility maps on the layout boundary. These feature maps (depth, correspondence, and covisibility) are encoded as a set of compact 1D horizon lines sampled on the panorama. This representation has the advantages of having more elements to register (e.g., 256 samples per panorama), more supervision signal for fine-grained estimation, and thus leading to better learning performance. Second, we build a non-linear registration module to compute the final relative camera pose. The module combines two horizon-depth maps with the horizon-correspondence and horizon-covisibility maps to obtain a set of covisible corresponding 2D boundary points in the ceiling projection, followed by a RANSAC-based pose estimation. Note that this non-linear space is more expressive and can encode a richer range of maps between two panoramas. The final complete layout is obtained simply by taking the union of two registered layouts.

We extensively validate our model by comparing with the state-of-the-art panorama registration method and multi-view layout estimation method on a large-scale indoor panorama dataset ZInD (Cruz et al., 2021). The experimental results demonstrate that our model is superior to competing methods by achieving significant performance boost in both panorama registration accuracy (mAA@5° $\uparrow$ 65.7%, mAA@10° $\uparrow$ 73.1%) and layout reconstruction accuracy (2D IoU $\uparrow$ 4.5%).

In summary, our contributions are as follows:

- We propose the first complete multi-view panoramic layout estimation framework. Our architecture jointly learns the layout and registration from data, is end-to-end trainable, and most importantly, does not rely on a pose prior.

- We devise a novel panorama registration framework to effectively tackle the wide baseline registration problem by exploiting the layout geometry and computing a fine-grained correspondence of the layout boundaries.
- We achieve state-of-the-art performance on ZInD (Cruz et al., 2021) dataset in both stereo panorama registration and layout reconstruction tasks.

2 Related Work

2.1 Single-view room layout estimation

There exist many methods to estimate the room layouts from just a single image taken inside an indoor environment. Methods that take only one perspective image include earlier attempts that relied on image clues and optimization (Hedau et al., 2009; Hoiem et al., 2007; Ramalingam & Brand, 2013) and later neural networks (Lee et al., 2017; Yan et al., 2020). Capturing the increasing availability and popularity of full 360° panoramic images, the seminal work by Zhang et al. (Zhang et al., 2014) proposed to take panoramas as native inputs for scene understanding. Recently, several methods were proposed to predict the room layouts from a single panorama using neural networks. A major difference between these methods is the assumption on the shape of the room layouts - from being strictly a cuboid (Zou et al., 2018), Manhattan world (Yang et al., 2019; Sun et al., 2019), to more recently general 2D layouts (Atlanta world) (Pintore et al., 2020). For our work, we choose to adopt the Manhattan assumption because more corresponding data is available. See Zou et al. (2021) for a thorough survey on predicting Manhattan room layouts from a single panorama. More recent methods delivered state-of-the-art performance by transforming the problem into a depth-estimation one (Wang et al., 2022) or by leveraging powerful transformer-based network architecture (Jiang et al., 2022). Although these single-view methods perform well in the cuboid and L-shape rooms, they tend to fail in the large-scale, complex and non-convex rooms where a single-view panorama covers only part of the whole space due to occlusion.

2.2 Panorama registration

Image registration, i.e., finding transformations between the cameras of two or multiple images taken of the same scene, is a key component of Structure-from-Motion (SfM). See Özyeşil et al. (2017) for a recent survey and Hartley & Zisserman (2003) for an extensive study. Registration problems can be categorized by: 1) the assumptions about the camera model, e.g., perspective (pinhole camera), weak-perspective, or orthographic, 2) the assumptions about the transformation, e.g., rigid, affine, or general non-rigid, and 3) the types of the image inputs, e.g., perspective images or full 360° panoramas, and with/without depths. In addition, the difficulty differs greatly on whether the images are taken densely or sparsely. Modern takes on registration problems often leverage state-of-art programs/libraries such as COLMAP (Schönberger & Frahm, 2016) and OpenMVG (Moulon et al., 2016). Our problem falls into a lesser-studied category: *registering rigid transforms between sparse panoramas*. While there exist methods that tackle sparse perspective image inputs (Salaün et al., 2017; Fabbri et al., 2020) and methods that handle panoramas natively (Pagani & Stricker, 2011; Taneja et al., 2012; Ji et al., 2020), our results show that we can improve upon the state-of-the-art panorama registration methods in our sparse view setting. A key bottleneck is that traditional SfM methods fail to handle the wide baseline registration problem where the views are far apart from each other. In Shabani et al. (2021), SfM of extremely sparse panoramas was tackled by matching room types and specific elements such as doors and windows. In contrast to previous methods that perform the registration in the global pixel-space, we propose a novel learning-based panorama registration framework that directly compute the registration between two panoramas without taking any prior knowledge as input.

2.3 Scene reconstruction using sparse panoramas

Attempts to reconstruct indoor scenes using just a handful of RGB panoramas as inputs Pintore et al. (2018; 2019) are nascent but promising since photographers are adapting 360° cameras into their workflows (e.g., Matterport 3D capture system (Chang et al., 2017)) and it is awkward to capture dense panoramic inputs due to camera/tripod setups. While previous methods assume that all the input panoramas are already registered, the PSMNet (Wang et al., 2022) introduces the first

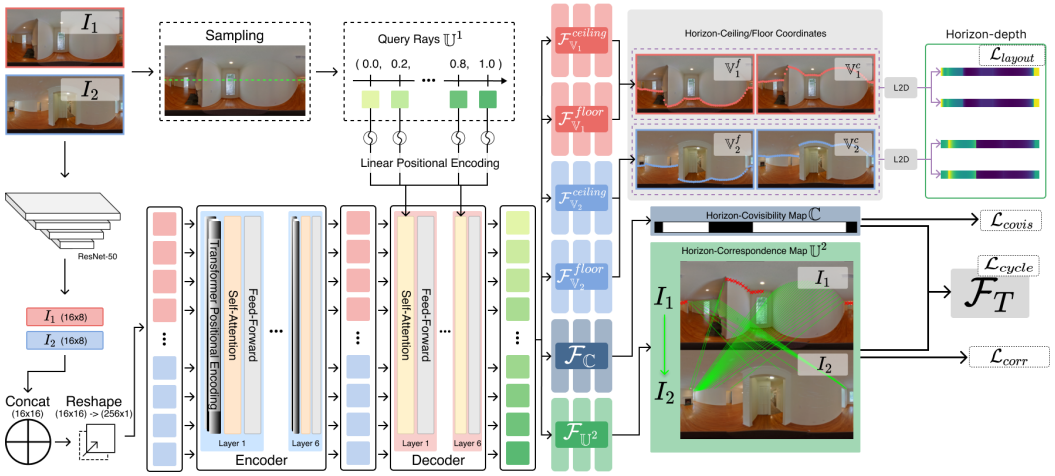

Figure 2: **GPR-Net architecture.** Our network architecture follows the standard vision transformer with an encoder-decoder scheme and multiple MLP heads. Given a pair of panoramas $\mathbf{I^1}$ and $\mathbf{I^2}$, our network first extracts feature maps as input tokens by ResNet-50 and feeds those features into the transformer $\mathcal{F}_T$. We then treat the output tokens as query rays $\mathbb{U}^1$ in panorama $\mathbf{I^1}$ and process the output of the transformer $\mathcal{F}_T$ by different MLP heads for horizon ceiling/floor coordinates, namely $\mathbb{V}^{j=(c,f)}_{k=(1,2)}$, horizon-covisibility map $\mathbb{C}$, and horizon-correspondence map $\mathbb{U}^2$, respectively. We perform L2D transformation on horizon ceiling/floor coordinates $\mathbb{V}^{j=(c,f)}_{k=(1,2)}$ to obtain horizon-depth map $\mathbb{D}$. Finally, we compute the layout loss $\mathcal{L}_{layout}$, covisibility loss $\mathcal{L}_{covis}$, correspondence loss $\mathcal{L}_{corr}$, and cycle-consistentcy loss $\mathcal{L}_{cycle}$ using the corresponding 1D output horizon maps.

learning-based framework that jointly estimates the room layout and registration given a pair of panoramas. However, it still has a major bottleneck that an initial approximate (noisy) registration must be given (e.g., either manually specified or computed by external methods) during both the training and inference stages. Our GPR-Net is also an end-to-end deep neural network that jointly learns the room layout and panorama registration. Most importantly, our model does not rely on a pose prior and is thus suitable for real-world application scenarios.

3 Methodology

3.1 Network Architecture

Figure 2 illustrates the GPR-Net architecture. GPR-Net uses building blocks from COTR (Jiang et al., 2021) and LED^2-Net (Wang et al., 2021). First, we feed two (vertically) axis aligned panoramas $\mathbf{I^1}$ and $\mathbf{I^2}$ into a ResNet-50 (He et al., 2015) feature extractor and generate two feature maps of size 16×8. We then concatenate these two feature maps *side-by-side*, generating a single feature map of size 16×16. Following the vision transformer architecture, we feed the extracted feature maps into the transformer $\mathcal{F}_T$ using each of the 16×16 pixels as an input token. The output tokens of the transformer encode information about boundary segments. We parametrize the boundary using an interval $[0, 1]$ and subdivide the interval into N equally sized segments. We associate each boundary segment with a query ray, e.g. for panorama $\mathbf{I^1}$ we have query rays $\mathbb{U}^1 = \{u^1_i\}^N_{i=1}$. The query rays are encoded by linear positional encoding. The output of the transformer is processed using multiple MLP heads to obtain corresponding 1D horizon features for latter post-process. Specifically, we build: (i) *layout MLP heads* $\mathcal{F}^{ceiling}_{\mathbb{V}_k}$ and $\mathcal{F}^{floor}_{\mathbb{V}_k}$ with $k = (1, 2)$. The outputs of $\mathcal{F}^{ceiling}_{\mathbb{V}}$ and $\mathcal{F}^{floor}_{\mathbb{V}}$ are respectively layout's horizon-ceiling coordinates, $\mathbb{V}^c = \{v^c_i\}^N_{i=1}$, and horizon-floor coordinates, $\mathbb{V}^f = \{v^f_i\}^N_{i=1}$. For each $\mathbb{V}^{j=(c,f)}_{k=(1,2)}$, we further exploit the Layout-to-Depth (L2D) transformation (Wang et al., 2021) to generate a corresponding horizon-depth map $\mathbb{D}^{j=(c,f)}_{k=(1,2)}$; (ii)

correspondence MLP head $\mathcal{F}_{\mathbb{U}^2}$ that outputs a horizon-correspondence map $\mathbb{U}^2 = \{u_i^2\}_{i=1}^N$, where u_i^2 indicates the correspondence between $u_i^1 \in \mathbf{I^1}$ and $u_i^2 \in \mathbf{I^2}$; and (iii) *covisibility MLP head* $\mathcal{F}_{\mathbb{C}}$ that outputs a horizon-covisibility map $\mathbb{C} = \{c_i\}_{i=1}^N$, where c_i is a value $\in [0, 1]$, encoding whether the $i-th$ element in $\mathbb{U}^2$ should be considered ($c_i = 1$) or not ($c_i = 0$) in the pose estimation.

3.2 LOSS FUNCTIONS

Here we elaborate on the layout, correspondence, covisibility, and cycle-consistency loss functions used for training our network.

Layout loss calculates the low-level geometry loss between the predicted horizon-depth maps $\mathbb{D}$ of input panoramas $\mathbf{I^1}$ and $\mathbf{I^2}$ and their ground-truth horizon-depth maps $\overline{\mathbb{D}}$. The loss is defined as follows:

$$\mathcal{L}_{layout} = \frac{1}{M} \sum_{j=(c,f)} \sum_{k=(1,2)} \|\mathbb{D}_k^j - \overline{\mathbb{D}}_k^j\|_1, \tag{1}$$

where M represents the size of the horizon-depth map.

Covisibility loss evaluates the predicted normalized horizon co-visibility map $\mathbb{C} = \{c_i\}_{i=1}^N$ with respect to the ground-truth horizon co-visibility map $\overline{\mathbb{C}} = \{\overline{c}_i\}_{i=1}^N$. The loss is defined as follows:

$$\mathcal{L}_{covis} = \frac{1}{N} \sum_{i=1}^{N} \alpha \overline{c}_i \cdot \log(c_i) + (1 - \overline{c}_i) \cdot \log(1 - c_i), \tag{2}$$

where α is the hyperparameter for weighting the positive samples.

Correspondence loss calculates the difference between the predicted horizon-correspondence map $\mathbb{U}^2 = \{u_i^2\}_{i=1}^N$ and the ground-truth horizon-correspondence map $\overline{\mathbb{U}}^2 = \{\overline{u}_i^2\}_{i=1}^N$. The loss is defined as follows:

$$\mathcal{L}_{corr} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \min(\|u_i^2 - \overline{u}_i^2\|_1, \|1 - \overline{u}_i^2 + u_i^2\|_1), & \text{if } \overline{c}_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where we use a cyclic loss instead of the simple L1 loss between the predicted and ground-truth correspondence to adopt the coordinate system of in equirectangular projection.

Cycle-consistency loss (Jiang et al., 2021) enforces the network outputs to be cycle-consistent and adapts the network to different ray casting positions in contrast to the uniformly sampled ray casting positions. We reverse the order of the two panoramas $\mathbf{I^1}$ and $\mathbf{I^2}$, and treat the ground-truth horizon-correspondence map $\{\overline{u}_i^2\}_{i=1}^N$ as input and the original input $\{u_i^2\}_{i=1}^N$ as target correspondence. We separate the cycle-consistency loss into two parts as follows:

$$\mathcal{L}_{covis}^{cycle} = \frac{1}{N} \sum_{i=1}^{N} \alpha \overline{c}_i \cdot \log(\mathcal{F}_{\mathbb{C}}(\mathcal{F}_T(\overline{u}_i^2))) + (1 - \overline{c}_i) \cdot \log(1 - \mathcal{F}_{\mathbb{C}}(\mathcal{F}_T(\overline{u}_i^2))) \tag{4}$$

and

$$\mathcal{L}_{corr}^{cycle} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \min(\|\mathcal{F}_{\mathbb{U}^2}(\mathcal{F}_T(\overline{u}_i^2)) - u_i^1\|_1, \|1 - u_i^1 + \mathcal{F}_{\mathbb{U}^2}(\mathcal{F}_T(\overline{u}_i^2))\|_1), & \text{if } \overline{c}_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where we still use the ground-truth horizon-covisibility map $\{\overline{c}_i\}_{i=1}^N$ as the target in $\mathcal{L}_{covis}^{cycle}$ since the order of $\{\overline{u}_i^2\}_{i=1}^N$ and $\{u_i^2\}_{i=1}^N$ is the same. The $\mathcal{L}_{cycle}$ is thus defined as follows:

$$\mathcal{L}_{cycle} = \mathcal{L}_{covis}^{cycle} + \mathcal{L}_{corr}^{cycle} \tag{6}$$

Finally, the overall loss function used in our network is defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{layout} + \lambda_2 \mathcal{L}_{corr} + \lambda_3 \mathcal{L}_{covis} + \lambda_4 \mathcal{L}_{cycle}, \tag{7}$$

where λ_1, λ_2, λ_3, and λ_4 are the hyperparameters for weighting the loss functions.

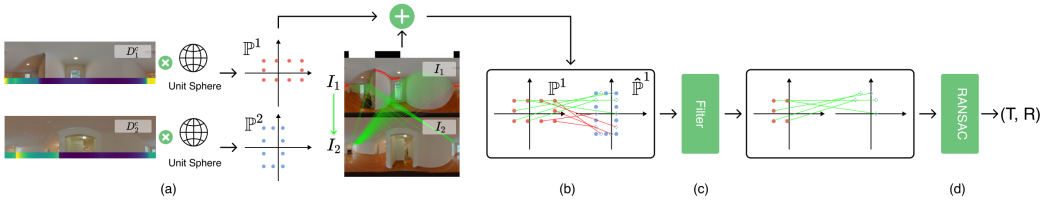

Figure 3: **Registration pipeline**. We start our registration process from the two horizon-depth maps $\mathbb{D}_1^c$ and $\mathbb{D}_2^c$. (a) To compute the two boundary point sets $\mathbb{P}^1$ and $\mathbb{P}^2$, we element-wisely multiply the unit sphere with the two horizon-depth maps $\mathbb{D}_1^c$ and $\mathbb{D}_2^c$. (b) Considering the non-uniformly distributed horizon-correspondence map $\mathbb{U}^2$, we extract the one-to-one point-wise correspondence $\mathbb{P}^1$ and $\hat{\mathbb{P}}^1$ by interpolating $\mathbb{P}^2$ with $\mathbb{U}^2$. Finally, (c) we filter out some of the in-covisible pairs via horizon co-visibility map $\mathbb{C}$ and (d) robustly estimate final pose parameters by RANSAC.

3.3 Non-linear registration

Figure 3 illustrates the pipeline of our non-linear registration. Given two estimated upper (ceiling) horizon-depth maps $\mathbb{D}_1^c$ and $\mathbb{D}_2^c$, we multiply individual horizon-depth maps element-wise with a uniformly-sampled unit sphere to obtain two sets of boundary points $\mathbb{P}^1 = \{p_i^1\}_{i=1}^M$ and $\mathbb{P}^2 = \{p_i^2\}_{i=1}^M$ on the layouts, where M is the number of samples on the unit sphere (Figure 3(a)). Note that the boundary points are on the XZ plane as we only consider the 3-DoF transformations in ZInD (Cruz et al., 2021). In order to compute a one-to-one point-wise correspondence between $\mathbb{P}^1$ and $\mathbb{P}^2$, we could exploit the estimated horizon-correspondence map $\mathbb{U}^2$. Since $\mathbb{U}^2$ is not necessary a uniform distribution, for each $p_i^1 \in \mathbb{P}^1$, we compute its corresponding point via interpolating $\mathbb{P}^2$ with $\mathbb{U}^2$ (Figure 3(b)). We denote the corresponding boundary points as $\hat{\mathbb{P}}^1 = \{\hat{p}_i^1\}_{i=1}^M$ We then filter out the matched pair in $\mathbb{P}^1$ and $\hat{\mathbb{P}}^1$ according to the horizon co-visibility map $\mathbb{C}$ (Figure 3(c)). The final pose parameters (i.e., translation T and rotation R) are computed via a RANSAC-based estimation (Figure 3(d)).

3.4 Layout fusion

Given the relative camera pose between input panoramas, we fuse two individual partial layouts into a complete one as follows. First, for each input panorama, we adopt the same post-processing procedure as LED^2-Net (Wang et al., 2021) that converts the estimated horizon-ceiling coordinates $\mathbb{V}_k^c$ and horizon-floor coordinates $\mathbb{V}_k^f$ into a 2D layout map L_k and a layout height H_k. Then, we register two 2D layout maps using the estimated relative camera pose and fuse them into a complete 2D layout map via a union operation $L_{fusion} = L_1 \cup L_2$. The final 3D layout is obtained by extruding L_{fusion} with the average layout height $(H_1 + H_2)/2$.

4 Results

In this section, we compare our method with state-of-the-art layout reconstruction and panorama registration approaches. We also conducted multiple ablation studies to validate the necessity of individual modules in our architecture.

4.1 Experimental Settings

Dataset. We conduct all the experiments on a public indoor panorama dataset, Zillow Indoor Dataset (ZInD), which contains 67,448 indoor panoramas. We follow the same procedure as PSMNet to select the panorama pair instances and obtain training (105256), validation (12376), and tests (12918) pairs. While we tried to match the PSMNet test protocol as closely as possible and exchanges multiple emails with the authors to that effect, we are still waiting for the authors of PSMNet to release their testing code and dataset split.

Table 1: **Quantitative comparisons with state-of-the-art methods on layout reconstruction.** The * symbol means that the numbers are reported in PSMNet Wang et al. (2022).

	w/ GT pose			w/o GT pose		
Method	2D IoU↑	δ^i ↑	3D IoU↑	2D IoU↑	δ^i ↑	3D IoU↑
LED2Net (Wang et al., 2021)	0.8364	0.9557	0.8131	0.5889	0.8777	0.5738
LGT-Net (Jiang et al., 2022)	0.8388	0.9537	0.8126	0.5831	0.8779	0.5661
PSMNet* (Wang et al., 2022)	0.8101	0.9238	-	0.7577	0.9217	-
GPR-Net (Ours)	**0.8449**	**0.9603**	**0.8211**	**0.8026**	**0.9452**	**0.7816**

Competing methods. We compare our method with the following state-of-the-art layout reconstruction models, said LED^2-Net (Wang et al., 2021), LGT-Net (Jiang et al., 2022), and PSMNet (Wang et al., 2022). Since LED^2-Net and LGT-Net are single-view layout estimation methods, we first estimate the layout for each view, register two input panoramas using OpenMVG, and then perform an union operation to obtain the final reslut. Note that in cases where OpenMVG fails to produce a registration, we use average ground-truth pose of the training datasets. To evaluate the performance on the panorama registration, we compare our GPR-Net with OpenMVG, a popular Structure-from-Motion library that supports stereo panorama matching. We followed the official settings for the feature extractor and correspondence matching strategy and applied 'incrementalv2' mode for SfM operation to achieve higher reconstruction rate.

Evaluation metrics. We used 2D IoU, 3D IoU, and δ^i for quantitative evaluation of the layout reconstruction. As for measuring the image registration quality, we used angular error between estimated relative pose and ground-truth in translation(Δt) and rotation(ΔR) (Brachmann & Rother, 2019) and mean-average-accuracy (Jiang et al., 2021) of translation and rotation at a 5° and 10° error threshold, denoted as T-mAA@5°, T-mAA@10°, R-mAA@5°, and R-mAA@10°, respectively.

Implementation details. We implemented our model in PyTorch and conducted experiments on a single NVIDIA V100 with 32GB VRAM. The resolution of the panoramas is resized to 512×256. We use the Adam optimizer with b1=0.9 and b2=0.999. The learning rates of the transformer and the ResNet-50 are 1e-4 and 1e-5, and the batch size is set to 8. We empirically set $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1, \lambda_4 = 1$ in the total loss function, $\alpha = 1$ in Equation 2, $M = 256$ in Equation 1, and $N = 256$ in Section 3.2.

4.2 Evaluation on layout reconstruction

In this experiment, we evaluate both the qualitative and quantitative performance of our model on the layout reconstruction task by comparing with baselines. The qualitative results are shown in Figure 4. In short, our method produces more accurate layout reconstruction than LED^2-Net and LGT-Net. Note that visual comparison with PSMNet is infeasible due to the lack of testing code and dataset. Our method also achieves the best performance against all the baselines across all evaluation metrics, as shown in Table 1. In the setting without ground-truth (GT) camera poses, our model shows an improvement over the PSMNet by 4.49% for 2D IoU and 0.02 for δ^i without needing a (noisy) pose prior. Please refer to the supplementary material for more visual comparisons.

Table 2: **Quantitative comparisons with state-of-the-arts methods on indoor panorama registration.** We evaluate all the methods on the ZInD dataset with our testing set.

	SfM successful pair		full testing pair			
Method	$\Delta t \downarrow$	$\Delta R \downarrow$	R-mAA@5° ↑	R-mAA@10° ↑	T-mAA@5° ↑	T-mAA@10° ↑
OpenMVG (Moulon et al., 2016)	17.6717	7.3707	0.1787	0.1969	0.1369	0.1627
GPR-Net (Ours)	**6.9907**	**2.5514**	**0.8633**	**0.9378**	**0.7667**	**0.8854**

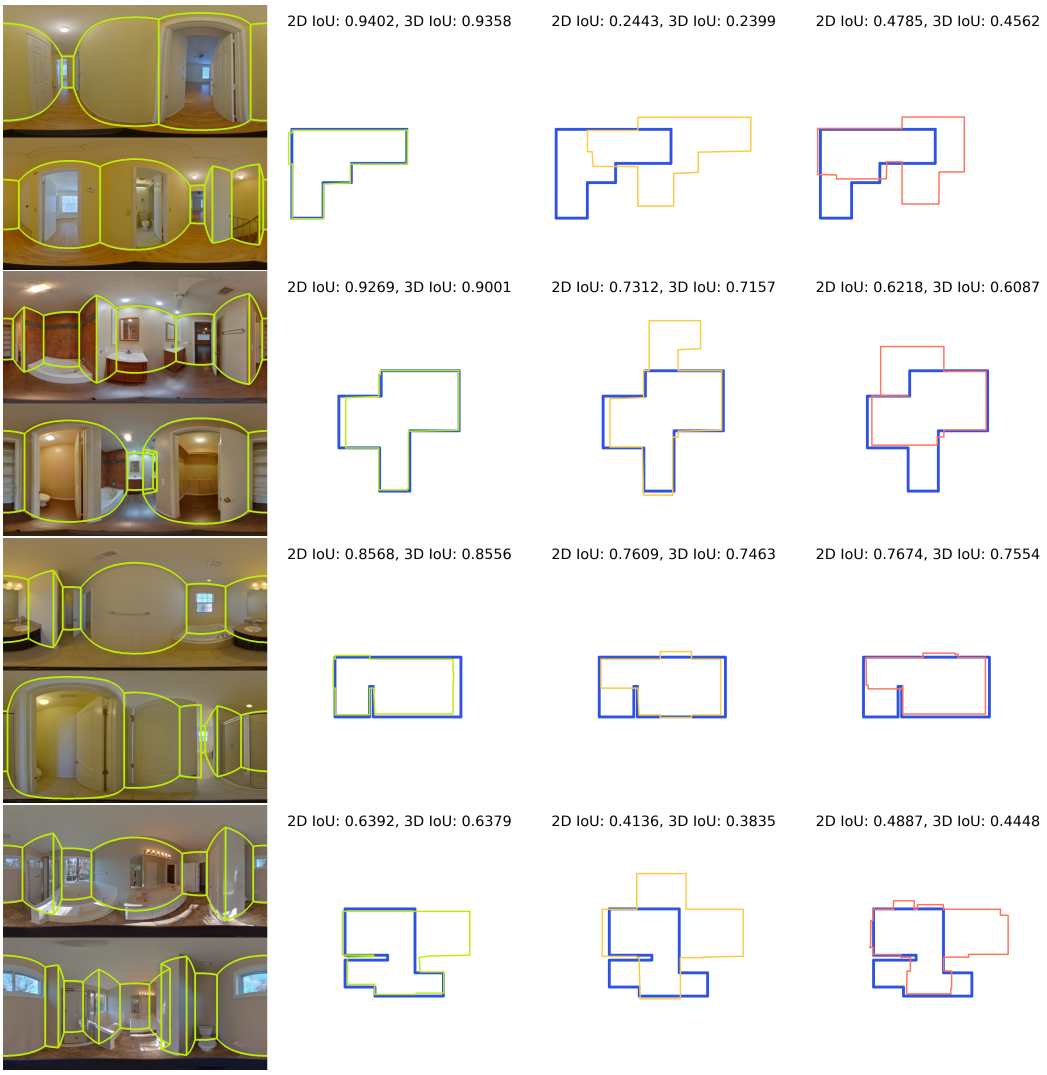

Figure 4: **Visual comparisons.** We show visual comparisons with other competing methods categorized by difficulty. From top to bottom, we select cases where our reconstruction accuracy is the range of top 10%, 20%, 50%, and bottom 10% in our test set. The first column shows two input panoramas with their estimated layouts. The remaining columns show the ground-truth layout, our layouts, LED^2-Net's layouts, and LGT-Net's layouts in blue, green, yellow, red, respectively.

4.3 Evaluation on indoor panorama registration

In this experiment, we compare GPR-Net with OpenMVG in terms of image registration quality on the indoor panoramas. As shown in Table 2, GPR-Net outperforms OpenMVG across all evaluation metrics. The main problem of OpenMVG is that it has many failure cases where the algorithm returns no registration. Therefore, we show a separated comparison on the subset of panorama pairs where OpenMVG successfully returns a result. Note that even in these cases, our model beats OpenMVG with a noticeable improvement in angular errors. Moreover, our model overwhelms OpenMVG when evaluating how many pairs are successfully registered within a certain error threshold.

Table 3: **Ray cast number vs. registration accuracy.**

	SfM successful pair		full testing pair			
Number of ray	$\Delta t \downarrow$	$\Delta R \downarrow$	R-mAA@5° ↑	R-mAA@10° ↑	T-mAA@5° ↑	T-mAA@10° ↑
256	**6.9907**	**2.5514**	**0.8633**	**0.9378**	**0.7667**	**0.8854**
512	16.9696	8.3937	0.6390	0.7183	0.5171	0.6416
1024	16.5873	7.2243	0.6825	0.7314	0.5901	0.6701

Table 4: **Ray cast number vs. layout reconstruction accuracy.**

	w/ GT pose			w/o GT pose		
Number of ray	2D IoU↑	δ^i ↑	3D IoU↑	2D IoU↑	δ^i ↑	3D IoU↑
256	**0.8449**	**0.9603**	**0.8211**	**0.8026**	**0.9452**	**0.7816**
512	0.7853	0.9472	0.7633	0.6917	0.8618	0.6730
1024	0.7619	0.9405	0.7407	0.6569	0.8437	0.6392

Table 5: **Joint optimize architecture vs. layout reconstruction accuracy.**

	w/ GT pose			w/o GT pose		
Method	2D IoU↑	δ^i ↑	3D IoU↑	2D IoU↑	δ^i ↑	3D IoU↑
w/o joint optimization	0.8364	0.9557	0.8131	0.7920	0.9413	0.7713
w/ joint optimization	**0.8449**	**0.9603**	**0.8211**	**0.8026**	**0.9452**	**0.7816**

4.4 Ablation Study

We conducted ablation studies to validate our method from different perspectives.

The influence of the number of cast rays. In this experiment, we start with the default setting of 256 cast rays and progressively add more rays during the inference. As shown in Table 4 and Table 3, we obtain the best performance in the default setting on both the layout reconstruction and panorama registration.

The effect of joint optimization architecture. In this experiment, we divided the layout prediction and the correspondence prediction into two individual models. Specifically, we use LED^2-Net as our layout prediction model and the vision transformer as our correspondence prediction model. As shown in Table 5, we obtain better accuracy using our joint prediction model on all the layout reconstruction metrics.

5 Conclusions

We present a first complete solution for room layout reconstruction from a pair of panorama images. In contrast to previous work, i.e. PSMNet, we do not rely on an approximate registration but can register the two panorama images directly. The major improvement over PSMNet comes from a novel Geometry-aware Panorama Registration Network (GPR-Net) that effectively tackles the wide baseline registration problem. We propose to exploit the layout geometry and compute fine-grained correspondences between the two layout boundaries, rather than directly computing the registration on global pixel-space. The main limitation of our method is that the layout fusion block that processes two layouts is very simple. We recommend the development of learned fusion modules as major avenue for future work.

References

Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*,

October 2019.

Angel Chang, Angela Dai, Thomas 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 (3DV)*, 2017.

Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2133–2143, June 2021.

Ricardo Fabbri, Timothy Duff, Hongyi Fan, Margaret H. Regan, David da Costa de Pinho, Elias Tsigaridas, Charles W. Wampler, Jonathan D. Hauenstein, Peter J. Giblin, Benjamin Kimia, Anton Leykin, and Tomas Pajdla. Trplp - trifocal relative pose from lines at points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. ISBN 0521540518.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1849–1856, Sept 2009. doi: 10.1109/ICCV.2009.5459411.

Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, Oct 2007. ISSN 1573-1405. doi: 10.1007/s11263-006-0031-y. URL https://doi.org/10.1007/s11263-006-0031-y.

Shunping Ji, Zijie Qin, Jie Shan, and Meng Lu. Panoramic slam from a multiple fisheye camera rig. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:169–183, 2020. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2019.11.014. URL https://www.sciencedirect.com/science/article/pii/S0924271619302758.

Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6207–6217, October 2021.

Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1654–1663, June 2022.

Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. *CoRR*, abs/1703.06241, 2017. URL http://arxiv.org/abs/1703.06241.

Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pp. 60–74. Springer, 2016.

A. Pagani and D. Stricker. Structure from motion using full spherical panoramic cameras. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 375–382, 2011. doi: 10.1109/ICCVW.2011.6130266.

G. Pintore, F. Ganovelli, R. Pintus, R. Scopigno, and E. Gobbetti. 3d floor plan recovery from overlapping spherical images. *Computational Visual Media*, 4:367–383, 2018.

Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. Automatic modeling of cluttered multi-room floor plans from panoramic images. *Computer Graphics Forum*, 38(7):347–358, 2019. doi: https://doi.org/10.1111/cgf.13842. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13842.

Giovanni Pintore, Marco Agus, and Enrico Gobbetti. AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan world assumption. In *Proc. ECCV*, August 2020. URL http://vic.crs4.it/vic/cgi-bin/bib-page.cgi?id='Pintore:2020:AI3'.

Srikumar Ramalingam and Matthew Brand. Lifting 3d manhattan lines from a single image. *2013 IEEE International Conference on Computer Vision*, pp. 497–504, 2013.

Y. Salaün, R. Marlet, and P. Monasse. Line-based robust sfm with little image overlap. In *2017 International Conference on 3D Vision (3DV)*, pp. 195–204, 2017. doi: 10.1109/3DV.2017.00031.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. 2016.

Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. URL https://aminshabani.github.io/publications/extreme_sfm/pdfs/iccv2021_2088.pdf.

Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. 2019.

Aparna Taneja, Luca Ballan, and Marc Pollefeys. Registration of spherical panoramic images with cadastral 3d models. 3DIMPVT '12, pp. 479–486, USA, 2012. IEEE Computer Society. ISBN 9780769548739. doi: 10.1109/3DIMPVT.2012.45. URL https://doi.org/10.1109/3DIMPVT.2012.45.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12956–12965, June 2021.

Haiyan Wang, Will Hutchcroft, Yuguang Li, Zhiqiang Wan, Ivaylo Boyadzhiev, Yingli Tian, and Sing Bing Kang. Psmnet: Position-aware stereo merging network for room layout estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8616–8625, June 2022.

Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, and F. Xu. 3d room layout estimation from a single rgb image. *IEEE Transactions on Multimedia*, 22(11):3014–3024, 2020. doi: 10.1109/TMM.2020.2967645.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pp. 668–686, 2014. doi: 10.1007/978-3-319-10599-4_43. URL https://doi.org/10.1007/978-3-319-10599-4_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 *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020.

Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. 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.

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. Feb 2021. ISSN 1573-1405. doi: 10.1007/s11263-020-01426-8. URL https://doi.org/10.1007/s11263-020-01426-8.

Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26:305–364, 2017. doi: 10.1017/S096249291700006X.