PSMNet: Position-aware Stereo Merging Network for Room Layout Estimation

Haiyan Wang[1,2†] Will Hutchcroft[1*] Yuguang Li[1*] Zhiqiang Wan[1] Ivaylo Boyadzhiev[1]
Yingli Tian[2] Sing Bing Kang[1]
[1]Zillow Group [2]The City College of New York
hwang005@citymail.cuny.edu, ytian@ccny.cuny.edu
{willhu,yuguangl,zhiqiangw,ivaylob,singbingk}@zillowgroup.com

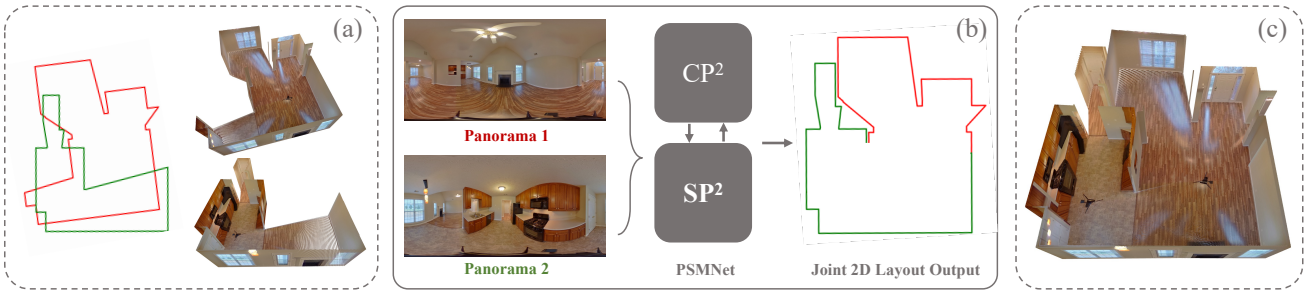

Figure 1: Estimating the complete layout of complex indoor spaces from a pair of 360° panoramas. We use GT data for the sake of demonstration. Due to occlusion, a single panorama may view only a portion of the whole space. (a) shows the 2D and 3D room layout components, representing only portion of the whole space that is visible to each panorama. In practice, the input relative pose may be only approximately known; this is represented by the noisy alignment between the two partially visible components. Our proposed end-to-end PSMNet shown in (b) takes the two panoramas as input and jointly estimates the complete visible room layout in 2D, while refining a given noisy relative pose. (c) visualizes the estimated layout in 3D. (a) and (c) are used for visualization. The input and output of PSMNet are shown in (b).

Abstract

In this paper, we propose a new deep learning-based method for estimating room layout given a pair of 360° panoramas. Our system, called Position-aware Stereo Merging Network or PSMNet, is an end-to-end joint layout-pose estimator. PSMNet consists of a Stereo Pano Pose (SP^2) transformer and a novel Cross-Perspective Projection (CP^2) layer. The stereo-view SP^2 transformer is used to implicitly infer correspondences between views, and can handle noisy poses. The pose-aware CP^2 layer is designed to render features from the adjacent view to the anchor (reference) view, in order to perform view fusion and estimate the visible layout. Our experiments and analysis validate our method, which significantly outperforms the state-of-the-art layout estimators, especially for large and complex room spaces.

1. Introduction

Image-based room layout estimation is an important step to constructing models of home interiors for a variety of applications, such as virtual tours, path planning, floor plan generation, and home insights on square footage and architectural style. Much work has been done in room layout estimation, and current techniques perform well on simple Manhattan and Atlanta-world layouts. However, their performance degrade for large and complex rooms, e.g., those that have more than 10 corners.

It is not unusual (at least in North America) to find room layouts that are significantly more complex than cuboids or L-shapes. Examples include large open spaces with merged kitchen, dining room, and living room. The prevalence of complex rooms is evidenced by the statistics of real residential homes in ZInD [4]. Figure 1 illustrates the difficulty of layout estimation for a complex indoor space with many self-occlusions. Here, single image solutions would not be adequate due to occlusion. This is because no panorama is able to see the entire open space. Using both panoramas would, in principle, be able to better extract the layout. In addition, given that reliability is distance dependent (due to reductions in resolution at farther distances), such dependence is reduced with multiple views.

In this paper, we recover the room layout from two 360° panoramas. This has its challenges, because relative camera pose of the panorama pair needs to be estimated jointly with layout. While techniques such as structure-from-motion

†Work done while Haiyan Wang was an intern at Zillow.
*Authors contributed equally.

exist, our goal is to generate layouts of complex rooms that may lack features due to occlusions. Wide baseline 2-view Structure from Motion (SfM) is still an open problem. In this work we assume that an input pose, potentially noisy, is provided. For example, this could be based on a rough user input [4] or matching corresponding semantic elements with noisy predictions [29].

Our solution is a joint pose-layout deep architecture to predict 2D room layout and refine a noisy 3 DOF relative camera pose in an end-to-end manner. Our system, called Position-aware Stereo Merging Network (PSMNet), consists of a transformer-based Stereo Pose Estimation (SP^2) network and a new pose-aware Cross-Perspective Projection (CP^2) module. CP^2 generates the final layout with the help of an attention-based merging model (inspired by SEBlock [12]) that weights regions based on certainty. The pose and layout modules share the same encoder for efficiency, and are trained end-to-end.

In our work, we make the same assumptions as ZInD [4]: input panoramas are both captured upright at approximate same height, and layouts are based on Atlanta world (horizontal floor and ceiling, and vertical walls). The ceiling height is used for visualization.

The contributions of our work are:

(i) First end-to-end joint layout-pose deep architecture (to our knowledge) for large and complex room layout estimation from a pair of panoramas.

(ii) New Cross-Perspective Projection (CP^2) module with attention-based merging for layout generation.

(iii) An integrated transformer-based relative Stereo Pano Pose (SP^2) network to refine noisy input pose.

(iv) State-of-the-art performance on a challenging, stereo panoramas dataset, sampled from ZInD [4].

2. Related Work

In this section, we review approaches relevant to our work. They are organized based on the following attributes for room layout estimation: (1) partial versus complete layouts, (2) single-view, and (3) multi-view 360° panoramas. More extensive surveys can be found in [19] and [23].

2.1. Partial vs Complete Rooms Layouts

Much work has been done on generating partial room layout from a single perspective image [2, 5, 11, 15, 16, 28, 41, 42]. Early, geometric-based approaches analyze lines and vanishing points [8, 11]. With the introduction of large-scale datasets [4, 11, 39, 40], most of the recent work is learning-based [5, 15, 41].

Extending [8] to multiple perspective views, [9] proposed a hybrid approach where low-level cues (extracted from structure-from-motion) are combined into a learnable Bayesian framework to build a multi-view consistent partial room layout. Extending further the input requirements, by using a small number of overlapping perspective RGB-D images, [17] proposed a geometric-based approach to fuse multiple, partial pieces into a complete room layout.

2.2. Single-view 360° Layout Estimation

PanoContext [38] was one of the first to study the effect of FoV for room layout estimation. Similar to other early work [36], they first convert a single panorama into overlapping perspective images to estimate per-pixel normals, by combining [16] and [11], which is later used to evaluate room layout hypothesis. LayoutNet [42] demonstrated the benefits of operating directly on the equirectangular panorama. They use an encoder-decoder CNN, similar to RoomNet [15], to estimate the corner and boundary probabilities for cuboid layout estimation. DuLa-Net [37] jointly exploited the equirectangular panorama and its perspective ceiling-view in an end-to-end differentiable network, using a novel equirectangular-to-perspective (E2P) feature fusion step.

HorizonNet [31] is a seminal approach that generates a compact 1D representation where each image column of the equirectangular panorama encodes the floor-wall, ceiling-wall, and wall-wall boundaries. A bidirectional RNN is used to learn short and long-term dependencies across the panorama. Many subsequent approaches [25, 32, 34, 35] adopted HorizonNet as their back-bone architecture. AtlantaNet [20] proposed floor and ceiling-view projections to combine the benefits of DuLa-Net and HorizonNet. They handle the more complex Atlanta-world cases [27]. Despite the increased 360° FoV, monocular layout estimation techniques are less effective for large open spaces or complex rooms with self-occlusion.

2.3. Multi-view 360° Layout Estimation

Most techniques on multi-view 360° layout estimation are focused on full floor-plan reconstruction from a sparse set of overlapping RGB panoramas [1, 22] or a dense sequence of RGB-D scans [3, 6, 7, 18]. The pure RGB approaches [1, 22] typically start with SfM [13] to determine relative panorama poses. However, this step tends to fail for sparsely captured panoramas with wide baselines, as demonstrated by [4, 29]. Assuming all images can be localized, a common approach is to first segment each input image into floor, wall and ceilings regions (akin to [36]), in combination with multi-view cues and constraints [1, 22, 24]. The multi-view segmentation maps are then projected, as 2D layouts, and fused together into a final floor-plan boundary [1], or per-room layout boundaries [21, 24].

Recently, [30] proposed a multi-view layout reconstruction for large indoor spaces, starting from multiple panorama images with known poses. Using ideas from DuLa-Net and HorizonNet, they first use [31] to obtain single-view layout predictions, which are then converted into ceiling-view

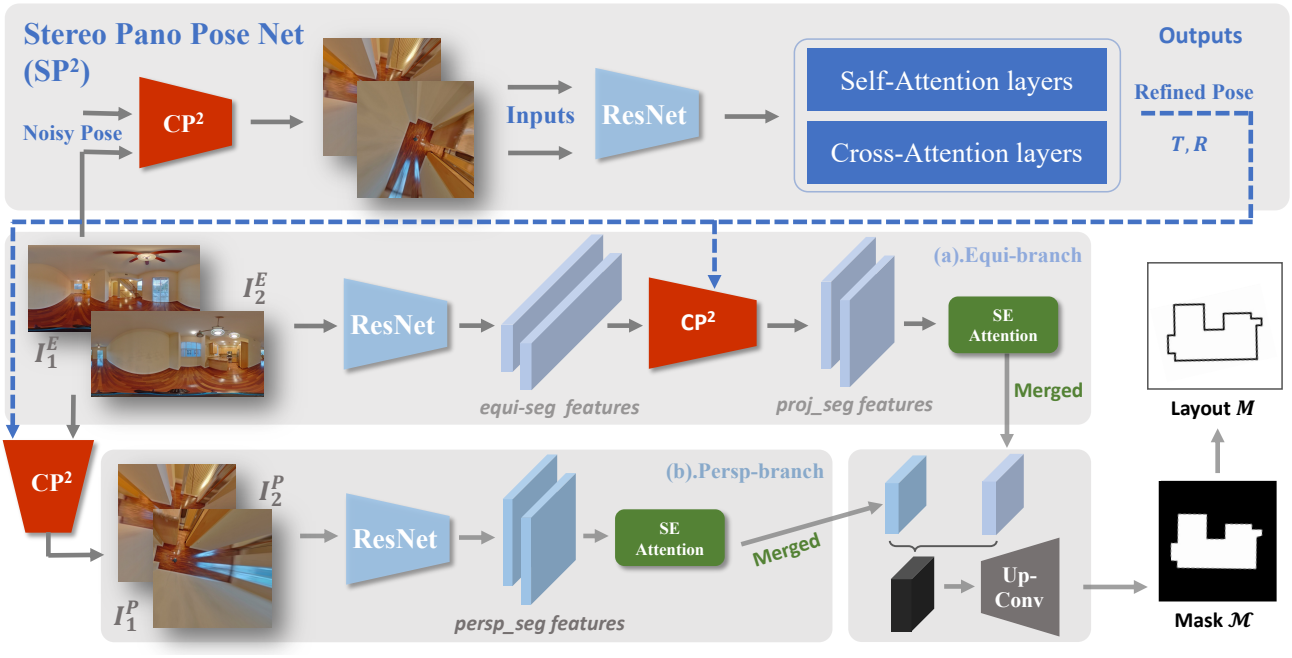

Figure 2: Our proposed PSMNet architecture. Its input is a pair of equirectangular panoramas with noisy relative pose, from which perspective projections are generated by the Cross-Perspective Projection layer (CP^2) as additional inputs to generate the room layout. Stereo Pano Pose (SP^2) Net is trained to refine the relative pose. The output of PSMNet is mask $\mathcal{M}$, which is then post-processed to generate the layout polygon M.

segmentation masks. Their key idea is to train a DNN to generate multiple ceiling-wall (boxification) line proposals from each view. Those are then fused using a graph-cut optimization to obtain a single multi-view consistent 2D layout. Their method can handle more than 2 views. However, they highly rely on the quality of the pre-computed single-view layouts as well as the given camera poses.

In contrast, we focus on 2-view layout reconstruction, with approximate poses. We propose an end-to-end fully differentiable DNN to jointly estimate multi-view consistent layout while refining pose.

3. Framework

In this section, we describe PSMNet (Figure 2) and the loss function we optimize. PSMNet includes the Cross-Perspective Projection layer (CP^2) and Stereo Pano Pose Net (SP^2), both of which are described in subsequent sections.

3.1. Architecture Design

PSMNet adopts a backbone similar to DuLa-Net [37] or AtlantaNet [20] with a dual-segmentation structure. The inputs to PSMNet are two 360° panoramas I_1^E, I_2^E with camera viewpoints V_1, V_2, respectively. Without loss of generality, let the first be the anchor view, with the other (secondary view) the target for pose estimation and feature fusion. Each image $I_n^E (n = 1, 2)$ is processed in equirectangular space and perspectively projected to the anchor view. The latter operation, which we call Cross-Perspective Projection (CP^2) layer , uses the relative pose estimated by the Stereo Pano Pose (SP^2) Network given the two panoramas.

Given a potentially noisy input pose, SP^{2}Net uses a transformer-based attention mechanism to refine the relative positions between stereo-view panoramas. We also extract two sets of segmentation features, namely *equi-seg features* and *persp-seg features*, with each set being the result of concatenating two views. The *equi-seg features* are further rendered to the anchor view in the perspective space in the same way as was done for the panoramas, resulting in *proj-seg features*. Note that *persp-seg features* and *proj-seg features* are camera aligned and thus can readily be merged.

Prior to segmentation feature merging, we apply an attention model inspired by [12] to extract an implicit confidence representation. As the concatenated segmentation features are generated from two separate camera views, each view's feature vector encodes content from differing regions of the room. As a result, the contributions of these features on the final merged feature are expected to be non-uniform (e.g., due to depth and texture variation). An SE-Attention model is used to estimate these contribution weights. The refined *persp-seg features* and *proj-seg features* are concatenated

and fed to a set of Up-Conv layers. The output of the last Up-Conv layer is binary mask $\mathcal{M}$ of the 2D room layout.

The final polygon layout is generated by our proposed Mostly Manhattan algorithm as follows. We first extract a dense contour from $\mathcal{M}$, which is then fit with line segments using the Douglas-Peucker algorithm [26]. While the majority of published work imposed a strong Manhattan constraint on the post-processed layout, we allow some walls to be non-Manhattan when a candidate wall is greater than a threshold γ away from one of the coordinate axes; more details are found in the supplementary material. As we do not estimate the ceiling height, the 3D layout can be extracted by extruding the layout with the ground truth ceiling height.

3.2. Loss Function Design

PM^{2}Net is jointly trained end-to-end on the pose and layout estimation tasks. The pose estimation is formulated as a regression problem; we compute the ℓ_1 loss between the ground truth and predicted pose parameters. We denote the rotation loss by $L_{\mathrm{p}}^{(R)}$ and translation loss by $L_{\mathrm{p}}^{(T)}$. The pose loss is given by

$$L_{\mathrm{p}} = \mu L_{\mathrm{p}}^{(R)} + (1-\mu) L_{\mathrm{p}}^{(T)}. \tag{1}$$

Layout estimation is cast as a segmentation process where we compute the cross-entropy losses between the predicted room shape mask and the ground truth in both equirectangular and perspective spaces, denoted by $L_l^{(E)}$ and $L_l^{(P)}$, respectively. The layout loss is

$$L_l = L_l^{(E)} + L_l^{(P)}. \tag{2}$$

The total loss for end-to-end training is

$$L_{\mathrm{tot}} = (1-\lambda) L_{\mathrm{p}} + \lambda L_l. \tag{3}$$

Note that in Eqs. (1) and (3), the hyper-parameters $\mu, \lambda \in [0, 1]$. For our experiments, both μ and λ are set as 0.5.

4. Cross-Perspective Projection (CP^2) Layer

PSMNet augments the equirectangular panoramas $I_n^E (n = 1, 2)$ with aligned perspective-projected top-down views as additional signals. These top-down views are generated by the CP^2 layer, using the first (anchor) viewpoint as reference; let these views be $I_n^P (n = 1, 2)$. The two panoramas are assumed to be axis aligned vertically [38]. We use normalized texture coordinates $(u_n^E, v_n^E, n = 1, 2)$ ranging from 0 to 1 to represent position in I_1^E and I_2^E.

The 3 DOF pose of the secondary relative to the anchor is $\{\Delta x, \Delta y, \Delta\theta\}$, indicating relative 2D position shift and horizontal angular difference. These are refined by SP^{2}Net as described in Section 5.

Let the focal lengths of $I_n^P (n = 1, 2)$ be $f_n^P (n = 1, 2)$. The panorama south pole can be found at the origin of I_1^P.

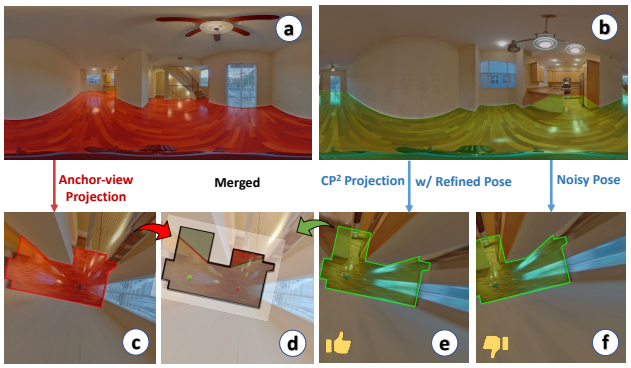

Figure 3: Illustration of our proposed CP^2. We plot visible floor area on the panorama and perspective projected images. (c) is the anchor-view projection from (a). (d) displays the estimated room layout from both views with jointly refined camera pose. (e) and (f) compare the projected adjacent panorama to the anchor view with and without the pose refinement, respectively.

I_2^P is projected from I_2^E to be in the same reference coordinate system as I_1^P. For a fixed field of view FoV_1 in the anchor view, f_1^P, FoV_2 (field of view in the secondary view), and f_2^P can be found as follows:

$$f_1^P = 0.5\lambda \cot(0.5 FoV_1), \tag{4}$$

$$FoV_2 = 2\tan^{-1}\left(\frac{H_1^E}{H_2^E}\tan(0.5 * FoV_1)\right), \tag{5}$$

$$f_2^P = 0.5\lambda \cot(0.5 FoV_2), \tag{6}$$

where $H_n^E (n = 1, 2)$ are the camera heights and λ is the width (in pixels) of $I_n^P (n = 1, 2)$. In our work, we assume $H_1^E = H_2^E$, which results in $FoV_1 = FoV_2$ and $f_1^P = f_2^P$.

Given translation $t_n = \{x_n, y_n\}$ and rotation θ_n, $n = 1, 2$, our CP^2 layer projects the panorama coordinates as follows:

$$u_n^E = \frac{atan2(p_x^n - x_n, p_y^n - y_n) - \theta_n}{2\pi}, \tag{7}$$

$$v_n^E = 1 - \frac{atan2(\left\|p_x^n - x_n, p_y^n - y_n\right\|_2, f_n^P)}{\pi}, \tag{8}$$

where $(p_x^n, p_y^n, n = 1, 2)$ is a point in the *joint floor coordinate* system. For the anchor view, $x_1 = y_1 = \theta_1 = 0$. For the secondary view, $x_2 = \Delta x, y_2 = \Delta y, \theta_2 = \Delta\theta$. The effect of CP^2 is illustrated in the Figure 3.

5. Stereo Panorama Pose (SP^2) Network

We assume that the relative pose between the two input panoramas is only approximately known; in practice, any pose estimate will be subject to noise. The SP^{2}Net component of PSMNet is responsible for refining the initial

pose estimate. More specifically, the goal of SP^{2}Net is to predict the pose refinement parameters $\Delta t = t_{gt} - t_c$ and $\Delta\theta = \theta_{gt} - \theta_c$, where $\Delta t = \{\Delta x, \Delta y\}$, (t_c, θ_c) is the input noisy pose and (t_{gt}, θ_{gt}) the ground truth pose.

First, the anchor panorama I_1^E is projected to perspective view I_1^P by the CP^2 layer, with the pose parameters all set to 0. The second input panorama I_2^E is projected to perspective view I_2^P by the same CP^2 layer using the input noisy pose (t_c, θ_c). The shared backbone ResNet-18 is then used to extract multi-scale features from I_1^P and I_2^P. The extracted features are denoted as F_1^P and F_2^P.

F_1^P and F_2^P are added with positional encodings, and each feature map is then flattened to a 1-D vector. The encoded features are passed through a transformer to extract position and context dependent local features. The transformer (inspired by [33]) consists of self-attention and cross-attention layers. The output features from the transformer are denoted as T_1^P and T_2^P.

Finally, T_1^P and T_2^P are concatenated along the channel dimension. The concatenated features C^P are fed into three convolutional layers to extract features that contain the relative pose information between the two input panoramas. The extracted features are flattened, and a fully connected layer is used to predict Δt and $\Delta\theta$.

6. Experiments

In this section, we report results of our approach on Zillow Indoor Dataset (ZInD) [4]. Given the variability of relative location of panoramas as well as the complexity of the room layout, we use *spatial overlap* and *co-visibility* (which we define shortly) to stratify our results.

Most room layout datasets such as PanoContext [38], Stanford 2D-3D [42], and Realtor360 [37] only feature a single panorama per room, making them not suitable for our work. We chose Zillow Indoor Dataset (ZInD) [4] for evaluation because it is the only large-scale, public dataset that has the multi-view panorama configuration for layout estimation, it is based on many real residential homes across many cities, and its rooms have significant geometric diversity (Manhattan and non-Manhattan, and significant spread in room size and number of room corners). We derive a stereo-view dataset from ZInD for our experiments. In total, there are 107,916, 13,189, and 12,348 pano pair instances in our train, test, and val splits, from 40,336, 5,138, and 4,993 unique panoramas, respectively.

6.1. Benchmarks and Evaluation Metrics

We compare our PSMNet with baselines built upon recent state-of-the-art layout estimation methods: HorizonNet [31], DulaNet [37], LED^{2}Net [35] and HoHoNet [32]. Since these methods only address single view layout estimation, we first derive the estimated result for each view and then perform a simple shape union across stereo views to get the merged result. Note that room layout recovery is based on what is *visible*; an occlusion edge shows up as a "wall".

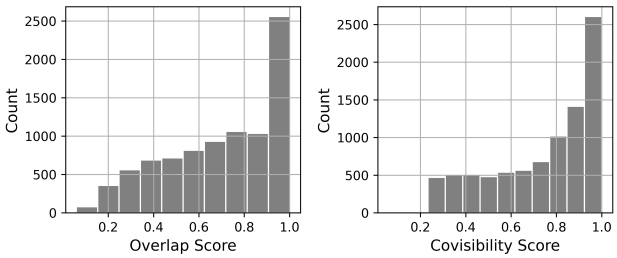

Figure 4: Data distribution of the spatial overlap and co-visibility scores.

In deriving single-view results, we found that using fully-Manhattan post-processing decreases the baseline performance. This is because ZInD contains many partially visible layouts, which introduces non-Manhattan occlusion "walls".

To increase the baseline performance, we instead apply post-processing which preserves non-Manhattan structure. For HorizonNet and HoHoNet, we sample the predicted contour at corner lines of sight to get the final layout polygon. For the segmentation based methods, we apply AtlantaNet's post-processing, which also preserves non-Manhattan walls. For all methods, we apply our Mostly Manhattan post-processing (Section 3.1) to the merged shape union to get the final room layout. The quality of the stereo-view layout estimate is evaluated using *2D IoU*. We also use δ_i [37], which measures accuracy in panorama pixel space.

Spatial overlap and *co-visibility* are employed in our experiments as measures of difficulty and to stratify results. For a panorama pair, *spatial overlap* measures the IoU between the ground truth single-view visible layout polygons. The higher the score, the more visible floor area the two panoramas have in common. As shown in Figure 4, we split the dataset into *Overlap-High* (> 0.9), *Overlap-Medium* (0.5 - 0.9) and *Overlap-Low* (< 0.5). In the test set, each split has 3,769, 5,644, and 3,776 data instances, respectively.

Since our method incorporates both the perspective and equirectangular projections, we additionally stratify by *co-visibility* [4], which measures visual overlap ($\in [0, 1]$) between two panoramas. Examples are split into *Covis-High* (> 0.9), *Covis-Medium* (0.5 - 0.9), and *Covis-Low* (< 0.5).

6.2. Implementation Details

PSMNet is implemented in PyTorch and trained with the Adam [14] optimizer on a single GPU for 200 epochs. We set the learning rate as 0.0001 and the batch size as 6. The backbone feature extraction network is ResNet18 [10]. The Manhattan threshold γ in Section 3.1 is set to 10.

The joint layout-pose network is trained with two configurations (one with ground truth pose and the other with noisy pose augmentation). For pose noise, in the training stage, we perturb the ground truth pose with noise sampled from a

Table 1: Quantitative evaluation stratified by spatial overlap at different levels of room complexity. Note that "Overlap-Low" indicates higher room complexity with more occlusions.

Pose	Methods	Overall		Overlap-High		Overlap-Medium		Overlap-Low	
		2D IoU (%)	δ_i	2D IoU (%)	δ_i	2D IoU (%)	δ_i	2D IoU (%)	δ_i
w/ GT	DulaNet [37]	64.03	0.8043	65.21	0.8185	62.14	0.8031	60.27	0.7980
	HorizonNet [31]	73.35	0.8663	82.08	0.8801	71.47	0.8678	69.20	0.8535
	HoHoNet [32]	74.25	0.8649	82.55	0.8816	72.43	0.8672	70.35	0.8486
	LED^{2}Net [35]	76.39	0.9056	83.68	0.9243	73.73	0.8736	72.08	0.8697
	PSMNet (Ours)	**81.01**	**0.9238**	**85.71**	**0.9349**	**80.13**	**0.9253**	**76.93**	**0.9074**
w/o GT	DulaNet [37]	59.30	0.7828	62.06	0.7699	57.97	0.7855	51.21	0.7936
	HorizonNet [31]	62.79	0.8354	70.98	0.8437	61.51	0.8355	58.24	0.8288
	HoHoNet [32]	63.31	0.8324	70.59	0.8390	62.03	0.8339	59.47	0.8253
	LED^{2}Net [35]	65.81	0.8566	71.06	0.8493	64.81	0.8574	63.14	0.8611
	PSMNet (Ours)	**75.77**	**0.9217**	**84.80**	**0.9371**	**74.73**	**0.9210**	**66.73**	**0.9040**

uniform distribution. We perform further data augmentation by randomly switching which panorama is selected as the anchor view. For all of the single-view baseline models, we re-train the network on single-view examples from the ZInD stereo dataset, with 200 epochs for fair comparison. After getting the single-view layout estimation results, we apply both the same ground truth pose and noisy pose to perform the merging process.

To assess our model's tolerance to noise, we synthetically generate noisy pose estimates by sampling. In practice, any pose estimator may be used, such as LayoutLoc [4]. 2-view wide baseline SfM remains a challenging open problem.

6.3. Quantitative Evaluation

In our evaluation of PSMNet on the ZInD stereo-view dataset, we apply the same ground truth (GT) and noisy poses as inputs for other baseline methods, as well as PSMNet, for an apples-to-apples comparison. Quantitative results reported in Table 1 show that PSMNet consistently outperforms the baseline methods with significant improvements especially for complex rooms.

Performance with GT pose. With known GT pose, the influence of pose refinement removed. As shown in the upper half of Table 1, with GT pose as input, PSMNet shows an overall improvement over the LED^{2}Net baseline of 4.62% for 2D IoU and 0.02 for δ_i. Overlap High demonstrates the most competitive baseline performance. With high visual overlap, the benefit of an additional view is reduced. Nevertheless, we improve upon all baselines. The advantage of PSMNet further increases as visual overlap is reduced (Overlap-Medium and Overlap-Low).

Performance with noisy pose. To assess noise tolerance, we sample pose noise from uniform distributions $\mathbf{U}(0, 40°)$ and $\mathbf{U}(0, 1m)$, for rotation and translation, respectively. As shown in the lower part of Table 1, when the input pose is noisy, our joint layout-pose estimation pipeline surpasses the baseline performance by a large margin. For Overlap-High, even without GT pose, PSMNet performs better than most baselines *with GT pose*. This is compelling illustration of the benefit of SP^2. Additional results for all methods stratified by *co-visibility* are shown in the supplementary. They are similarly positive for PSMNet.

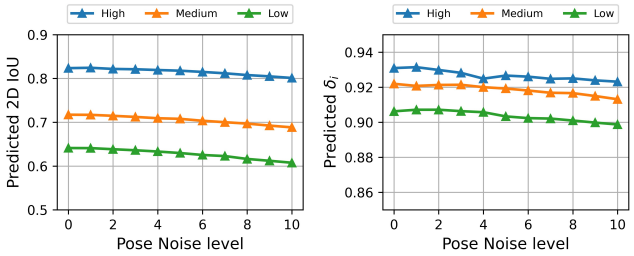

Figure 5: Illustration of the robustness of our proposed PSMNet under various level of pose noise for Overlap-High, Overlap-Medium, and Overlap-Low groups.

Effect of pose noise. Our simulated coarse poses are generated by adding $\pm\theta^{\circ}_{err}$ rotational errors as well as translation errors of fixed magnitudes in a random direction. Figure 5 shows our system performance with pose noises ranging from $0 \rightarrow (0m, 0°)$ to $10 \rightarrow (1m, 40°)$, with increments of $(0.1m, 4°)$. We compute both 2D IoU and δ_i on the stratified ZInD dataset. PSMNet demonstrates robust performance over a range of input pose noises, with higher than 80% IoU on *Overlap-High* and more than 60% IoU on *Overlap-Low*, even when the pose noise increases significantly to $(1m, 40°)$. The δ_i plot shows a similar trend.

6.4. Qualitative Evaluation

In Figure 6, we show sample estimated layouts for PSMNet compared with the LED^{2}Net baseline. In the first two columns, the accuracy of both our layout and pose refinement is demonstrated by the alignment with the true boundaries in

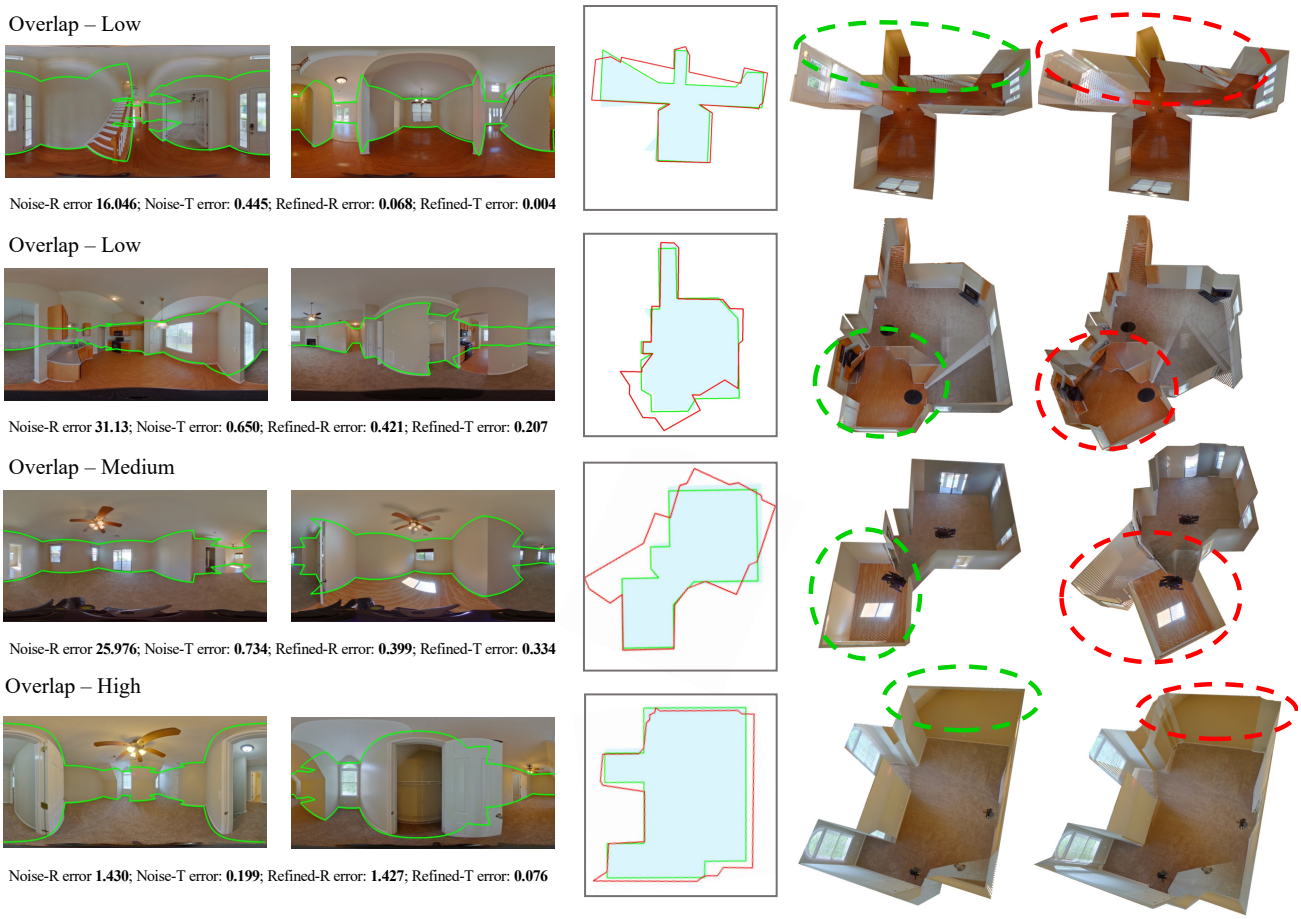

Figure 6: Position-aware layout estimation results on the ZInD dataset with noisy pose. The first two columns show the estimated visible room layout on each panorama in green. In the middle top-down plot we stack the predicted room shape in green over the ground truth mask in cyan, while the LED^{2}Net baseline result is shown in red. The fourth column demonstrates the 3D layouts of our PSMNet and the last column is the results of LED^{2}Net.

the equirectangular images, for both the anchor and adjacent view alike. Note that while they are captured in segmentation, we do not post-process additional "internal" polygons for islands, pillars, or separating walls that arise in large spaces. This can be seen by the absence of partial boundary in row 2. We further compare PSMNet, the LED^{2}Net baseline, and GT, displayed in top-down projection, in column 3. Columns 4 and 5 show the 3D layouts for PSMNet and LED^{2}Net, extruded by GT ceiling height for display. The benefits of SP^2, and by contrast the consequences of pose noise, are striking. We highlight significant differences; the benefit of end-to-end learning of pose and layout is most noticeable when comparing the cohesive PSMNet layouts, with the poorly merged regions of the LED^{2}Net baseline.

6.5. Ablation Study on Network Components

We conduct experiments to investigate the affect of individual components in our PSMNet architecture. Specifically,

Table 2: Evaluation of different variants of PSMNet.

CP^2	SP^2	SE Attention	2D IoU (%)	δ_i
×	×	×	60.41	0.8031
✓	×	×	66.39	0.8701
✓	✓	×	70.92	0.8883
✓	×	✓	69.14	0.8723
✓	✓	✓	**72.38**	**0.9003**

we consider the following variants:

(i) Remove the proposed CP^2, SP^2, and SE Attention layers. Instead of cross-perspective rendering, we do a direct perspective rendering for the second view of input panorama I_2^E.

(ii) Remove the SP^2 and the SE Attention layers, while just using the proposed CP^2 to perform cross-perspective projection based on the coarse pose. There is no further

Table 3: Comparison of different post processing methods.

Pose	Methods	Most Manhattan PP		AtlantaNet PP	
		2D IoU	δ_i	2D IoU	δ_i
w/ GT	DulaNet [37]	64.03	0.8043	62.06	0.7899
	HorizonNet [31]	73.35	0.8663	71.36	0.8785
	HoHoNet [32]	74.25	0.8649	73.25	0.8732
	LED^{2}Net [35]	76.39	0.9056	75.14	0.8849
	PSMNet (Ours)	**81.01**	**0.9238**	**77.69**	**0.9159**
w/o GT	DulaNet [37]	59.30	0.7828	58.90	0.7634
	HorizonNet [31]	62.79	0.8354	60.17	0.8145
	HoHoNet [32]	63.31	0.8324	61.85	0.8204
	LED^{2}Net [35]	65.81	0.8566	64.09	0.8423
	PSMNet (Ours)	**75.77**	**0.9217**	**73.22**	**0.8926**

pose refinement.

(iii) Replace the SE Attention layers with standard convolution layers to process extracted features.

(iv) Remove the pose refinement model SP^2, instead using the coarse pose directly as input to the CP^2 layer.

Variant performance is reported in Table 2 with highlighted the best (bold) and worst (underline) layout estimations. CP^2 is the most critical component of our model which makes use of the refined pose from SP^2 in order to associate the adjacent view to the anchor view. With CP^2 added, we see that both SP^2 and SE Attention bring additional substantial gains to our model, with SP^2 proving to be slightly more effective than SE Attention.

We further examine the effect of our proposed Mostly Manhattan post processing algorithm. As mentioned in Section 6.1, due to ZInD's complexity where most visible layouts go beyond the Manhattan world, fully-Manhattan post processing methods such as in HorizonNet [31], and DulaNet [37] do not work well. By comparison, the post processing method introduced by AtlantaNet [20] is better suited for general visible layouts. Here, we compare the performance of our proposed Mostly Manhattan post-processing to AtlantaNet's post-processing method, when applied to our network as well as the baselines. The results are reported in Table 3. Our Mostly Manhattan method consistently outperforms the AtlantaNet post processing method w/ or w/o GT pose. Also note that even when we apply AtlantaNet post processing to our network's output, we still achieve a better performance compared to the baseline models.

7. Discussion

In Figure 7, we share an example which highlights limitations and challenges of our task. Here, the room is not only complex, but there is also a low overlap score (only 0.27) between the two panoramas. As a result, there are few common features between the two views. The GT and predicted PSMNet layouts are shown in Figure 7 (a) and (b); in (b), boundary misalignment can be seen in panorama 2. We

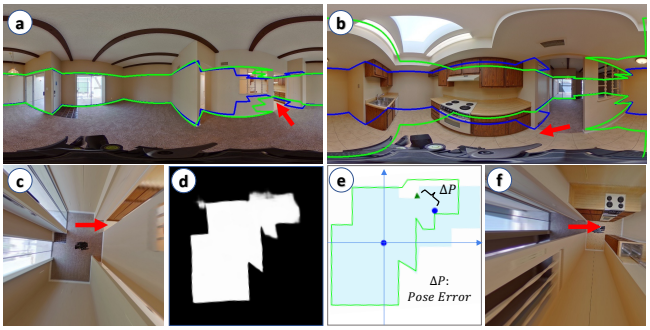

Figure 7: Illustration of a challenging example, with an overlap score of just 0.27.

highlight the narrow path connecting the kitchen and living room (with arrows), which causes low co-visibility. This challenge is further compounded in the perspective views, Figure 7 (c) and (f), with limited common floor and wall texture between the image pair.

Figure 7 (e) displays the error (ΔP) in the refined position of panorama 2. This pose error results in feature misalignment inside our network, which ultimately leads to a noisy predicted segmentation, shown in Figure 7 (d). Figure 7 (e) further visualizes the final predicted layout and GT floor segmentation, where we observe a direct correlation between the relative pose error and shifted layout boundary. This particular example also contains a flaw in the data, where the GT is missing a portion of floor boundary around the divider between kitchen and living room (which our model recognizes). This means that in actuality the co-visibility and overlap scores are even lower than computed.

8. Conclusion

We have introduced a novel end-to-end approach for jointly estimating complex room layout from stereo-view panoramas, while refining a noisy relative pose. We adopt a dual-projection backbone architecture to extract features from both equirectangular and perspective-view images. For pose refinement, we propose a transformer-based Stereo Pano Pose (SP^2) Network to derive implicit correspondence, and predict refinement parameters by a fully-connected layer. A novel Cross-Perspective Projection (CP^2) is crucially designed to project the adjacent panorama view to the anchor view, as well as to align multi-scale equirectangular features for merging in the central segmentation branch. To weight the contribution of features from both views, we apply SE-Attention inspired by [12]. To evaluate the performance of our method, we introduce baselines based upon currently available state-of-the-art single-perspective layout estimators. Our model demonstrates significant improvements in layout estimation accuracy on a new stereo-view visible layout dataset, derived from ZInD, which will be released to the community.

References

[1] Ricardo Silveira Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 628–635, 2014. 2

[2] Yu-Wei Chao, Wongun Choi, Caroline Pantofaru, and Silvio Savarese. Layout estimation of highly cluttered indoor scenes using geometric and semantic cues. In *ICIAP*, 2013. 2

[3] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2661–2670, 2019. 2

[4] 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*, pages 2133–2143, 2021. 1, 2, 5, 6

[5] Erick Delage, Honglak Lee, and A. Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:2418–2428, 2006. 2

[6] Hao Fang, Florent Lafarge, Cihui Pan, and Hui Huang. Floorplan generation from 3d point clouds: A space partitioning approach. *Isprs Journal of Photogrammetry and Remote Sensing*, 175:44–55, 2021. 2

[7] Hao Fang, Cihui Pan, and Hui Huang. Structure-aware indoor scene reconstruction via two levels of abstraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021. 2

[8] Alex Flint, Christopher Mei, David William Murray, and Ian D. Reid. A dynamic programming approach to reconstructing building interiors. In *ECCV*, 2010. 2

[9] Alex Flint, David William Murray, and Ian D. Reid. Manhattan scene understanding using monocular, stereo, and 3d features. *2011 International Conference on Computer Vision*, pages 2228–2235, 2011. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[11] Varsha Hedau, Derek Hoiem, and David Alexander Forsyth. Recovering the spatial layout of cluttered rooms. *2009 IEEE 12th International Conference on Computer Vision*, pages 1849–1856, 2009. 2

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2, 3, 8

[13] Florian Kangni and Robert Laganière. Orientation and pose recovery from spherical panoramas. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[14] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 5

[15] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4875–4884, 2017. 2

[16] David C. Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009. 2

[17] C. Lin, Changjian Li, and Wenping Wang. Floorplan-jigsaw: Jointly estimating scene layout and aligning partial scans. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5673–5682, 2019. 2

[18] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floornet: A unified framework for floorplan reconstruction from 3d scans. In *ECCV*, 2018. 2

[19] Bincy P Mathew. Review on room layout estimation from a single image. *International Journal of Engineering Research and*, 9, 2020. 2

[20] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. Atlantanet: Inferring the 3d indoor layout from a single 360 image beyond the manhattan world assumption. In *European Conference on Computer Vision*, pages 432–448. Springer, 2020. 2, 3, 8

[21] Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. Automatic modeling of cluttered multi-room floor plans from panoramic images. *Computer Graphics Forum (proceeding of Pacific Graphics 2019)*, 2019. 2

[22] Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and E. Gobbetti. 3d floor plan recovery from overlapping spherical images. *Computational Visual Media*, 4:367–383, 2018. 2

[23] Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. State-of-the-art in automatic 3d reconstruction of structured indoor environments. *Computer Graphics Forum*, 39(2):667–699, 2020. 2

[24] Giovanni Pintore, Ruggero Pintus, Fabio Ganovelli, Roberto Scopigno, and E. Gobbetti. Recovering 3d existing-conditions of indoor structures from spherical images. *Comput. Graph.*, 77:16–29, 2018. 2

[25] Shivansh Rao, Vikas Kumar, Daniel Kifer, C. Lee Giles, and Ankur Mali. Omnilayout: Room layout reconstruction from indoor spherical panoramas. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3701–3710, 2021. 2

[26] Alan Saalfeld. Topologically consistent line simplification with the douglas-peucker algorithm. *Cartography and Geographic Information Science*, 26(1):7–18, 1999. 4

[27] Grant Schindler and Frank Dellaert. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *CVPR 2004*, 2004. 2

[28] Alexander G. Schwing, Tamir Hazan, Marc Pollefeys, and Raquel Urtasun. Efficient structured prediction for 3d indoor scene understanding. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2815–2822, 2012. 2

[29] 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)*, pages 5703–5711, October 2021. 2

[30] Si Sio-Keong, Su Jheng-Wei, Peng Chi-Han, Chen Kuo-Wei, Chang Felix, Yao Chih-Yuan, and Chu Hung-Kuo. Recon-

structing 3d indoor layout from multiple panoramic images. In *CGW*, 2021. 2

[31] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 2, 5, 6, 8

[32] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 2, 5, 6, 8

[33] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 5

[34] Phi Vu Tran. Sslayout360: Semi-supervised indoor layout estimation from 360° panorama. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15348–15357, 2021. 2

[35] 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*, pages 12956–12965, 2021. 2, 5, 6, 8

[36] Hao Yang and Hui Zhang. Efficient 3d room shape recovery from a single panorama. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5422–5430, 2016. 2

[37] 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*, pages 3363–3372, 2019. 2, 3, 5, 6, 8

[38] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV*, 2014. 2, 4, 5

[39] Yinda Zhang, Fisher Yu, Shuran Song, Pingmei Xu, Ari Seff, and Jianxiong Xiao. Large-scale scene understanding challenge: Room layout estimation. 2

[40] Jia Zheng, Junfei Zhang, J. Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, 2020. 2

[41] Wenzhao Zheng, Jiwen Lu, and Jie Zhou. Structural deep metric learning for room layout estimation. In *ECCV*, 2020. 2

[42] 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*, pages 2051–2059, 2018. 2, 5