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

Supplementary Material

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

I. Introduction

In this supplementary document, we provide additional details of our PSMNet architecture, discuss our Mostly Manhattan post-processing algorithm in greater detail, and provide additional experimental results (quantitative and qualitative).

II. The PSMNet Architecture

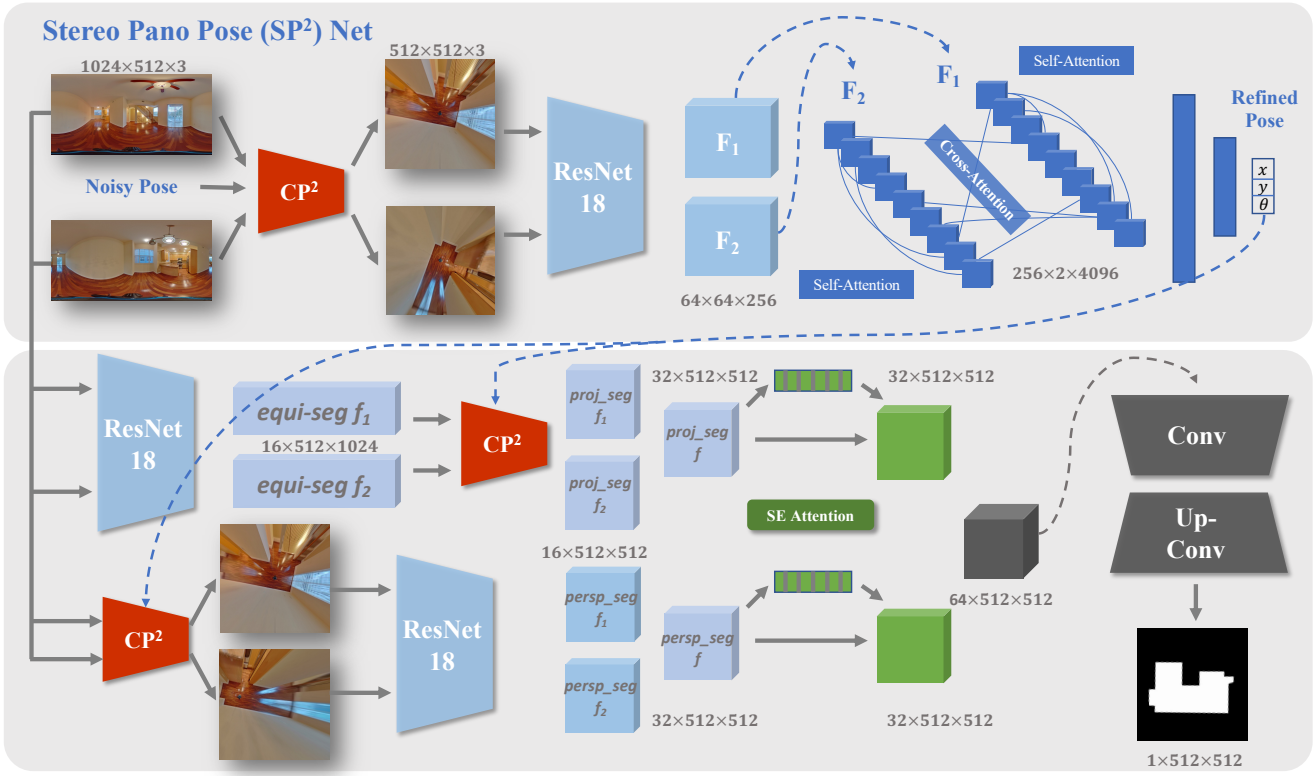

Figure I. Our proposed PSMNet architecture.

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

Figure I shows a more detailed end-to-end structure of PSMNet. The network consists of two sub-networks, the Stereo Pano Pose (SP^2) Network and the layout estimation network. SP^2 Network takes two panoramas of size $1024 \times 512 \times 3$ as input. Given an input noisy pose, the CP^2 layer projects the two panoramas into the perspective space, with size $512 \times 512 \times 3$. They are fed into a ResNet18 backbone to extract the semantic feature vectors F_1 and F_2 which have size $64 \times 64 \times 256$. In conjunction with the positional encoding, the extracted features F_1 and F_2 are further processed by the self-attention and cross-attention layers, generating the position-aware features of size $256 \times 2 \times 4096$. Several convolution and fully connected layers are adopted to estimate the refined relative pose x, y, θ between two panoramas. The refined pose is then used to provide accurate projection for the layout estimation network. Two *equi-seg* features ($16 \times 512 \times 1024$) are directly extracted from the input panoramas and further projected to the *proj-seg* features ($32 \times 512 \times 512$) with the CP^2 layer. Further, two *persp-seg* features ($16 \times 512 \times 1024$) are extracted from the perspective projected images produced by the CP^2 layer, with the help of the refined pose. Next, the *proj-seg* and *persp-seg* features are merged separately by the SE Attention layer and concatenated together to produce a feature of size $64 \times 512 \times 512$. Several convolution and up-convolution layers are employed to process the fused features and generate the final segmentation mask.

III. Mostly Manhattan Post Processing Algorithm

To apply our post-processing, we first perform contour extraction on the predicted segmentation using OpenCV, and simplify the largest contour returned using the Douglas–Peucker algorithm [2]. We then estimate the vanishing angle from the equirectangular image of the anchor panorama. This is used to rotate the extracted contour so that most (Manhattan) walls will align with one of the coordinate axes. We subsequently remove any similar points that are within a distance threshold of one another. We then loop over wall segments and delete intermediate vertices if any two wall segments are within 30° of co-linearity. We then form wall line equations, snapping any wall to align with a Manhattan coordinate axis if within a threshold of 20°. We generate the final polygon by intersecting neighboring wall lines to obtain the set of vertices.

IV. More on Experimentation

IV-A. Quantitative Evaluation on Co-visibility Score

Table I. Quantitative evaluation stratified by co-visibility at different levels of room complexity. Note that "Covis-Low" indicates higher room complexity with more occlusions.

Pose	Methods	Overall		Covis-High		Covis-Medium		Covis-Low	
		2D IoU (%)	δ_i	2D IoU (%)	δ_i	2D IoU (%)	δ_i	2D IoU (%)	δ_i
w/ GT	DulaNet [6]	64.03	0.8043	65.89	0.8099	61.03	0.8012	62.73	0.8031
	HorizonNet [3]	73.35	0.8663	76.42	0.8791	71.28	0.8576	73.29	0.8661
	HoHoNet [4]	74.25	0.8649	76.98	0.8795	72.34	0.8560	74.34	0.8629
	LED^2Net [5]	76.39	0.9056	78.52	0.9167	73.45	0.8710	76.18	0.8680
	PSMNet (Ours)	**81.01**	**0.9238**	**83.69**	**0.9364**	**79.38**	**0.9037**	**79.20**	**0.9094**
w/o GT	DulaNet [6]	59.30	0.7828	60.68	0.7829	59.04	0.7729	60.48	0.7923
	HorizonNet [3]	62.79	0.8354	66.18	0.8392	60.93	0.8273	61.98	0.8451
	HoHoNet [4]	63.31	0.8324	65.90	0.8348	61.60	0.8237	63.20	0.8450
	LED^2Net [5]	65.81	0.8566	67.48	0.8500	64.45	0.8481	66.20	0.8796
	PSMNet (Ours)	**75.77**	**0.9217**	**81.16**	**0.9336**	**73.47**	**0.9015**	**69.33**	**0.9013**

As mentioned in the Section 6.3 of the paper, we also demonstrate the layout estimation results on the ZInD [1] stereo-view dataset stratified by the *co-visibility* score. *Co-visibility* score is defined in [1]. It measures the visual overlap between two viewpoints in the panoramic space. The score ranges from 0 to 1; the lower of the score, the greater the challenge for our task. Specifically, the dataset is split into three portions: *Covis-High* (> 0.9), *Covis-Medium* (0.5 - 0.9), and *Covis-Low* (< 0.5). The quantitative evaluation is reported in Table I. Overall the results show a trend similar to the *spatial overlap* results reported in Table 1 of the main paper.

For layout performance with GT pose as input, we find that the baseline methods achieve higher performance on *Covis-Low* than they do on the *Covis-Medium* split, i.e., view overlap does not correlate with baseline performance. We hypothesize that this is because the baseline methods do not need to reason jointly about the two views; they need only predict separately from each perspective and merge in post. As a result, the combined layout performance largely depends on single-view

Figure II. Qualitative evaluation of the SP^2 model. (a),(b): a pair of (Overlap-Medium) panoramas (**ground truth**, **prediction**). (c): two original single view GT layouts (**anchor view**, **adjacent view**). (d): the effect of pose refinement, starting with a noisy pose (**black**); in (**blue**) is our refined pose obtained from SP^2, visualized on top of the adjacent view layout. The refined pose (**blue**) is much closer to the anchor view layout (**red**), enabling 2-view layout estimation to be learned jointly with our CP^2 branch.

estimate quality, which is not a function of overlap between views. A similar situation happens when applying the noisy pose, which is shown in the bottom part of Table I. For the baseline methods, noisy pose presents the most significant challenge as they have no means of refinement. With no visual overlap-dependent pose refinement or joint layout estimation, the co-visibility score again is not necessarily correlated with performance. On the other hand, for PSMNet, which does jointly reason on both views, the performance is more directly affected by visual (and spatial) overlap.

IV-B. Evaluation of SP^2

The initial pose is necessary due to the limited overlap and complex occlusion between image pairs (especially for the Overlap Low & Medium splits). Two-view SfM, with large to extreme base-line, is a non-trivial open research problem. In practice, the initial pose is easy to obtain, e.g. in production pipelines (see [1]); *manual* pair-wise calibration is a much simpler task, where two (manually selected) corresponding points (e.g., on the floor) are enough to calibrate an upright pair of 360° cameras, but the recovered pose may be noisy. Thus, we propose SP^2 to refine the initial noisy pose and, in cooperation with CP^2, estimate the complex room layout. There are no other deep learning methods that estimate relative pose between two panoramas (without depth-based supervision). The benefit of SP^2 is the use of CP^2 layer to estimate fine-grained pose from noisy pose via the projected two perspective images. Quantitative and qualitative evaluation of SP^2 are reported in Table II and Figure II, which show the efficacy of SP^2. The refinement ability of SP^2 increases as spatial overlap increases (from Low to High).

Table II. Quantitative evaluation of the SP^2 model: a/b shows the initial and refined pose respectively (R and T stand for rotation and translation).

Error	Overall	High	Medium	Low
Mean R (°)	19.88 / 1.15	19.74 / 0.81	20.01 / 1.23	19.85 / 1.48
Mean T (m)	0.42 / 0.09	0.44 / 0.05	0.4 / 0.08	0.4 / 0.16

IV-C. More Visual Results

Here, we present additional visual comparison between our proposed PSMNet in green and the LED^{2}Net [5] baseline in red. The ground truth is shown in blue. As in the main paper, we choose *spatial overlap* to stratify the dataset. Figure III shows the layout estimation results which are merged by adopting the GT pose. The results demonstrate that PSMNet's ability to reason jointly about both views is superior to direct merging of separately predicted partial views, even when the pose is known exactly.

Figure IV demonstrates the results when applying the noisy pose to perform the merging. This more clearly shows the superiority of our joint layout-pose estimation network, as we see the lack of pose refinement capability in the LED^{2}Net baseline further degrades layout estimation quality.

References

[1] 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. 2, 3

[2] Alan Saalfeld. Topologically consistent line simplification with the douglas-peucker algorithm. *Cartography and Geographic Information Science*, 26(1):7–18, 1999. 2

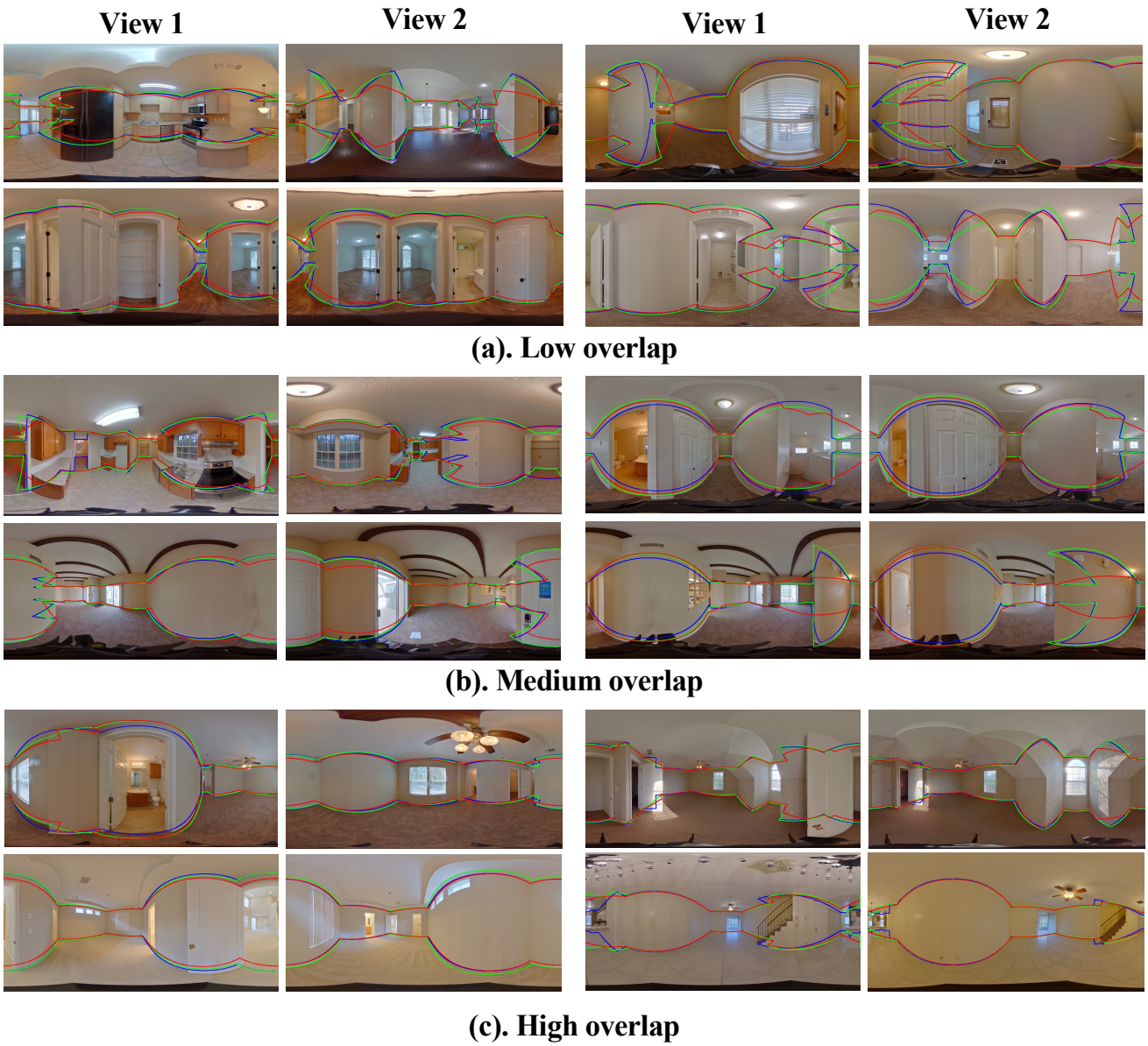

Figure III. Position-aware layout estimation results on the ZInD dataset with gt pose.

[3] 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

[4] 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] 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, 3

[6] 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

Figure IV. Position-aware layout estimation results on the ZInD dataset with noisy pose.