Disentangling Orthogonal Planes for Indoor Panoramic Room Layout Estimation with Cross-Scale Distortion Awareness Supplementary Material

Zhijie Shen, Zishuo Zheng, Chunyu Lin†, Lang Nie, Kang Liao, Shuai Zheng, and Yao Zhao
Institute of Information Science, Beijing Jiaotong University, China
Beijing Key Laboratory of Advanced Information Science and Network Technology
{zhjshen, zszheng, cylin, nielang, kang_liao, zs1997, yzhao}@bjtu.edu.cn

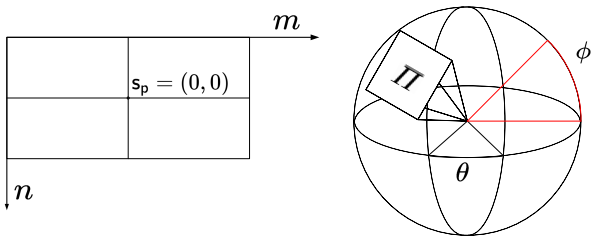

Figure 1. Illustration of the three coordinate systems.

1. Distortion-Aware Projection

Definition. Following [2, 5], we denote a unit sphere as S with its surface as S^2. For each point s (where $\mathsf{s} = (\phi, \theta) \in S^2$ and $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\theta \in [-\pi, \pi]$) on the spherical surface, it is uniquely defined. Define the virtue tangent plane as $\mathit{\Pi}$ and it is located at $\mathsf{s}_{\mathit{\Pi}} = (\phi_{\mathit{\Pi}}, \theta_{\mathit{\Pi}})$. Denote the coordinates of a point on $\mathit{\Pi}$ as $\mathsf{x} \in \mathbb{R}^2$, $\mathit{\Pi}_0$ represents that the virtue tangent plane located as $\mathsf{s} = (0, 0)$. Define the sampling locations as $\mathsf{s}_{(j,k)}$, where $j, k \in \{-1, 0, 1\}$. Denote the equirectangular image as I with the points on it as $\mathsf{p} = (m, n)$. We set the coordinates of the geometric center of the image to $\mathsf{s}_{\mathsf{p}} = (0, 0)$ (Fig. 1 left).

Projection Between Spherical Surface And The Virtue Tangent Planes. We utilize the step sizes of the equirectangular image at the equator (denoted as $\Delta_\theta, \Delta_\phi$) to sample the points. $\forall \quad \mathsf{s} = (\phi, \theta) \in S^2$, the coordinates of the nine points to sample can be written as:

$$\begin{aligned} \mathsf{s}_{(0,0)} &= (\phi, \theta) \\ \mathsf{s}_{(\pm1,0)} &= (\phi \pm \Delta_\phi, \theta) \\ \mathsf{s}_{(0,\pm1)} &= (\phi, \theta \pm \Delta\theta) \\ \mathsf{s}_{(\pm1,\pm1)} &= (\phi \pm \Delta_\phi, \theta \pm \Delta\theta) \end{aligned} \tag{1}$$

And the coordinates $\mathsf{x}_{(j,k)}$ of the nine points to sample on the virtue tangent plane located at s can be calculated via the gnomonic projection:

$$\begin{aligned} \mathsf{x}_{(0,0)} &= (\phi, \theta) \\ \mathsf{x}_{(\pm1,0)} &= (\phi \pm \tan\Delta_\theta, \theta) \\ \mathsf{x}_{(0,\pm1)} &= (\phi, \theta \pm \tan\Delta_\phi) \\ \mathsf{x}_{(\pm1,\pm1)} &= (\phi \pm \tan\Delta_\theta, \theta \pm \sec\Delta_\theta \tan\Delta_\phi) \end{aligned} \tag{2}$$

The projection between the tangent plane $\mathsf{x}_{(j,k)}$ and the points s on the spherical surface S^2 can be calculated via the inverse gnomonic projection:

$$\begin{cases} \phi(0,0) = \sin^{-1}(\cos v \sin\phi + \frac{y \sin v \cos\phi}{\rho}) \\ \theta(0,0) = \theta + \tan^{-1}(\frac{x \sin v}{\rho \cos\phi \cos v - y \sin\phi \sin v}) \end{cases} \tag{3}$$

where $x = \phi$, $y = \theta$, $\rho = \sqrt{x^2 + y^2}$ and $v = \tan^{-1}\rho$.

$$\begin{cases} \phi(\pm1,0) = \sin^{-1}(\cos v \sin(\phi \pm \tan\Delta_\theta) + \frac{y \sin v \cos(\phi \pm \tan\Delta_\theta)}{\rho}) \\ \theta(\pm1,0) = \theta + \tan^{-1}(\frac{x \sin v}{\rho \cos(\phi \pm \tan\Delta_\theta) \cos v - y \sin(\phi \pm \tan\Delta_\theta) \sin v}) \end{cases} \tag{4}$$

where $x = \phi \pm \tan\Delta_\theta$, $y = \theta$, $\rho = \sqrt{x^2 + y^2}$ and $v = \tan^{-1}\rho$.

$$\begin{cases} \phi(0,\pm1) = \sin^{-1}(\cos v \sin\phi + \frac{y \sin v \cos\phi}{\rho}) \\ \theta(0,\pm1) = \theta \pm \tan\Delta_\phi + \tan^{-1}(\frac{x \sin v}{\rho \cos\phi \cos v - y \sin\phi \sin v}) \end{cases} \tag{5}$$

where $x = \phi$, $y = \theta \pm \tan\Delta_\phi$, $\rho = \sqrt{x^2 + y^2}$ and $v = \tan^{-1}\rho$.

$$\begin{cases} \phi(\pm1,0) = \sin^{-1}(\cos v \sin(\phi \pm \tan\Delta_\theta) \\ \quad + \frac{y \sin v \cos(\phi \pm \tan\Delta_\theta)}{\rho}) \\ \theta(\pm1,0) = \theta \pm \sec\Delta_\theta \tan\Delta_\phi \\ \quad + \tan^{-1}(\frac{x \sin v}{\rho \cos(\phi \pm \tan\Delta_\theta) \cos v - y \sin(\phi \pm \tan\Delta_\theta) \sin v}) \end{cases} \tag{6}$$

where $x = \phi \pm \tan\Delta_\theta$, $y = \theta \pm \sec\Delta_\theta \tan\Delta_\phi$, $\rho = \sqrt{x^2 + y^2}$ and $v = \tan^{-1}\rho$.

Projection Between Spherical Surface And The Equirectangular Image. $\forall \quad \mathsf{p} = (m, n)$, its related spherical point $\mathsf{s} = (\phi, \theta)$ can be calculated as:

$$\theta = 2\pi \frac{(m - \frac{W}{2})}{W} \tag{7}$$

$$\phi = \pi(\frac{n - \frac{H}{2}}{H}) \tag{8}$$

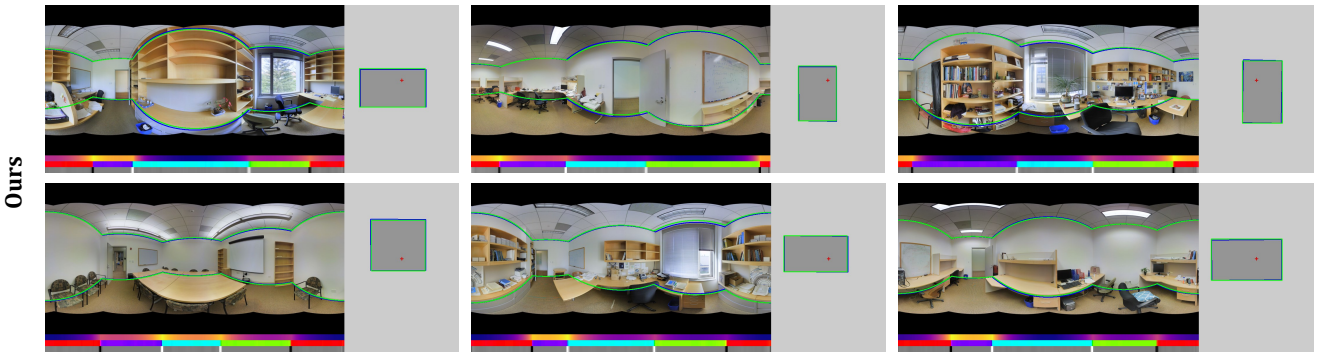

Figure 2. Qualitative results on Stanford 2D-3D [1].

Figure 3. Qualitative results on PanoContext [8].

where H/W represents the height/width of the equirectangular image. Hence, we can calculate the coordinates of the nine most relevant points on the equirectangular images. And the computed coordinates are those prepared points locations mentioned in the submission.

2. Channel-Wise Graph Attention

The dependencies among different channels are generally unconfined. It inevitably leads to information redundancy among channels. In the ideal case, we prefer that each channel can capture distinct aspects of information. To this end, we propose a discriminative channels generation mechanism via graph convolution to enforce each channel to concentrate on different parts of information. Specifically, for channel feature matrix $X \in \mathbb{R}^{C \times W_{\tilde{s}}}$ with C channels, let each channel $X_i \in \mathbb{R}^{1 \times W_{\tilde{s}}}$ be represented as a node in the channel-wise graph, the formula of the discriminative channels generation mechanism can be represented as:

$$X' = \mathbf{L}XW = (\mathbf{I} - \mathbf{A})XW \tag{9}$$

where $\mathbf{L}$ and $\mathbf{A}$ are the symmetric normalized Laplacian matrix and normalized adjacency matrix of the channel-wise graph, and I and W are the identity matrix and learnable weights, respectively. Instead of the common form of graph convolution $\mathbf{A}XW$ used for image-wise or object-wise graphs in previous works, it can be seen from Eq.(9) that a node will subtract the information from the neighbor nodes rather than aggregation. It encourages each channel to retain as much information as possible that is different from its neighbors.

As an important part of the generation of the discriminative channels, channel-wise graph construction is strictly determined by whether Eq.(9) can play its proper role. In practice, a weighted channel-aware measure $\mathrm{Sim}(\cdot)$ is used to construct the channel-wise graph $\tilde{A}_{ij} = \mathrm{Sim}(X_i, X_j)$, which is defined as:

$$\mathrm{Sim}(X_i, X_j) = (X_i W_s) C_a (X_j W_s)^\top \tag{10}$$

where W_s is a projection matrix, $C_a = \mathrm{diag}[c_a]$ is a weighted diagonal matrix, c_a as the weight vector is derived from X. There are many different options to get c_a. Without loss of generality, we adopt the popular hybrid pooling module to obtain C_a conveniently. Specifically, the sequence $X^\top \in \mathbb{R}^{W_{\tilde{s}} \times C}$ is first fed into one average pooling and one max pooling, respectively. Then, the addition operation is used to fuse the outputs of two branches as the final weight vector $c_a \in \mathbb{R}^{W_{\tilde{s}} \times 1}$. Therefore, the normalized channel-wise graph construction can be formulated as follows:

$$\mathbf{A} = \mathbf{D}^{-\frac{1}{2}} (XW_s) C_a (XW_s)^\top \mathbf{D}^{-\frac{1}{2}} \tag{11}$$

where $\mathbf{D} = \mathrm{diag}[d_1, d_1, \cdots, d_n]$ is the degree matrix with $d_i = \sum_{j=0}^{N} \tilde{A}_{ij}$. Finally, the formula of the discriminative

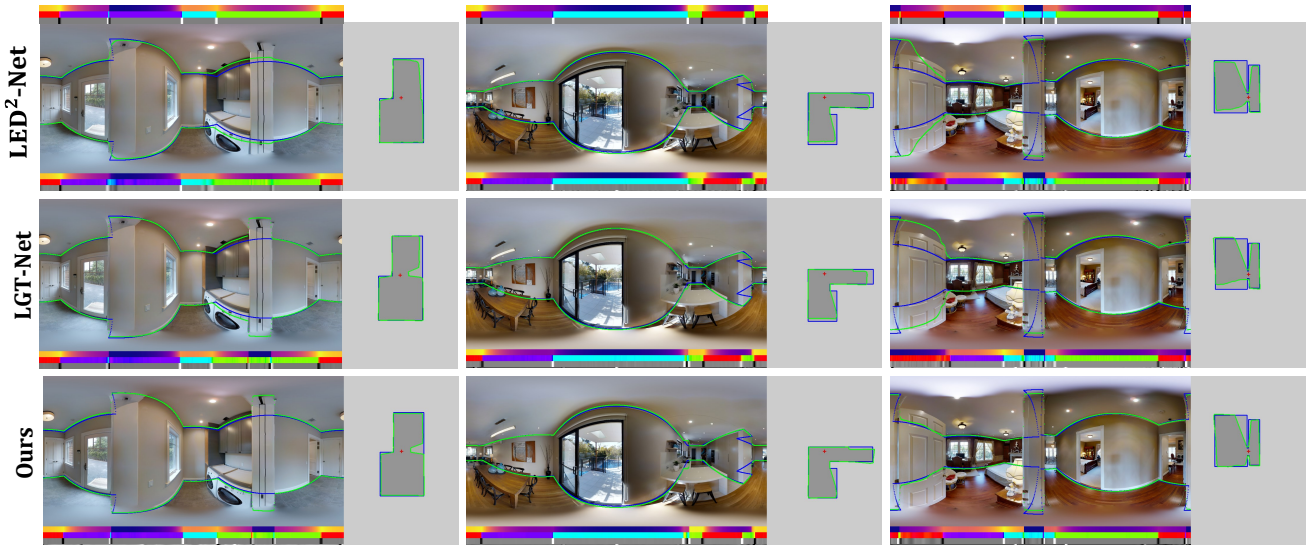

Figure 4. Qualitative comparison with LGT-Net [4] and LED^2-Net [7] on MatterportLayout [9].

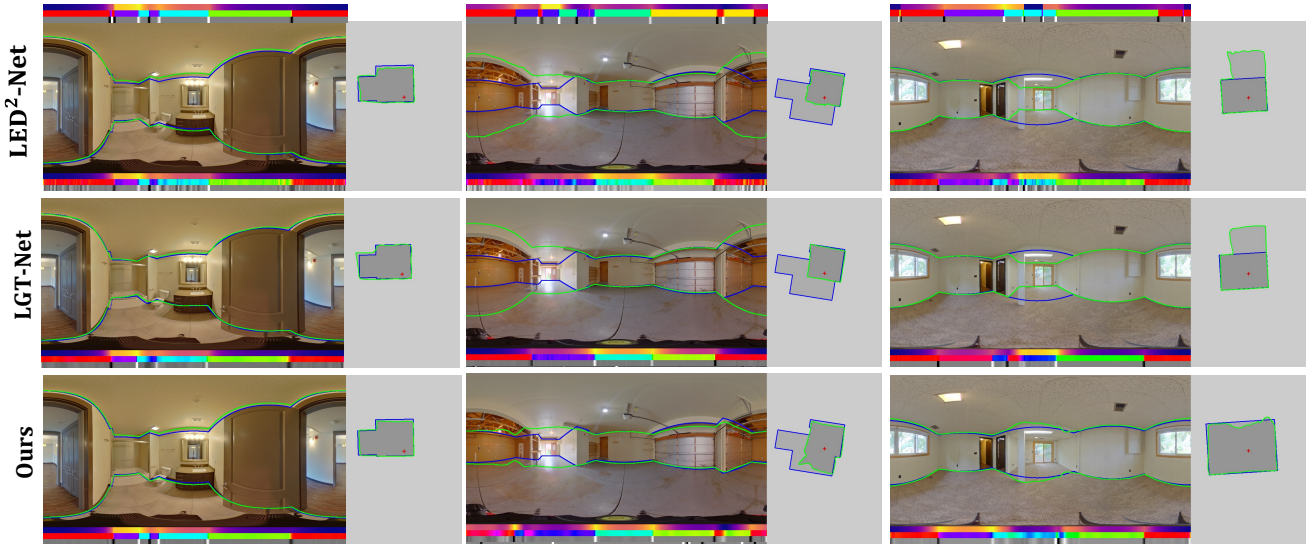

Figure 5. Qualitative comparison with LGT-Net [4] and LED^2-Net [7] on ZInD [3].

channels generation mechanism can be rewritten as:

$$feat = (\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}(XW_s)C_a(XW_s)^\top\mathbf{D}^{-\frac{1}{2}})XW \quad (12)$$

3. More Qualitative Results

We provide more qualitative results in this section. Specifically, we exhibit our qualitative results on the two cuboid datasets (while previous work [4] do not provide them) in Fig. 2 and Fig. 3, respectively. Besides, we provide more qualitative results on the two general room layout datasets (shown in Fig. 4 and Fig. 5). Similar to the submission, the boundaries of the room layout on a panorama are shown on the left, and the floor plan is on the right. Ground truth is best viewed in Blue lines, and the prediction in Green. The predicted horizon depth, normal, and gradient are visualized below each panorama, and the ground truth is in the first row. We do not employ the post-processing strategy for all the listed methods.

4. More features visualization

The features visualization results (on the right) and our prediction results are shown (on the left) in Fig. 6. The boundaries of the room layout on a panorama, as well as the floor plane are exhibited in the figure. Ground truth is best viewed in Blue lines, and the prediction in Green. The predicted horizon depth, normal, and gradient are visualized below each panorama. From the figure, we can observe that the features without disentangling orthogonal planes show ambiguous attention due to the confusing semantics. In contrast, our disentangled vertical plane features are more dis-

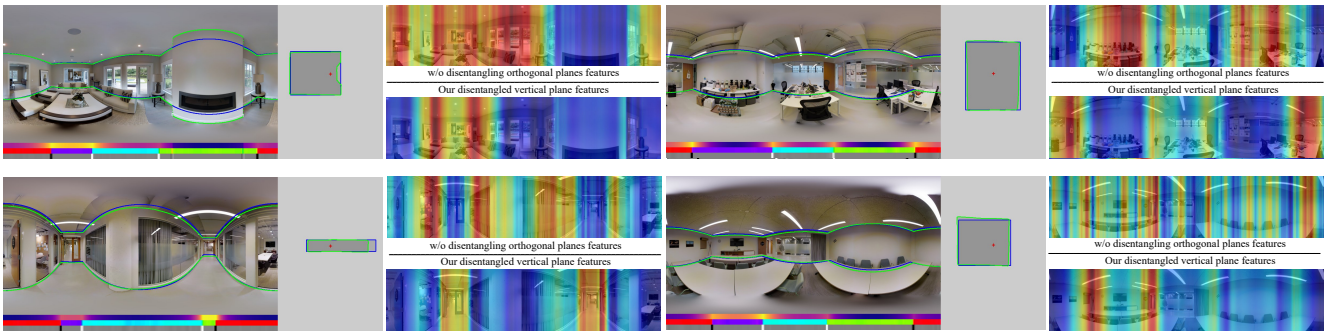

Figure 6. More features visualization. To make the features best viewed, we resize these 1D feature sequences to a 2D representation via a vertical bilinear interpolation operation.

Method	GMac	Param.(M)	Time(ms)	Fps
HorizonNet[6]	71.8	81.56	51	20
LED^2-Net[7]	71.8	81.56	56	18
LGT-Net[4]	81.5	119.34	31	32
Ours	81.8	137.05	41	24

Table 1. Complexity comparison.

criminative and give more attention to the layout corners. The procedure of disentangling orthogonal planes frees our 1D sequence of the vertical plane from the negative effect of the indoor furniture (or illumination), yielding more effective attention to those layout-relevant locations rather than those regions with rich texture.

5. Complexity

We exhibit the complexity comparison results of the methods in Tab. 1. From the table, we can observe that even though our method disentangles the compressed sequences into two 1D representations, the parameters just increased by 17M. In addition, compared with the solution [4] that we strictly followed, the calculation complexity of our model has hardly increased. Our inference time is slightly higher than the LGT-Net's [4] but still outperforms LED^2-Net [7]. In conclusion, our approach does not do serious harm to the computational complexity while yielding better performance.

6. Overview

The contributions are summarized as follows:

- We propose to disentangle orthogonal planes to capture an explicit geometric cue for indoor 360° room layout estimation, with a soft-flipping fusion strategy to assist this procedure.
- We design a cross-scale distortion-aware assembling mechanism to perceive distortion distribution as well as integrate shallow geometric structures and deep semantic features.
- On popular benchmarks, our solution outperforms other SoTA schemes, especially on the metric of intersection over the union of 3D room layouts.

We propose to disentangle orthogonal planes to capture geometric cues in 3D space. Specially, we introduce a vertical flip-fusion strategy to leverage the symmetry property of indoor room layout. Besides, our experimental results demonstrate that dealing with distortion, as well as integrating shallow and deep features, can enhance performance. The code will be publicly available upon acceptance. We hope our work can contribute to this field.

References

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

[2] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*, 2018. 1

[3] 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 *CVPR*, 2021. 3

[4] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *CVPR*, 2022. 3, 4

[5] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 depth estimation. In *ECCV*, 2022. 1

[6] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *CVPR*, 2019. 4

[7] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led 2-net: Monocular 360° layout estimation via differentiable depth rendering. In *CVPR*, 2021. 3, 4

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

[9] 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. *IJCV*, 2021. 3