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

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

Abstract

Based on the Manhattan World assumption, most existing indoor layout estimation schemes focus on recovering layouts from vertically compressed 1D sequences. However, the compression procedure confuses the semantics of different planes, yielding inferior performance with ambiguous interpretability.

To address this issue, we propose to disentangle this 1D representation by pre-segmenting orthogonal (vertical and horizontal) planes from a complex scene, explicitly capturing the geometric cues for indoor layout estimation. Considering the symmetry between the floor boundary and ceiling boundary, we also design a soft-flipping fusion strategy to assist the pre-segmentation. Besides, we present a feature assembling mechanism to effectively integrate shallow and deep features with distortion distribution awareness. To compensate for the potential errors in pre-segmentation, we further leverage triple attention to reconstruct the disentangled sequences for better performance. Experiments on four popular benchmarks demonstrate our superiority over existing SoTA solutions, especially on the 3DIoU metric. The code is available at https://github.com/zhijieshen-bjtu/DOPNet.

1. Introduction

Indoor panoramic layout estimation refers to reconstructing 3D room layouts from omnidirectional images. Since the panoramic vision system captures the whole-room contextual information, we can estimate the complete room layout with a single panoramic image. However, inferring the 3D information from a 2D image is an ill-posed problem. Besides, the 360° field-of-view (FoV) of panoramas brings severe distortions that increase along the latitude. Both issues are challenging for indoor layout estimation.
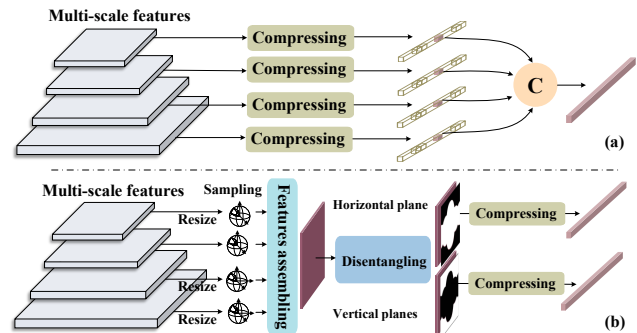
---

†Corresponding author

Figure 1. (a) The commonly used architecture. (b) The proposed one. Compared with the traditional pipeline, ours has two advantages: (1) Disentangling the 1D representation into two separate sequences with different plane semantics. (2) Adaptively integrating shallow and deep features with distortion awareness via a feature assembling mechanism rather than simple concatenation.

Different from outdoor scenarios, the indoor room has the following properties: (1) The indoor scenes conform to the Manhattan World assumption (The floors and ceilings are all flat planes, and the walls are perpendicular to them). (2) The room layout is always described as the corners or the floor boundary and ceiling boundary. These characteristics could be used as potential priors to guide the design of a reasonable layout estimation method.

Based on the Manhattan World assumption, previous approaches [19, 20, 8, 14, 21] prefer to estimate the layout from a 1D sequence. They advocate compressing the extracted 2D feature maps in height dimension to obtain a 1D sequence, of which every element in this sequence share the same distortion magnitude (Fig. 1a). We argue that this compressed representation does not eliminate the panoramic distortions because there is no explicit distortion processing when extracting 2D feature maps before compression. Moreover, this strategy roughly mixes the vertical and horizontal planes together, confusing the semantics of different planes that are crucial for layout estimation.

On the other hand, some researchers devoted themselves

to adopting different projection formats to boost the performance, e.g., the bird's view of the room [29] and cubemap projection [27]. These projection-based schemes weaken the negative effect of the distortions. Nevertheless, frequent projection operations between different formats raise computational complexity. In addition, there exists an inevitable domain gap between the feature maps from different formats.

To address the above problems, we propose a novel architecture (Fig.1b) that disentangles the orthogonal planes in advance to capture an explicit geometric cue. Following [21], our room layout can be recovered from the predicted horizon-depth map and the room height. Therefore, the "clean" depth-relevant features and height-relevant features can both help with the layout estimation. To obtain such "clean" features free from the disturbance of decorations and furniture, we disentangle the widely used 1D representation into two separate sequences — the horizontal and vertical ones. Specifically, we pre-segment the vertical plane (i.e., walls) and horizontal planes (i.e., floors and ceilings) from the whole-room contextual information. Then these orthogonal planes are compressed into two 1D representations. Especially, based on the symmetry property between the floor boundary and ceiling boundary, we design a soft-flipping fusion strategy to assist this process.

Moreover, we propose an assembling mechanism to fuse multi-scale features with distortion awareness effectively. To eliminate the negative effect of distortion, we compute the attention among distortion-relevant positions following distortion distribution patterns. Meanwhile, cross-scale interaction is carried out to integrate shallow geometric structures and deep semantic features. Considering the error of pre-segmentation, we further leverage triple attention to reconstruct the two 1D sequences. Particularly, we adopt graph-based attention to generate discriminative channels, self-attention to rebuild long-range dependencies, and cross-attention to provide the missing information for different sequences.

To demonstrate our effectiveness, we conduct extensive experiments on four popular datasets — MatterportLayout [29], Zind [6], Stanford 2D-3D [1], and PanoContext [26]. The qualitative and quantitative results show that the proposed solution outperforms other SoTA methods. Our contributions can be 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.

2. Related Work

2.1. 360° Room Layout Estimation

Based on Manhattan World assumption [5] that all walls and floors are aligned with global coordinate system axes and are perpendicular to each other, many approaches utilize convolutional neural networks (CNNs) to extract useful features to improve layout estimation accuracy. Specifically, Zou et al. [28] propose LayoutNet to predict the corner/boundary probability maps directly from the panoramas and then optimize the layout parameter to generate the final predicted results. Furthermore, they annotate the Stanford 2D-3D dataset [1] with awesome layouts for training and evaluation. Yang et al. [24] propose Dula-Net to predict a 2D floor plane semantic mask from both the equirectangular view and the perspective view of the ceilings. The modified version, LayoutNet v2, and Du la-Net v2, which have better performance than the original version, are presented by Zou et al. [29]. Fernandez et al. [7] present to utilize equirectangular convolutions (EquiConvs) to generate corner/edge probability maps. Sun et al. propose HorizonNet [19] and HoHoNet [20] to simplify the layout estimation processes by representing the room layout with a 1D representation. Besides, they use Bi-LSTM and multi-head self-attention to build long-range dependencies and refine the 1D sequences.

Recently, several approaches have adopted this 1D representation and achieved impressive performance. For example, Rao et al. [15] establish their network based on HorizonNet [19]. They replace standard convolution operation with spherical convolution to reduce distortions and take Bi-GRU to reduce computational complexity. Wang et al. [21] combine the geometric cues across the layout and propose LED2-Net that reformulates the room layout estimation as predicting the depth of walls in the horizontal direction. Not constrained to Manhattan scenes, Pintore et al. [14] introduce AtlantaNet to predict the room layout by combining two projections of the floor and ceiling planes. Driven by the self-attention mechanism, many transformer-based methods [25, 22, 13] proposed to build long-range dependencies. For example, Jiang et al. [8] represent the room layout by horizon depth and room height and introduce a Transformer to enforce the network to learn the geometric relationships. However, the popular 1D representations could confuse the semantics of different planes, thus causing the layout estimation challenging. We disentangle this 1D representation by pre-segmenting orthogonal planes.

2.2. 360° Image Distortion Mitigation

When converting spherical data to the equirectangular projection format, severe distortions are introduced. Some recent studies focus on designing spherical customized convolutions to make the network adapt to panoramic distortions and offer superior results compared to standard convolutional networks. Su *et al.* [18] introduce SphConv to remove distortions by changing the regular planar convolutions kernel size towards each row of the 360° images. Cohen *et al.* [2] propose a spherical CNN that transforms the domain space from Euclidean S2 space to a SO(3) group to reduce the distortion and encodes full rotation invariance in their network. Coors *et al.* [3] addresses it by defining the convolution kernels on the tangent plane and keeping the convolution domain consistent in each kernel. Therefore, the non-distorted features can be extracted directly on 360° images. Rao *et al.* [15] first applied spherical convolution operation to room layout task and showed significant improvements. Different from the above approaches, Shen *et al.* [16] propose the first panorama Transformer (PanoFormer) and divide the patches on the tangent plane to remove the negative effect of distortions. But most recent layout estimation works do not treat the distortions seriously because they think that the 1D representations can solve the distortions well. On the contrary, we propose a feature assembling mechanism with cross-scale distortion awareness to deal with distortions. Extensive experimental results demonstrate the necessity of handling distortions before generating 1D representations.

3. Approach

3.1. Overview

As illustrated in Fig. 2, the proposed framework consists of three stages: feature extraction followed by a feature assembling mechanism, orthogonal plane disentanglement, and 1D representation reconstruction. Specifically, we adopt a backbone (basically ResNet) to extract a series of features at 4 different scales from a single panorama. Then a feature assembling mechanism is designed to fuse the multi-scale features and free them from panoramic distortions. In the next stage, we introduce a soft-flipping fusion strategy to explore the symmetry between the floor boundary and the ceiling boundary. After that, we segment the vertical and horizontal planes from the fused features followed by vertical compression, yielding two 1D representations. In the last stage, the disentangled features are reconstructed with triple attention to be more discriminative and informative. Finally, the horizon-depth map and room height are estimated from the reconstructed sequences.

3.2. Cross-scale Distortion Awareness

In previous works [19, 20, 21, 8], researchers prefer to follow the architecture of HorizonNet [19] to extract multi-scale feature maps of an equirectangular image. These feature maps are merged together via vertical compression and concatenation [12]. However, the popular architecture just concatenates multi-scale features together and neglects the inherent distortion distribution, resulting in an inferior panoramic feature extraction capability. To address the above issues, we propose a novel feature assembling mechanism to deal with distortions, as well as ingeniously integrate shallow and deep features.

Eliminating Distortions. Motivated by previous works [7, 16], we first polymerize the most relevant features via a virtue tangent plane on the panoramic sphere projection. Based on the projection formula [16, 17, 4] (we discuss it in detail in the supplementary material), we can get the distortion-free sampling coordinates $p \in \mathbb{R}^{H \times W \times 9 \times 2}$ (height, width, number of points, number of coordinates) in the 2D feature maps. The learnable offsets Δp are employed to adjust the sampling locations adaptively. Then we obtain gathered features $f_{df} \in \mathbb{R}^{C \times H \times W \times 9}$ from the original features $f \in \mathbb{R}^{C \times H \times W}$ as follows:

$$f_{df} = \sum_{q=1}^{H \times W} \sum_{k=1}^{9} Sample(f, p_{q,k} + \Delta p_{q,k}), \qquad (1)$$

where q represents the qth point to gather the relevant features; k indexes the kth related features.

Integrating Shallow and Deep features. We select the 3th scale ($\tilde{s}$) as the reference scale and resize other feature maps to this scale. Specifically, we utilize the learnable sampling coordinates $(p_{q,k} + \Delta p_{q,k})$ to gather the distortion-free feature maps for every feature with resized scales following Eq.1. Let $f^s \in \mathbb{R}^{C \times H_s \times W_s}$ be the original feature map at a certain scale s. The gathered multi-scale feature maps $f_{m\tilde{s}} \in \mathbb{R}^{C \times H_{\tilde{s}} \times W_{\tilde{s}} \times 9 \times 4}$ can be represented as:

$$f_{m\tilde{s}} = \sum_{s=1}^{4} \sum_{q=1}^{H_{\tilde{s}} \times W_{\tilde{s}}} \sum_{k=1}^{9} \\ Sample(resize(f^s, \tilde{s}), p_{q,k}^{\tilde{s}} + \Delta p_{q,k}^{\tilde{s}}). \qquad (2)$$

Then, the designed cross-scale distortion-aware feature assembling mechanism can be applied as:

$$f'_{m\tilde{s}} = \ \text{CSDA}(f_{m\tilde{s}}) = \sum_{l=1}^{L} A_l \cdot reshape(f_{m\tilde{s}}), \qquad (3)$$

where L indexes the heads of the self-attention; $A_l \in \mathbb{R}^{L \times H_{\tilde{s}} W_{\tilde{s}} \times 36}$ is the learnable self-attention weights from $f_{m\tilde{s}}$. To complete the calculation, we reshape $f_{m\tilde{s}}$ into the size of $\mathbb{R}^{L \times H_{\tilde{s}} W_{\tilde{s}} \times 36 \times \frac{C}{L}}$. The final assembling features are denoted as $f'_{m\tilde{s}} \in \mathbb{R}^{C \times H_{\tilde{s}} \times W_{\tilde{s}}}$.
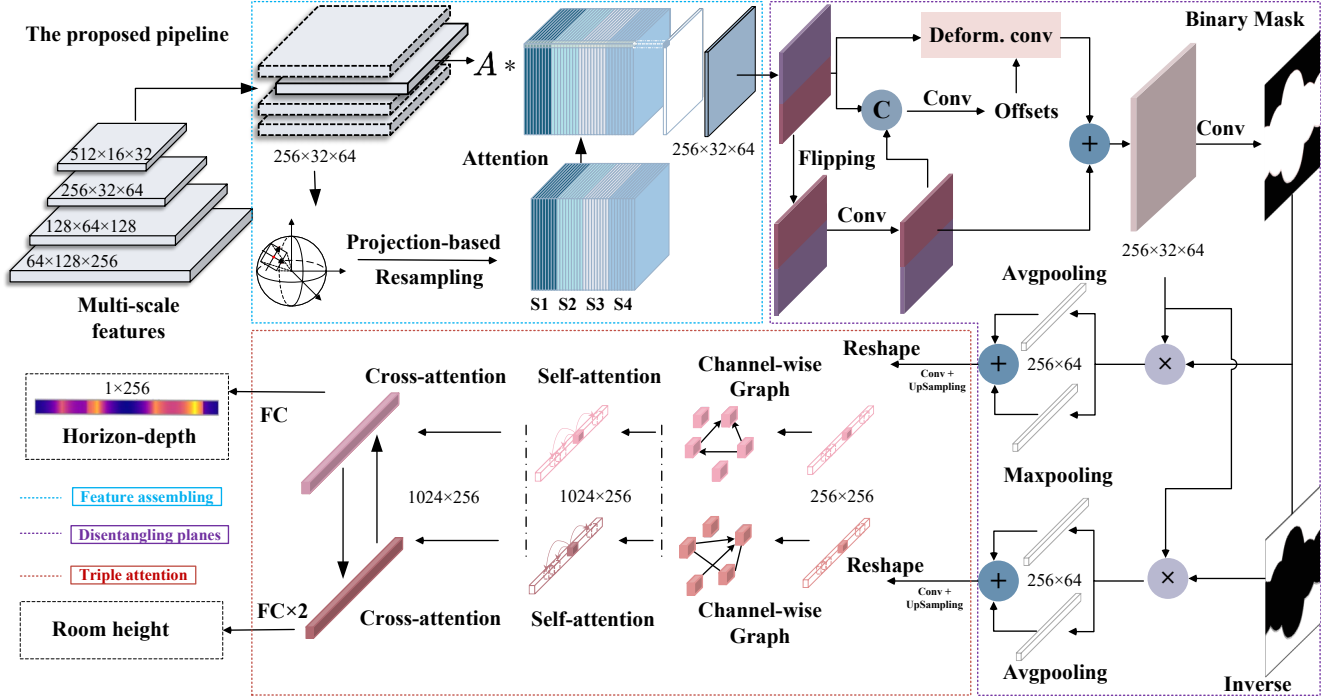
Figure 2. Overview of the proposed framework. The feature assembling mechanism is designed to deal with distortions as well as integrate shallow and deep features (circled in Blue). Then we disentangle orthogonal planes to produce two 1D representations with distinguished plane semantics (circled in Purple). And triple attention is deployed to reconstruct the 1D representations (circled in Red). Our scheme leverages a single panorama with a resolution of 512×1024 as input and predicts the room height and the horizon depth.
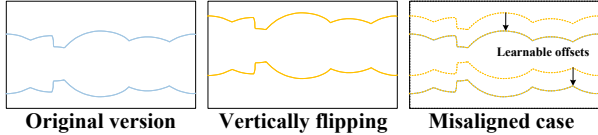

Figure 3. Illustration of the misaligned case when leveraging the symmetry property.

3.3. Disentangling Orthogonal Planes

In order to explicitly capture the geometric cues, we propose to disentangle the orthogonal planes. Particularly, we segment the vertical plane and horizontal planes from the whole scenario before compressing the 2D feature maps.

Symmetry in Manhattan World. It is challenging to learn accurate layout boundaries in a complex scene due to the disturbance of furniture and illumination, e.g., occlusion. To address it, we propose to leverage the symmetry of the Manhattan World assumption (e.g., the floor boundary and ceiling boundary are strictly symmetric in 3D space) to provide complementary information. To exploit this symmetry property, we introduce a soft-flipping fusion strategy.

We first vertically flip the feature maps to get their symmetrical version. In fact, the floor boundary and ceiling boundary are not strictly symmetric in an image because the shooting position is not in the exact middle of the floor and ceiling. To this end, we adopt a deformable convolution with 3×3 kernel size to adaptively adjust the symmetri-

cal version. Next, we fuse the original feature and its "soft" symmetrical version to provide more informative boundary cues. We denote the fused features as $f_u \in \mathbb{R}^{C \times H_{\tilde{s}} \times W_{\tilde{s}}}$.

Segmenting Horizontal/Vertical Planes. To disentangle the 1D representation, we pre-segment the horizon/vertical planes in advance to capture an explicit geometric cue. Specifically, we generate the pseudo labels for the orthogonal planes with layout GT. Then, we use the generated labels to encourage the network to learn a binary mask $S \in \mathbb{R}^{H_{\tilde{s}} \times W_{\tilde{s}}}$ via a simple segmentation head. The two segmented feature maps can be denoted as:

$$
\begin{aligned}
f_h^p &= (f_u \oplus f_{m\tilde{s}}') \otimes sigmoid(S), \\
f_v^p &= (f_u \oplus f_{m\tilde{s}}') \otimes (1 - sigmoid(S))
\end{aligned}
\tag{4}
$$

where f_h^p / f_v^p represents the horizon/vertical plane features, respectively; $\oplus$ is the element-wise sum and $\otimes$ is the element-wise product. Following previous works, we further vertically compress 2D feature maps to two 1D sequences $\mathbf{Q}_h, \mathbf{Q}_v \in \mathbb{R}^{W_{\tilde{s}} \times C}$ with differnet plane semantics.

3.4. Reconstructing 1D Representations

Generating Discriminative Channels. There are generally unconfined dependencies among channels, resulting in confused semantic cues [10]. To this end, we propose a discriminative channels generation mechanism via graph convolution to enforce each channel to concentrate on distin-

guishing features. Different from normal pixel-wise [23] or object-wise [11] graphs, the introduced channel-wise one tends to guide the node to subtract the information from the neighbor nodes rather than aggregation. The formula of the discriminative channels generation mechanism can be represented as:

$$f' = \mathbf{L}fW = (\mathbf{I} - \mathbf{A})fW \qquad (5)$$

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. Please refer to the supplementary materials for more details about the channel-wise graph attention. Following Eq. 5, $\mathbf{Q}_h$ and $\mathbf{Q}_v$ are first reconstructed to be more discriminative in channels (represented as $\mathbf{Q}'_h, \mathbf{Q}'_v \in \mathbb{R}^{W_{\bar{s}} \times C}$).

Rebuilding Long-Range Dependencies. We employ the standard self-attention to reconstruct the intra-sequence long-range dependencies. The attention formula of the horizontal plane sequence can be written as:

$$Attention(\mathcal{Q}'_h, \mathcal{K}'_h, \mathcal{V}'_h) = softmax(\frac{\mathcal{Q}'_h(\mathcal{K}'_h)^T}{\sqrt{d_k}}\mathcal{V}'_h), \quad (6)$$

where $\mathcal{Q}'_h, \mathcal{K}'_h, \mathcal{V}'_h$ are all learned from $\mathbf{Q}'_h$ via a fully connected layer. Similarly, the calculation for the vertical plane sequence is defined as:

$$Attention(\mathcal{Q}'_v, \mathcal{K}'_v, \mathcal{V}'_v) = softmax(\frac{\mathcal{Q}'_v(\mathcal{K}'_v)^T}{\sqrt{d_k}}\mathcal{V}'_v). \quad (7)$$

It captures global interactions to adapt to the large FoV of panoramas, contributing to better performance. We denote the two sequences with reconstructed long-range dependencies as $\mathbf{Q}''_h, \mathbf{Q}''_v \in \mathbb{R}^{W_{\bar{s}} \times C}$.

Providing Missing Residuals. To compensate for inevitable errors in pre-segmentation, we introduce cross-attention to provide the missing residuals. For the sequence about the horizontal plane, we extract the potential beneficial residuals as follows:

$$Attention(\mathcal{Q}''_h, \mathcal{K}''_v, \mathcal{V}''_v) = softmax(\frac{\mathcal{Q}''_h(\mathcal{K}''_v)^T}{\sqrt{d_k}}\mathcal{V}''_v). \quad (8)$$

For the other:

$$Attention(\mathcal{Q}''_v, \mathcal{K}''_h, \mathcal{V}''_h) = softmax(\frac{\mathcal{Q}''_v(\mathcal{K}''_h)^T}{\sqrt{d_k}}\mathcal{V}''_h). \quad (9)$$

where $\mathcal{Q}''_h, \mathcal{K}''_h, \mathcal{V}''_h$ are all learned from $\mathbf{Q}''_h$ ($\mathcal{Q}''_v, \mathcal{K}''_v, \mathcal{V}''_v$ from $\mathbf{Q}''_v$) via a fully connected layer. Then we add the learned residuals with the two 1D sequences to implement the reconstruction. After that, two separate fully connected layers are employed to output the predicted horizon depth from the vertical plane sequence and the room height value from the horizontal plane one, respectively.

3.5. Objective Function

Our objective function consists of two parts: one for room layout estimation and the other for plane segmentation. For the first part, we strictly follow [8] and it can be denoted as $\mathcal{L}_{layout}$. For the segmentation part, we apply binary cross-entropy loss $\mathcal{L}_{segment}$ as follows:

$$\mathcal{L}_{segment} = \frac{1}{N}\sum_{i \in N}(d_i \log \bar{d}_i + (1D_i)\log(1 - \bar{d}_i)) \quad (10)$$

where $\bar{d}_i$ is the ground truth of generated segmentation label, and d_i is the predicted value. Ultimately, we formulate the objective function as follows:

$$\mathcal{L} = \lambda\mathcal{L}_{segment} + \mathcal{L}_{layout} \quad (11)$$

where λ is set to 0.75 to balance different constraints.

4. Experiments

In this section, we validate the effectiveness of our solution and compare it with existing SoTA approaches on four popular datasets. Concretely, we conduct experiments on a single GTX 3090 GPU, and the batch size is set to 16 for training. The proposed approach is implemented with PyTorch. We choose Adam [9] as the optimizer and keep the default settings. The initialized learning rate is 1×10^{-4}. As in previous works [19, 8], we adopt standard left-right flipping, panoramic horizontal rotation, luminance change, and pano stretch for data augmentation during training.

4.1. Datasets

Four datasets are used for our experimental validation: Stanford 2D-3D [1], PanoContext [26], MatterportLayout [29] and ZInd [6].

PanoContext [26] and Stanford 2D-3D [1] are two commonly used datasets for indoor panoramic room layout estimation that contain 514 and 552 cuboid room layouts, respectively. Especially, Stanford 2D-3D [1] is labeled by Zou *et al.* [28] and has a smaller vertical FoV than other datasets. Besides, MatterportLayout [29] is also annotated by Zou *et al.* [29], which contains 2,295 general room layouts. The final ZInd [6] dataset includes cuboid, general Manhattan, non-Manhattan, and non-flat ceilings layouts, which mimic the real-world data distribution better. The splits of ZInd [6] consists of 24,882 (for training), 3,080 (for validation), and 3,170 (for testing) panoramas. We strictly follow the same training/validation/test splits of the four datasets as in previous works for a fair comparison.

4.2. Evaluation Metrics

We employ the commonly used standard evaluation metrics for a fair comparison, including corner error (CE), pixel error (PE), intersection over the union of floor shapes

Method	3DIoU(%)	CE(%)	PE(%)
Train on PanoContext + Whole Stnfd.2D3D datasets			
LayoutNetv2[29]	85.02	0.63	1.79
Dula-Netv2[29]	83.77	0.81	2.43
HorizonNet[19]	82.63	0.74	2.17
LGT-Net[8]	85.16	-	-
LGT-Net [w/ Post-proc][8]	84.94	0.69	2.07
Ours	**85.46**	-	-
Ours [w/ Post-proc]	85.00	**0.69**	2.13
Train on Stnfd.2D3D + Whole PanoContext datasets			
LayoutNetv2[29]	82.66	0.83	2.59
Dula-Netv2[29]	**86.60**	0.67	2.48
HorizonNet[19]	82.72	0.69	2.27
AtlantaNet[14]	83.94	0.71	2.18
LGT-Net[8]	85.76	-	-
LGT-Net [w/ Post-proc][8]	86.03	0.63	2.11
Ours	85.47	-	-
Ours [w/ Post-proc]	85.58	0.66	**2.10**

Table 1. Quantitative comparison results with the current SoTA solutions evaluated on Stanford 2D-3D and PanoContext[6] dataset.

(2DIoU), and 3D room layouts (3DIoU). Among them, 3DIoU yields a better reflection of the accuracy of the layout estimation in 3D space. RMSE and $\delta_{1.25}$ indicate the performance of depth estimation, e.g. the horizon-depth map,

4.3. Cuboid Room Results

We follow LGT-Net [8] to use the combined dataset scheme mentioned in Zou *et al.* [29] to evaluate our network on cuboid datasets. We denote the combined dataset that contains training splits of PanoContext [6] and whole Stanford 2D-3D datasets as "Train on PanoContext + Whole Stnfd.2D3D datasets" in Tab. 1. Similarly, "Train on Stnfd.2D3D + Whole PanoContext" in Tab. 1 represents that we train our network on the combined dataset that contains training splits of Stanford 2D-3D and the whole PanoContext dataset. The scheme is commonly used in previous works [28, 8]. We also report the results with a post-processing strategy in DuLa-Net [24] that is denoted as "Ours [w/ Post-proc]".

Comparison results. We exhibit the quantitative comparison results on cuboid room layouts in Tab. 1. From the first group in Tab. 1, we can observe that our approach outperforms all the other SoTA schemes with respect to 3DIoU. But from the second group, Dula-Net v2 [29] offers better 3DIoU than ours. Dula-Net v2 [29] employs a perspective view (i.e., cubemap) that is more effective for panoramas with small vertical FoV. However, once applied to more general room layout datasets, its performance degrades significantly (it is proved with the other two datasets in Sec. 4.4). On the contrary, the proposed approach shows more general performance.

Method	2DIoU(%)	3DIoU(%)	RMSE	δ_1
LayoutNetv2[29]	78.73	75.82	0.258	0.871
Dula-Netv2[29]	78.82	75.05	0.291	0.818
HorizonNet[19]	81.71	79.11	0.197	0.929
AtlantaNet[14]	82.09	80.02	-	-
HoHoNet[20]	82.32	79.88	-	-
LED^2-Net[21]	82.61	80.14	0.207	0.947
DMH-Net [27]	81.25	78.97	-	0.925
LGT-Net[8]	83.52	81.11	0.204	**0.951**
Ours	**84.11**	**81.70**	**0.197**	0.950

Table 2. Quantitative comparison results with the current SoTA solutions evaluated on MatterportLayout[29] dataset.

Method	2DIoU(%)	3DIoU(%)	RMSE	δ_1
HorizonNet[19]	90.44	88.59	0.123	0.957
LED^2-Net[21]	90.36	88.49	0.124	0.955
LGT-Net[8]	91.77	89.95	0.111	0.960
Ours	**91.94**	**90.13**	-	-

Table 3. Quantitative comparison results with the current SoTA solutions evaluated on ZInD[6] dataset.

4.4. General Room Results

MatterportLayout [29] dataset and ZInD [6] dataset provide more general indoor room layouts, which is much more challenging than the cuboid room layout datasets. Tab. 2/Tab. 3 exhibits the evaluation on MatterportLayout [29]/ZInD [6] datasets. The results of LED^2-Net [21] and HorizonNet [19] are from [8] that we strictly followed. Particularly, Jiang *et al.* utilize their official codeto re-train and re-evaluate with the standard evaluation metrics.

Comparison results. We observe that Dula-Net [24] demonstrates much worse performance on the general room layout dataset (MatterportLayout), indicating that these perspective view-based methods are hard to be adapted to the more general indoor scenarios. Moreover, these 1D sequence-based methods are better than those 2D convolution-based schemes. However, they cannot produce accurate 3D layouts from a 1D sequence due to confused plane semantics, even if they introduce more powerful relationship builders (e.g., Bi-LSTM, Transformer). Hence, these methods give worse 3DIoU. Compared with them, our approach offers better performance than all other approaches with respect to 3DIoU because ours captures an explicit 3D geometric cue by disentangling orthogonal planes, which benefits recovering layouts from clear semantics (Fig. 6).

We show the qualitative comparisons in Fig. 4. From the figure, we can observe that our method offers better 3D results (floor plan). In Fig. 5, our results are much similar to the post-processing version. Both indicate our proposal

(a) Qualitative comparison on MatterportLayout [29] dataset.
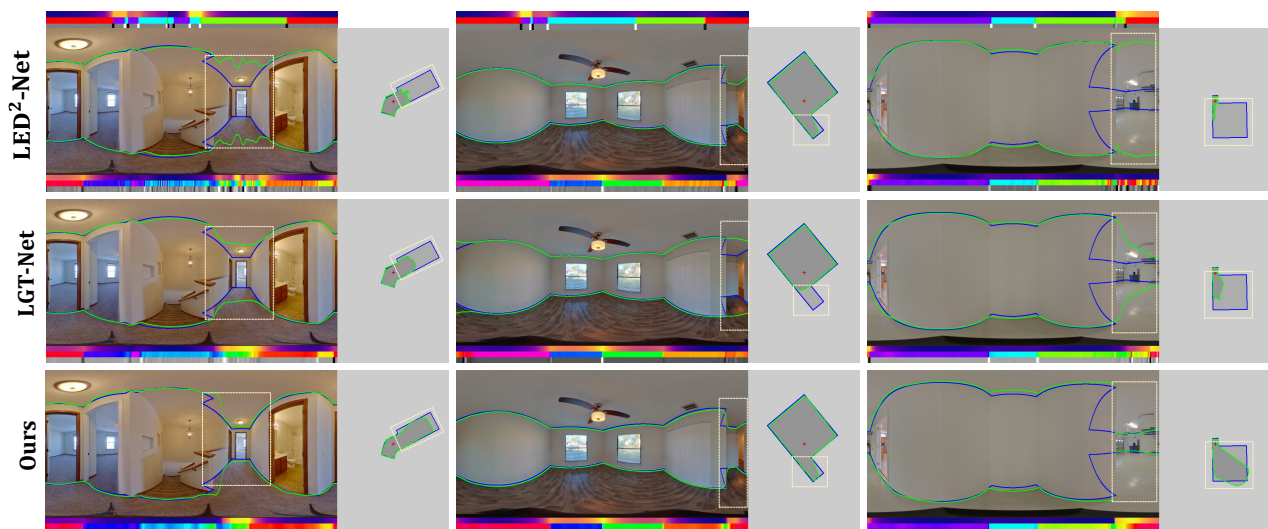
(b) Qualitative comparison on ZInD [6] dataset.

Figure 4. Qualitative comparison results evaluated on general layout datasets, MatterportLayout [29] and ZInD [6]. We compare our method with LED²-Net [21] and LGT-Net [8]. The compared methods are all not employed the post-processing strategy. 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 labeled the significant differences with dashed lines.

that disentangling orthogonal planes to capture an explicit 3D geometric cue is an effective strategy.

4.5. Ablation study

We exhibit ablation studies in Tab. 4, where each component of our model is evaluated on MatterportLayout [29]. To demonstrate the effectiveness of the proposed feature assembling mechanism, we first ablate cross-scale interaction in feature assembling (denoted as "w/o Cross-scale interaction"). Then, we further remove the distortion elimination (denoted as "w/o Feature assembling"). To exhibit the

benefit of orthogonal plane disentanglement, we put off the proposed disentangling orthogonal planes procedure (denoted as "w/o Disentangling planes") and soft-flipping fusion, respectively. Finally, we show the effect of each attention mechanism (denoted as "w/o Discriminative channels", "w/o Long-range dependencies", and "w/o residuals", respectively). Specially, we train each model around 500 epochs (not the best ones). Hence, the results of "Ours [Full]" are slightly lower than those in Tab. 2.

Feature assembling mechanism. From Tab. 4, the pipeline without cross-scale interaction shows inferior per-

(a) Standford 2D-3D [1]
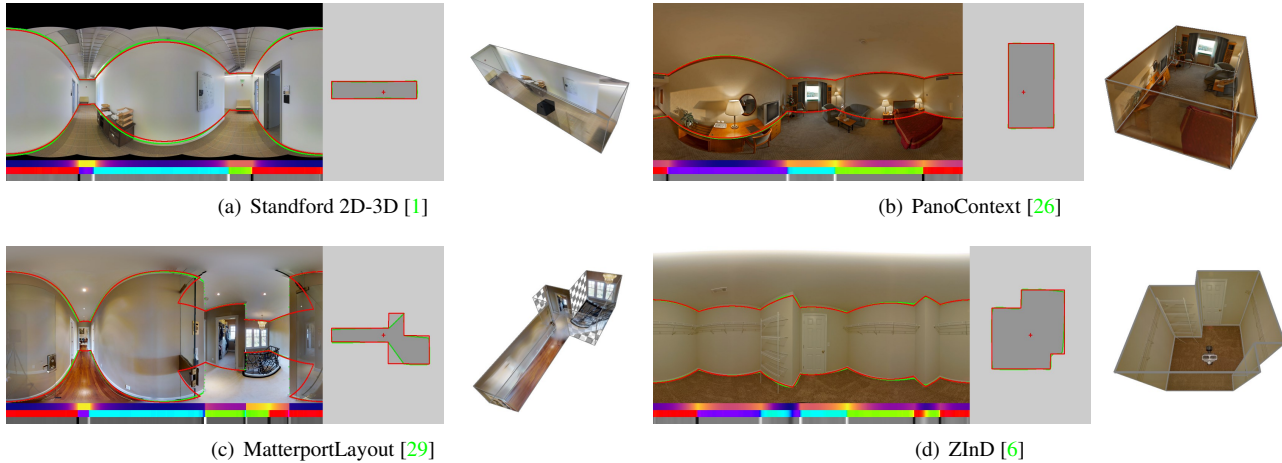
(b) PanoContext [26]

(c) MatterportLayout [29]

(d) ZInD [6]

Figure 5. The 3D visualization results. We exhibit the predicted boundaries (best viewed in Green lines) and the ones with post-processing of the prediction (best viewed in Red lines) in the panoramas.

Figure 6. We exhibit the features of the 1D representation. The one without orthogonal plane disentanglement is shown on the left, and ours is on the right. Without disentangling, the left contains redundant and confusing features that are not good for layout estimation. In contrast, our disentangled vertical plane features are more discriminative, showing more attention to the layout corners.

Method	2DIoU(%)	3DIoU(%)
w/o Cross-scale interaction	83.24	80.71
w/o Feature assembling	82.36	80.10
w/o Disentangling planes	83.23	80.60
w/o Flipping fusion	83.01	80.35
w/o Discriminative channels	82.74	80.23
w/o Long-range dependencies	83.04	80.97
w/o Residuals	82.89	80.45
Ours [Full]	**83.46**	**81.34**

Table 4. Ablation study on MatterportLayout [29] dataset.

formance to the full model. Further removing the distortion elimination part, the pipeline gives worse results. It is proved that both dealing with distortions and integrating cross-scale features are essential for layout estimation.

Disentangling orthogonal planes. Since the geometric cues are essential for inferring 3D information from 2D images, we propose to disentangle orthogonal planes. To prove the effectiveness, we remove that stage but preserve the flipping fusion strategy (the effectiveness of this strategy is validated separately). The results in Tab. 4 show that this stage can make the performance more competitive. Besides, the embedding of symmetry property also contributes to the promotion of the predicted results.

Reconstructing 1D representations. The dependencies of the 1D sequences have changed when disentangling the orthogonal planes. Hence, we propose to reconstruct the 1D

representations. We verify the effectiveness of each attention mechanism in turn. From Tab. 4, we can observe that all three reconstruction operations (generating discriminative channels, rebuilding long-range dependencies, and providing the missing residuals) can all benefit the overall performance. Significantly, when removing the channel-wise graph, the pipeline's performance decreases significantly, demonstrating that enforcing the network concentration on discriminative channel information can capture effective information by avoiding redundancy.

5. Conclusion

In this paper, we propose a novel panoramic indoor room layout estimation approach. Current approaches generate a 1D representation by a vertical compression operation. We argue that this strategy confuses the semantic cues of different planes. To address this issue, 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 the performance. Experiments demonstrate that our algorithm significantly outperforms current SoTA methods.

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, 5, 8

[2] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv:1801.10130*, 2018. 3

[3] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*, 2018. 3

[4] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*, 2018. 3

[5] James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, 1999. 2

[6] 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. 2, 5, 6, 7, 8

[7] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 2020. 2, 3

[8] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[10] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. *Advances in Neural Information Processing Systems*, 2018. 4

[11] Chen Lin, Shuai Zheng, Zhizhe Liu, Youru Li, Zhenfeng Zhu, and Yao Zhao. Sgt: Scene graph-guided transformer for surgical report generation. In *MICCAI*, 2022. 5

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[14] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. At-lantanet: Inferring the 3d indoor layout from a single 360 image beyond the manhattan world assumption. In *ECCV*, 2020. 1, 2, 6

[15] Shivansh Rao, Vikas Kumar, Daniel Kifer, C Lee Giles, and Ankur Mali. Omnilayout: Room layout reconstruction from indoor spherical panoramas. In *CVPR*, 2021. 2, 3

[16] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, 2022. 3

[17] Zhijie Shen, Chunyu Lin, Lang Nie, Kang Liao, and Yao Zhao. Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap. In *ICME*, 2021. 3

[18] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in Neural Information Processing Systems*, 2017. 3

[19] 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. 1, 2, 3, 5, 6

[20] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021. 1, 2, 3, 6

[21] 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. 1, 2, 3, 6, 7

[22] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2

[23] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation by iterative affinity learning. *IJCV*, 128(6):1736–1749, 2020. 5

[24] 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 *CVPR*, 2019. 2, 6

[25] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021. 2

[26] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV*, 2014. 2, 5, 8

[27] Yining Zhao, Chao Wen, Zhou Xue, and Yue Gao. 3d room layout estimation from a cubemap of panorama image via deep manhattan hough transform. In *ECCV*, 2022. 2, 6

[28] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *CVPR*, 2018. 2, 5, 6

[29] 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. 2, 5, 6, 7, 8