LGT-Net: Indoor Panoramic Room Layout Estimation with Geometry-Aware Transformer Network

Zhigang Jiang[1,2] Zhongzheng Xiang[2] Jinhua Xu[1*] Ming Zhao[2]
[1]East China Normal University [2]Yiwo Technology
zigjiang@gmail.com even_and_just@126.com jhxu@cs.ecnu.edu.cn zhaoming@123kanfang.com

Abstract

3D room layout estimation by a single panorama using deep neural networks has made great progress. However, previous approaches can not obtain efficient geometry awareness of room layout with the only latitude of boundaries or horizon-depth. We present that using horizon-depth along with room height can obtain omnidirectional-geometry awareness of room layout in both horizontal and vertical directions. In addition, we propose a planar-geometry aware loss function with normals and gradients of normals to supervise the planeness of walls and turning of corners. We propose an efficient network, LGT-Net, for room layout estimation, which contains a novel Transformer architecture called SWG-Transformer to model geometry relations. SWG-Transformer consists of (Shifted) Window Blocks and Global Blocks to combine the local and global geometry relations. Moreover, we design a novel relative position embedding of Transformer to enhance the spatial identification ability for the panorama. Experiments show that the proposed LGT-Net achieves better performance than current state-of-the-arts (SOTA) on benchmark datasets. The code is publicly available at https://github.com/zhigangjiang/LGT-Net.

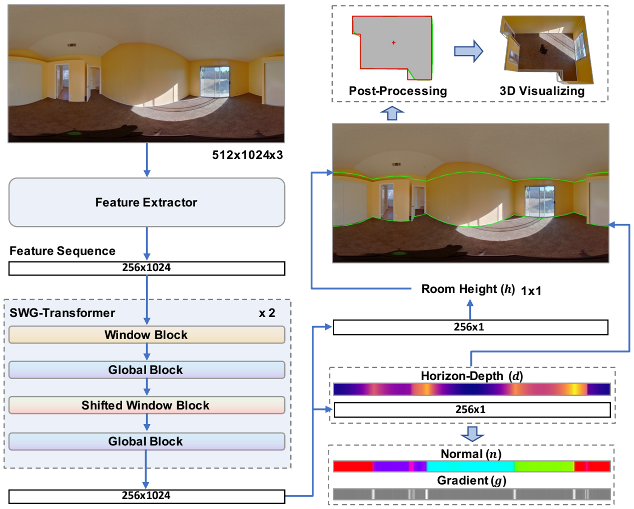


Figure 1. Overall architecture of the proposed LGT-Net. The network estimates the room layout from a single panorama using the *omnidirectional*-geometry aware loss of horizon-depth and room height and the *planar*-geometry aware loss of normals and gradients of normals. We visualize the predicted boundaries (green) by the horizon-depth and room height, and the floor plan (red) with post-processing by Manhattan constraint, finally output the 3D room layout.

1. Introduction

The goal of estimating the 3D room layout by an indoor RGB image is to locate the corners or the floor-boundary and ceiling-boundary, as shown in Fig. 3a, which plays a crucial role in 3D scene understanding [24]. The panoramic images have wider (360°) field of view (FoV) than perspective images and contain the whole-room contextual information [30]. With the development of deep neural networks and the popularity of panoramic cameras in recent years, 3D room layout estimation by a single panorama has made great achievements [23, 28, 32].

Most room layouts conform to the Atlanta World assumption [20] with horizontal floor and ceiling, along with vertical walls [18]. Thus the room layout can be represented by floor-boundary and room height, as shown in Fig. 3a. However, previous approaches [23, 24, 26] estimate the room height by ceiling-boundary. And the networks predict the floor-boundary and ceiling-boundary with the same output branch, which affects each other since they need to predict both horizontal shape and vertical height of room layout. Meanwhile, most previous approaches [23, 28, 32] use Manhattan constraint [3] or directly simplify boundaries [18] in post-processing without considering the planar attribute of the walls to constrain the network output results. In addition, for models [23, 24, 26] which formulate the room layout estimation task as 1D sequence prediction, a sequence processor is needed to model the geometry relationship. Bidirectional Long Short-Term Memory (Bi-

[*]Corresponding author.

LSTM) [11, 21] is used in [23, 26]. Transformer [25] is an efficient framework for sequence processing and has made great success in natural language processing (NLP) tasks. Vision Transformer (ViT) [5] has demonstrated strong abilities in the computer vision field recently. Nevertheless, there is no specially designed Transformer architecture for panoramas as we know.

Due to the above problems, we propose an efficient network called LGT-Net for panoramic room layout estimation. It contains a feature extractor to convert the panorama to feature sequence and a Transformer architecture as sequence processor. Our proposed network directly predicts the room height and floor-boundary by two branches in the output layer, as shown in Fig. 1. Inspired by Wang *et al*. [26], we represent the floor-boundary by horizon-depth. Thus, we propose an *omnidirectional*-geometry aware loss function that computes the errors of horizon-depth and room height, which brings better geometry awareness of the room layout in both horizontal and vertical directions. In addition, we observe the planar attribute of the walls and the turning attribute of the corners. Thus we propose to use the *planar*-geometry aware loss function of normal consistencies and gradient of normal errors to supervise these attributes.

Moreover, we design a novel Transformer architecture called SWG-Transformer as the sequence processor for our network, which consists of (Shifted) Window Blocks [15] and Global Blocks to combine the local and global geometry relations, as shown in Fig. 1. With the attention mechanism [16], our SWG-Transformer can better process the left and right borders of the panoramas than Bi-LSTM. In addition, we design a novel relative position embedding [13, 19, 22] of Transformer architecture to enhance the spatial identification ability for the panoramas.

In order to demonstrate the effectiveness of our proposed approach, we conduct extensive experiments on benchmark datasets, including ZInD [4] dataset. Meanwhile, we conduct ablation study on MatterportLayout [33] dataset in the following aspects: loss function, network architecture, and position embedding of Transformer to demonstrate the effectiveness of each component. Experiments show that our proposed approach performs better than SOTA. The main contributions of our work are as follows:

- We represent the room layout by horizon-depth and room height and output them with two branches of our network. Furthermore, we compute the horizon-depth and room height errors to form *omnidirectional*-geometry aware loss function and compute normal and gradient errors to form *planar*-geometry aware loss function.
- We show that exploiting Transfomer as a sequence processor is helpful for panoramic understanding. And our proposed SWG-Transformer can better establish the local and global geometry relations of the room layout.
- We specially design a relative position embedding of Transformer to enhance the spatial identification ability for the panoramas.

2. Related Work

Panoramic Room Layout Estimation Previous approaches mainly follow the Manhattan World assumption [3] or the less restrictive Atlanta World assumption [20] to estimate the room layout from a panorama and constrain post-processing.

Convolutional neural networks (CNNs) have been used to estimate the room layout with better performance. Zou *et al*. [32] propose LayoutNet to predict probability maps of boundaries and corners and use layout parameter regressors to predict the final layout. Meanwhile, they extend the cuboid layout annotations of the Stanford [1] dataset. Yang *et al*. [28] propose Dula-Net to predict floor and ceiling probability maps under both the equirectangular view and the perspective view of the ceiling. Fernandez *et al*. [8] propose to use equirectangular convolutions (EquiConvs) to estimate the room layout. Sun *et al*. [23] simplify the layout estimation task from 2D dense prediction to 1D sequence prediction. They propose HorizonNet to extract the sequence by a feature extractor based on ResNet-50 [10], then use Bi-LSTM as a sequence processor to establish the global relations. We also use a framework composed of a feature extractor and a sequence processor. Zou *et al*. [33] propose improved version, LayoutNet v2 and Dula-Net v2, which have better performance on cuboid datasets than original approaches, and propose the general MatterportLayout dataset. However, their experiments show that HorizonNet [23] is more efficient on general datasets. Pintore *et al*. [18] propose AtlantaNet to predict floor and ceiling boundary probability maps by same network instance and directly simplify [6] output boundaries as post-processing.

Recently, Wang *et al*. [26] propose LED^2-Net [26] to formulate the room layout estimation as predicting depth on the horizontal plane (horizon-depth), and they can pre-train on synthetic Structured3D [31] dataset with deep information. Sun *et al*. [24] propose HoHoNet to improve HorizonNet by re-designing the feature extractor with the Efficient Height Compression (EHC) module and employing multi-head self-attention (MSA) [25] as a sequence processor instead of Bi-LSTM.

Geometry Awareness Wang *et al*. [26] propose a geometry-aware loss function of the room layout estimation by horizon-depth, which is only effective on horizontal direction. Hu *et al*. [12] propose to use losses of normal and gradient of depth to improve the performance for depth estimation on perspective images. Eder *et al*. [7] propose plane-

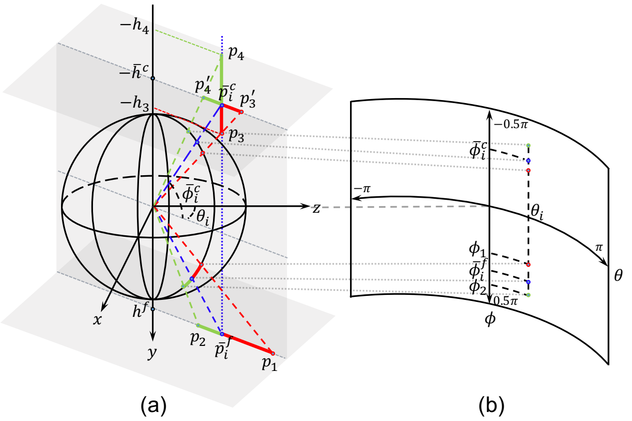


Figure 2. The mapping relationships between 3D points and a panorama. **(a)** The coordinate relations in 3D space, where h^f ($\bar{h}^c$) is the ground truth distance from camera center to the floor (ceiling). **(b)** The longitude and latitude (θ, ϕ) relations on a panorama.

aware loss that leverages curvature, normal, and point-to-plane distance to improve the performance for depth estimation on panoramic images. These works inspire us to propose a more effective geometry awareness loss function.

Transformer Recently, ViT shows that Transformer architecture can compete with CNNs in visual classification tasks. Moreover, improved ViT networks (*e.g.*, T2T-ViT [29], PVT [27], and Swin-Transformer [15]) demonstrate that Transformer architecture is capable of surpassing CNNs. Inspired by Swin-Transformer, we exploit window partition to reduce computation and enhance local modeling ability in SWG-Transformer. However, using window partition alone leads to lower global modeling ability. Thus, our proposed SWG-Transformer consists of (Shifted) Window Blocks and Global Blocks to combine the local and global geometry relations.

3. Approach

Our proposed approach aims to estimate the 3D room layout from a single panorama. We first describe the room layout representation with horizon-depth and room height and show that they can achieve *omnidirectional*-geometry awareness (Sec. 3.1). Then, we introduce our proposed loss function, which consists of *omnidirectional*-geometry aware loss and *planar*-geometry aware loss (Sec. 3.2). Finally, we describe the network architecture of LGT-Net and use the SWG-Transformer to establish the local and global geometry relations of the room layout (Sec. 3.3).

3.1. Panoramic Room Layout Representation

We represent the room layout by the floor-boundary and room height, as shown in Fig. 3a. We adopt a sampling approximation scheme to compute the floor-boundary. Specifically, sample N points $\{p_i\}_{i=1}^N$ with equal longitude intervals from the polygon of floor-boundary, where N is 256 by default in our implementation. The longitudes of the sampling points are denoted as $\{\theta_i = 2\pi(\frac{i}{N} - 0.5)\}_{i=1}^N$. Then, we convert the points $\{p_i\}_{i=1}^N$ to the horizon-depth sequence $\{d_i = D(p_i)\}_{i=1}^N$:

$$\begin{aligned} p &= (x, y, z), \\ D(p) &= \sqrt{x^2 + z^2}. \end{aligned} \tag{1}$$

Thus, we can estimate the room layout by predicting the horizon-depth sequence and room height.

The floor-boundary on the ground plane is sensitive in the horizontal direction, as shown in Fig. 3a. HorizonNet [23] predicts latitudes of ceiling and floor boundaries and calculates errors. However, when the latitude errors of two sampling points are equal (*e.g.*, $|\phi_1 - \bar{\phi}^f_i| = |\phi_2 - \bar{\phi}^f_i|$), the corresponding horizon-depth errors may be different (*e.g.*, $|D(p_1) - D(\bar{p}^f_i)| > |D(p_2) - D(\bar{p}^f_i)|$), as shown in Fig. 2. Thus, we predict horizon-depth and calculate errors to make better geometry awareness of the room layout in the horizontal direction.

Moreover, the room height is sensitive in the vertical direction, as shown in Fig. 3a. LED^2-Net [26] also predicts latitudes but converts the latitudes of floor (ceiling) boundary to horizon-depth by projecting to ground truth floor (ceiling) plane to compute errors. During inference, it calculates the room height by the consistency between the horizon-depth of ceiling and floor boundaries. However, when the ceiling horizon-depth errors of the two sampling points are equal (*e.g.*, $|D(p'_3) - D(\bar{p}^c_i)| = |D(p'_4) - D(\bar{p}^c_i)|$), the corresponding room height errors may be different (*e.g.*, p'_3 and p'_4 are converted to p_3 and p_4 by the consistency of ground truth horizon-depth $D(\bar{p}^c_i)$, and $|h_3 - \bar{h}^c| < |h_4 - \bar{h}^c|$), as shown in Fig. 2. Thus, we directly predict the room height and compute error to make better geometry awareness of the room layout in the vertical direction.

As a result, we propose an *omnidirectional*-geometry aware loss function that computes the errors of horizon-depth and room height. Tab. 4 shows the improvement in our approach.

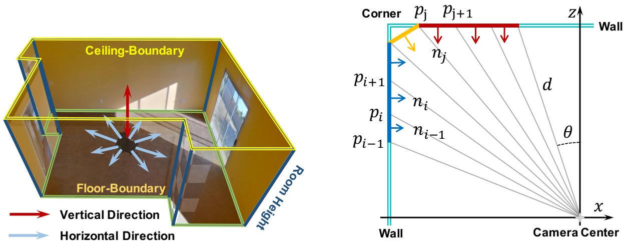
3.2. Loss function

Horizon-Depth and Room Height For the horizon-depth and room height, we apply the L1 loss:

$$\begin{aligned} \mathcal{L}_d &= \frac{1}{N} \sum_{i \in N} |d_i - \bar{d}_i|, \\ \mathcal{L}_h &= |h - \bar{h}|, \end{aligned} \tag{2}$$

where $\bar{d}_i$ ($\bar{h}$) is the ground truth horizon-depth (room height), and d_i (h) is the predicted value.

Normals As shown in Fig. 3, each wall is a plane, but the positions on the same wall may have different horizon-depth (*e.g.*, $D(p_{i-1}) \neq D(p_i)$). However, the normals at



(a) Omnidirectionl Geometry Awareness.

(b) Planar Geometry Awareness.

Figure 3. Illustration of geometry awareness for the room layout. **(a)** The horizontal and vertical directions influence the room layout. We propose *omnidirectional*-geometry aware loss function by horizon-depth and room height. **(b)** The *planar*-geometry awareness by normals.

different positions on the same wall plane are consistent. Thus we use normal consistency to supervise the planar attribute of the walls. Specifically, when the network predicts the horizon-depth sequence $\{d_i\}_{i=1}^{N}$, we convert each horizon-depth d_i to the corresponding 3D point p_i and obtain the normal vector n_i that is always perpendicular to the y-axis. Then we compute the cosine similarity to get the normal loss:

$$\begin{aligned} p_i &= (d_i \sin(\theta_i),\ h^f,\ d_i \cos(\theta_i)), \\ n_i &= M_r\left(\frac{p_{i+1} - p_i}{\|p_{i+1} - p_i\|}\right)^T, \\ \mathcal{L}_n &= \frac{1}{N}\sum_{i\in N}(-n_i \cdot \bar{n}_i), \end{aligned} \tag{3}$$

where M_r is the rotation matrix of $\frac{\pi}{2}$, $\bar{n}_i$ is the ground truth of normal vector, and n_i is the predicted normal vector.

Gradients of normals The normals change near the corners, as shown in Fig. 3b. In order to supervise the turning of corners, we compute the angle between n_{i-1} and n_{i+1} to represent gradient g_i of normal angle, then apply the L1 loss:

$$\begin{aligned} g_i &= \arccos(n_{i-1} \cdot n_{i+1}), \\ \mathcal{L}_g &= \frac{1}{N}\sum_{i\in N}|g_i - \bar{g}_i|, \end{aligned} \tag{4}$$

where $\bar{g}_i$ and g_i are the ground truth and predicted gradients, respectively.

Total Loss The loss terms related to horizon-depth and room height enhance the *omnidirectional*-geometry awareness. And the loss terms corresponding to the normals and gradients enhance the *planar*-geometry awareness. Therefore, to enhance both aspects, we use a total loss function as follows:

$$\mathcal{L} = \lambda\mathcal{L}_d + \mu\mathcal{L}_h + \nu(\mathcal{L}_n + \mathcal{L}_g), \tag{5}$$

where $\lambda, \mu, \nu \in \mathbb{R}$ are hyper-parameters to balance the contribution of each component loss.

3.3. Network

Our proposed LGT-Net consists of a feature extractor and a sequence processor, as shown in Fig. 1. The feature extractor extracts a feature sequence from a panorama. Then, our proposed SWG-Transformer processes the feature sequence. In the end, our network respectively predicts the horizon-depth sequence and a room height value by two branches in the output layer.

Feature Extractor In our implementation, the feature extractor uses the architecture proposed in HorizonNet [23] based on ResNet-50 [10]. The architecture takes a panorama with dimension of $512 \times 1024 \times 3$ (height, width, channel) as input and gets 2D feature maps of 4 different scales by ResNet-50. Then, it compresses the height and up samples width N of each feature map to get 1D feature sequences with same dimension $\mathbb{R}^{N\times\frac{D}{4}}$ and connect them, finally outputs a feature sequence $\mathbb{R}^{N\times D}$, where D is 1024 in our implementation. Moreover, we can also use the EHC module proposed by Sun *et al.* [24] or Patch Embedding [5] of ViT [5] (described in Sec. 4.4) as the feature extractor to extract the feature sequence.

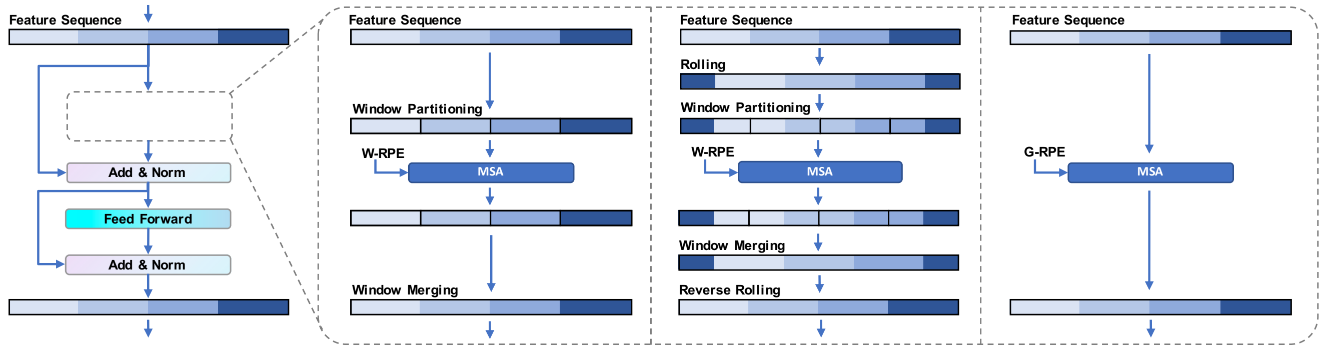
SWG-Transformer In our proposed SWG-Transformer, each loop contains four successive blocks, in the following order: Window Block, Global Block, Shifted Window Block, Global Block. The default loop is repeated twice (×2) for a total of 8 blocks, as shown in Fig. 1. Each block follows the basic Transformer [25] encoder architecture, as shown in Fig. 4a, and the difference lies in the operations before and after MSA. Moreover, the dimension of the sequence and corresponding positions of tokens are the same in the input sequence and output sequence of each block.

In Window Block, we use window partition for the input feature sequence and get $\frac{N}{N_w}$ window feature sequences $\mathbb{R}^{N_w\times D}$ before MSA, where N_w denotes the window length and is set to 16 by default in our implementation. The window partition enhances local geometry relations and reduces the computation when calculating self-attention. Moreover, the window feature sequences are merged after the MSA, as shown in Fig. 4b.

Shifted Window Block aims to connect adjacent windows to enhance information interaction, and it is based on the Window Block. We roll the input feature sequence with $\frac{N_w}{2}$ as its offset before the window partition. To restore the original positions of feature sequence after merging the window feature sequences, we perform a reverse roll operation, as shown in Fig. 4c.

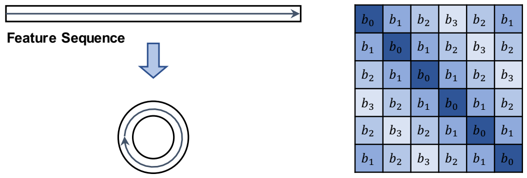
In Global Window Block, operations like window partitioning and rolling are unnecessary. It follows the original Transformer [25] encoder architecture and aims to enhance the global geometry relations, as shown in Fig. 4d.

Position Embedding Since the pure attention module is insensitive to positions of distinguishing tokens, the spatial



(a) Basic Transformer Encoder (b) Window Block (c) Shifted Window Block (d) Global Block

Figure 4. Illustration of SWG-Transformer Blocks. W-RPE and G-RPE are integrated into MSA for each block. **(a)** All blocks are based on the original Transformer [25] encoder. **(b)** Window Block needs to partition and merge windows before and after MSA. **(c)** Shifted Window Block needs to roll and reverse roll sequence feature before and after Window Block operation. **(d)** Global Block does not add additional operations.


Feature Sequence
b_0 b_1 b_2 b_3 b_2 b_1
b_1 b_0 b_1 b_2 b_3 b_2
b_2 b_1 b_0 b_1 b_2 b_3
b_3 b_2 b_1 b_0 b_1 b_2
b_2 b_3 b_2 b_1 b_0 b_1
b_1 b_2 b_3 b_2 b_1 b_0

(a) (b)

Figure 5. **(a)**The feature sequence from panorama is a circular structure. **(b)** Illustration of relative position bias matrix for Global Block.

recognition ability is weakened. Thus, in computing self-attention, we use the relative position embedding of T5 [19] to enhance the spatial identification ability. Specifically, we denote the input sequence of MSA as $X = \{x_i\}_{i=1}^{M}$, where M is the sequence length and $x_i \in \mathbb{R}^D$. A bias matrix $B \in \mathbb{R}^{M\times M}$ is added to Scaled Query-Key product [25]:

$$\begin{aligned} \alpha_{ij} &= \frac{1}{\sqrt{D}}(x_i W^Q)(x_j W^K)^T + B_{ij}, \\ \text{Attention}(X) &= \text{Softmax}(\alpha)(XW^V), \end{aligned} \tag{6}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D\times D}$ are learnable project matrices, each bias B_{ij} comes from a learnable scalar table.

In (Shifted) Window Block, $M = N_w$. We denote the learnable scalar table as $\{b_k\}_{k=-N_w+1}^{N_w-1}$, and B_{ij} corresponds to b_{j-i}. This scheme is denoted as W-RPE and integrated into MSA, as shown in Fig. 4b and Fig. 4c.

In Global Block, $M = N$. As shown in Fig. 5a, the feature sequence is a circular structure. If we use a scheme similar to Window Block and denote the learnable scalar table as $\{b_k\}_{k=-N+1}^{N-1}$, it will result in the same distance represented twice from different directions. Specifically, B_{ij} corresponds to b_{j-i} and also corresponds to b_{j-N-i}. Thus, we propose a *symmetric* representation of only distance and denote the learnable scalar table as $\{b_k\}_{k=0}^{n}$, where $n = \frac{N}{2}$. When $|j - i| \le \frac{N}{2}$, B_{ij} corresponds to $b_{|j-i|}$, otherwise B_{ij} corresponds to $b_{N-|j-i|}$. A visualization of bias matrix is shown in Fig. 5b. We denote this scheme as G-RPE and integrate it into MSA, as shown in Fig. 4d.

4. Experiments

We implement LGT-Net using PyTorch [17] and use the Adam optimizer [14] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the learning rate is set to 0.0001. We train the network on a single NVIDIA GTX 1080 Ti GPU for 200 epochs on ZInD [4] dataset and 1000 epochs on other datasets, with batch size of 6. We adopt the same data augmentation approaches mentioned in Horizon-Net [23], including standard left-right flipping, panoramic horizontal rotation, luminance change, and pano stretch during training. In addition, we set hyper-parameters as $\lambda = 0.9, \mu = 0.1, \nu = 0.1$ in Eq. (5).

4.1. Datasets

PanoContext and Stanford 2D-3D PanoContext [30] dataset contains 514 annotated cuboid room layouts. Stanford 2D-3D [1] dataset contains 552 cuboid room layouts labeled by Zou *et al.* [32] and has a smaller vertical FoV than other datasets. We follow the same training/validation/test splits of LayoutNet [32] to evaluate these two datasets.

MatterportLayout MatterportLayout [33] dataset is a subset of Matterport3D [2] dataset. It contains 2,295 general room layouts labeled by Zou *et al.* [33]. We follow the same training/validation/test splits for evaluation.

ZInD To the best of our knowledge, ZInD [4] dataset is currently the largest dataset with room layout annotations. It better mimics the real-world data distribution since it includes cuboid, more general Manhattan, non-Manhattan, and non-flat ceilings layouts. ZInD [4] dataset contains 67448 panoramas from 1575 real unfurnished residential

Method	3DIoU(%)	CE(%)	PE(%)
Train on PanoContext + Whole Stnfd.2D3D datasets			
LayoutNet v2 [33]	85.02	**0.63**	**1.79**
DuLa-Net v2 [33]	83.77	0.81	2.43
HorizonNet [23]	82.63	0.74	2.17
Ours	**85.16**	-	-
Ours [w/ Post-proc]	84.94	0.69	2.07
Train on Stnfd.2D3D + Whole PanoContext datasets			
LayoutNet v2 [33]	82.66	0.83	2.59
DuLa-Net v2 [33]	**86.60**	0.67	2.48
HorizonNet [23]	82.72	0.69	2.27
AtlantaNet [18]	83.94	0.71	2.18
Ours	85.76	-	-
Ours [w/ Post-proc]	86.03	**0.63**	**2.11**

Table 1. Quantitative results of cuboid layout estimation evaluated on PaonContext [30] (top) and Stanford 2D–3D [1] (bottom) datasets.

homes[1] and separates a "simple" subset that every room layout does not have any contiguous occluded corners. We experiment on the "simple" subset and use the "raw" layout annotations, and follow the official training/validation/test splits at the per-home level. In addition, we filter 0.8% of layout annotations that do not contain the camera center. In total, we have the training, validation, and test splits consisting of 24882, 3080, and 3170 panoramas, respectively.

4.2. Evaluation Metrics

We use the standard evaluation metrics proposed by Zou *et al.* [32]: intersection over union of floor shapes (2DIoU) and 3D room layouts (3DIoU), corner error (CE), and pixel error (PE). Meanwhile, we evaluate the depth accuracy with root mean squared error (RMSE) by using the camera height of 1.6 meters and the percentage of pixels (δ_1) where the ratio between prediction depth and ground truth depth is within a threshold of 1.25 mentioned in Zou *et al.* [33].

4.3. Cuboid Room Results

Since data in a single dataset is limited, it may lead to bias. We use a combined dataset scheme mentioned in Zou *et al.* [33] for training. The combined dataset contains a training split of the current evaluation dataset and another whole dataset. We provide the quantitative results of the cuboid layout in Tab. 1. In addition, some baseline results include post-processing. We also report results with a post-processing of DuLa-Net [28] (denoted as "Ours [w/ Post-proc]"). Meanwhile, CE and PE values are reported.

PanoContext LayoutNet v2 [33] gives slightly better CE and PE performance than ours. And we argue that its 2D convolution for corner location and the post-processing method of gradient ascent is more effective for cuboid layouts. However, our approach offers better performance than all the other SOTA approaches with respect to 3DIoU.

Method	2DIoU(%)	3DIoU(%)	RMSE	δ_1
LayoutNet v2 [33]	78.73	75.82	0.258	0.871
DuLa-Net v2 [33]	78.82	75.05	0.291	0.818
HorizonNet [23]	81.71	79.11	**0.197**	0.929
AtlantaNet [18]	82.09	80.02	-	-
HoHoNet [24]	82.32	79.88	-	-
LED^2-Net [26]	82.61	80.14	0.207	0.947
Ours	**83.52**	**81.11**	0.204	**0.951**
Ours [w/ Post-proc]	**83.48**	**81.08**	0.214	0.940

Table 2. Quantitative results of general layout estimation evaluated on MatterportLayout [33] dataset.

Method	2DIoU(%)	3DIoU(%)	RMSE	δ_1
HorizonNet [23]	90.44	88.59	0.123	0.957
LED^2-Net [26]	90.36	88.49	0.124	0.955
Ours [w/ Pure ViT]	88.93	86.19	0.146	0.950
Ours	**91.77**	**89.95**	**0.111**	**0.960**

Table 3. Quantitative results of general layout estimation evaluated on ZInd [4] dataset.

Stanford 2D-3D Dula-Net v2 [33] gives slightly better 3DIoU than ours, and we argue that it uses perspective view, which is more effective for panoramas with small vertical FoV. However, our approach offers better performance than similar approaches [23, 26] predicting on equirectangular view.

4.4. General Room Results

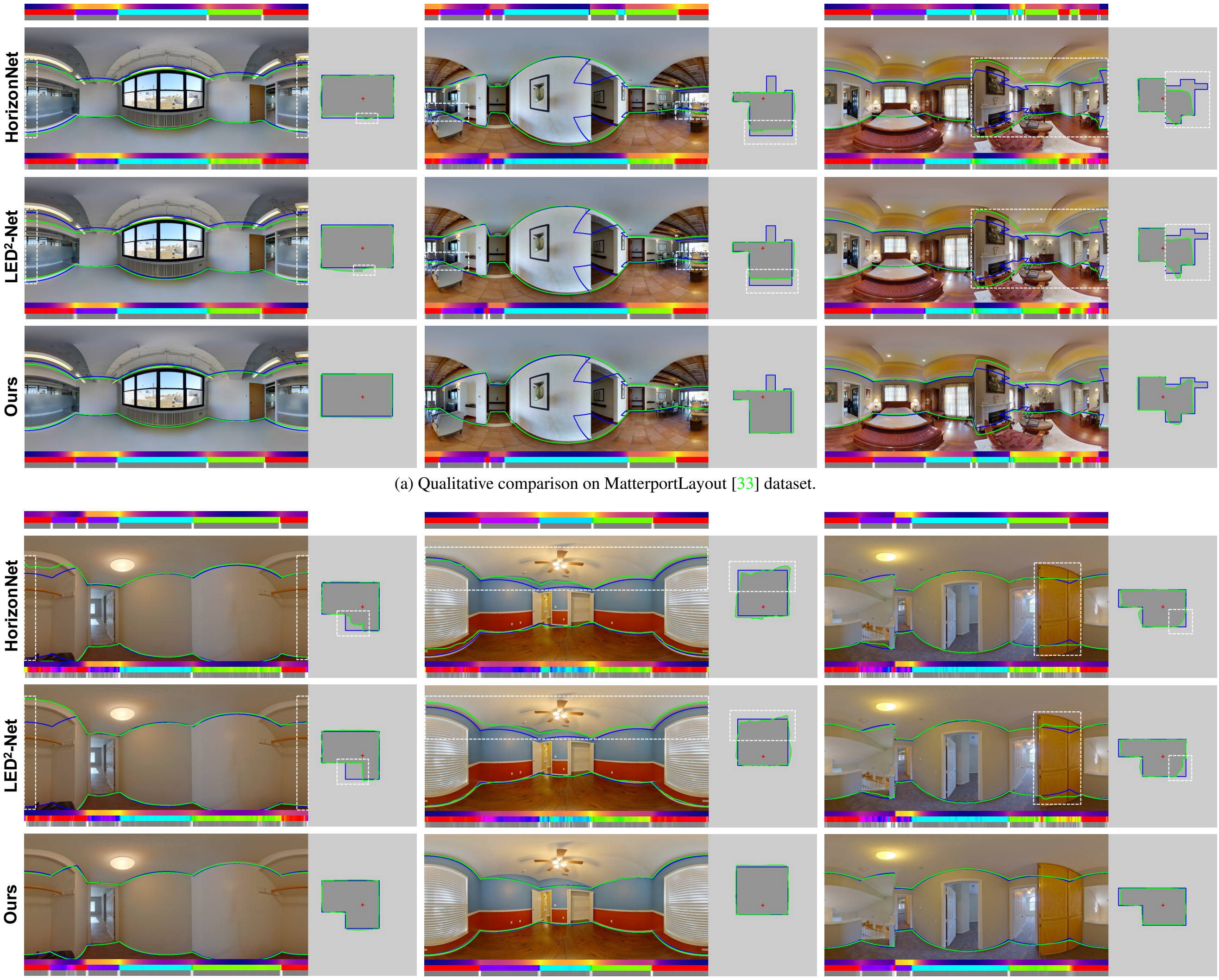
MatterportLayout Evaluation of MatterportLayout [33] dataset is shown in Tab. 2. The results of LED^2-Net [26] are obtained from their official code[2] with re-training and re-evaluating by the standard evaluation metrics. Moreover, we also report results with the post-processing of DuLa-Net [28] (denoted as "Ours [w/ Post-proc]"). Our approach offers better performance than all other approaches with respect to 2DIoU, 3DIoU, and δ_1.

We observe that similar approaches [23, 24, 26] of extracting the 1D feature sequence on equirectangular view are better than those using 2D convolutions [28, 32]. In our opinion, Bi-LSTM [11, 21] and our SWG-Transformer are based on 1D horizontal feature sequence, which are better at establishing relations of the room layout.

Qualitative comparisons are shown in Fig. 6a. The first column shows that HorizonNet [23] and LED^2-Net [26] predict discontinuous layouts at the left and right borders of the panorama because they use Bi-LSTM [11, 21] to process the feature sequence and need to span the entire sequence

[1] https://github.com/zillow/zind

[2] https://github.com/fuenwang/LED2-Net



(a) Qualitative comparison on MatterportLayout [33] dataset.

(b) Qualitative comparison on ZInd [4] dataset.

Figure 6. Qualitative comparison of general layout estimation. We show the room layouts without post-processing by HorizonNet [23], LED^2-Net [26], and ours on MatterportLayout [33] dataset (top) and ZInd [4] dataset (bottom). We show the boundaries of room layout on panorama (left) and the floor plan (right). The blue lines are ground truth, and the green lines are prediction. Moreover, we visualize the predicted horizon-depth, normal, and gradient below each panorama and the ground truth in the first row. The dashed white lines highlight the errors generated by the baselines.

while processing tokens at the first and last position. However, our proposed SWG-Transformer treats tokens equally at all positions. The second and third columns show that our approach better estimates the boundaries far from the camera center and those of complex room layouts. Meanwhile, the visualizations of floor plans, normals, and gradients show that our approach offers better results by *planar*-geometry awareness.

ZInd Evaluation on ZInd [4] dataset is shown in Tab. 3. The results of HorizonNet [23] and LED^2-Net [26] are obtained from their official codes[2][3] with training and evaluating by the standard evaluation metrics. Our approach has higher accuracy than all other approaches under all settings. Moreover, similar to the idea of ViT [5], we split the panorama into patches by Patch Embedding [5] and feed them into our proposed SWG-Transformer (denoted as “Ours [w/ Pure ViT]”). The results show that such ViT architecture achieves comparable performance on the large dataset.

Qualitative comparisons are shown in Fig. 6b. The first column shows that our SWG-Transformer can better process the left and right borders of the panoramas. The second column shows that our proposed *omnidirectional*-geometry awareness has advantage on non-flat ceilings layouts since our approach is not affected by ceiling-boundary. The third column shows that our approach performs better with furni-

[3] https://github.com/sunset1995/HorizonNet

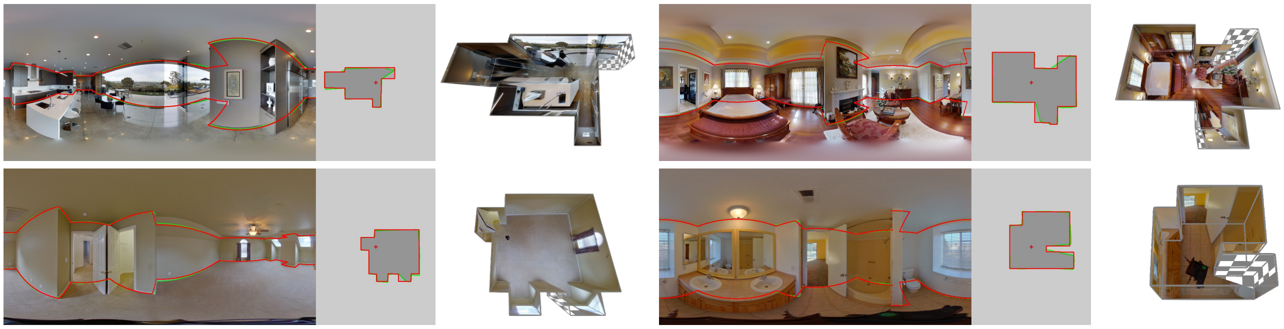


Figure 7. The 3D visualization results of our approach on MatterportLayout [33] dataset (first row) and ZInd [4] dataset (second row). The green lines are predicted boundaries by our network, and the red lines are results with post-processing of the prediction.

Method	2DIoU(%)	3DIoU(%)	RMSE	δ_1
w/o Height	82.82	80.44	0.205	0.945
w/o Nomal+Gradient	84.24	81.86	0.196	0.954
w/o Gradient	84.27	81.89	**0.194**	0.954
w/ Pure ViT	64.05	60.44	0.434	0.782
w/o Global Block	83.02	80.40	0.212	0.947
w/ Bi-LSTM	83.98	81.32	0.201	0.950
w/o Window Block	83.96	81.47	0.197	**0.958**
w/o PE	83.78	81.50	0.197	0.951
w/ APE	83.90	81.55	0.201	0.951
Ours [Full]	**84.38**	**82.01**	**0.194**	0.955

Table 4. Ablation study on MatterportLayout [33] dataset.

ture occlusion than other approaches [23, 26].

The 3D visualization results of our approach on MatterportLayout [33] dataset and ZInd [4] dataset are shown in Fig. 7. These examples show that our approach is effective in room layout estimation. See supplemental material for more qualitative results and quantitative results of different corners number and cross-dataset evaluation.

4.5. Ablation Study

Ablation study is shown in Tab. 4. We reported results of the best performance of each configuration on the test split of MatterportLayout [33] dataset. It should be noted that all experiments of ablation study select the best epoch in the test split. Thus, the results of "Ours [full]" are higher than the corresponding quantitative results.

Loss Function We replace the loss function in our approach with floor and ceiling horizon-depth errors like LED^2-Net [26] (denoted as "w/o Height") and show that our proposed *omnidirectional*-geometry aware loss of horizon-depth and room height significantly improves performance. Moreover, our experiments without the normal and gradient errors (denoted as "w/o Normal+Gradient" and "w/o Gradient") show that our proposed *planar*-geometry aware loss by normals and gradients of normals improves the performance.

Network Architecture We experiment with ViT architecture (denoted as "w/ Pure ViT") and show that ViT architecture does not achieve comparable performance in MatterportLayout [33] dataset. We argue that ViT architecture relies on large datasets like ZInd [4] to perform better. Moreover, our experiments without Global Blocks or (Shifted) Window Blocks (denoted as "w/o Global Block" and 'w/o Window Block") demonstrate that using Window Blocks or Global Blocks alone leads to lower performance. We replace SWG-Transformer with Bi-LSTM [11, 21] (denoted as "w Bi-LSTM") and show that our SWG-Transformer offers better performance than Bi-LSTM.

Position Embedding We experiment without position embedding (denoted as "w/o PE") and only use absolute position embedding [9] with learnable parameters (denoted as "w/ APE"). These experiments show that absolute position embedding does not bring much improvement, but our designed relative position embedding offers the best performance. We believe that since the contexts of panoramas constantly change in the horizontal direction, it is difficult to map the changes with a fixed absolute position embedding.

5. Conclusions

In this paper, we proposed an efficient model, LGT-Net, for 3D room layout estimation. Horizon-depth and room height offer omnidirectional geometry awareness. Normals and gradients of normals offer planar geometry awareness. Moreover, the proposed SWG-Transformer with the noval relative position embedding can better establish the local and global geometry relations of the room layout. We evaluate our approach on both cuboid and general datasets and show better performance than the baselines.

Acknowledgement This work is supported by the National Natural Science Foundation of China under Project 61175116.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 5, 6

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017. 5

[3] James M Coughlan and Alan L Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 941–947. IEEE, 1999. 1, 2

[4] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2133–2143, 2021. 2, 5, 6, 7, 8

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2, 4, 7

[6] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2):112–122, 1973. 2

[7] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In *2019 International Conference on 3D Vision (3DV)*, pages 76–84. IEEE, 2019. 2

[8] 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*, 5(2):1255–1262, 2020. 2

[9] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017. 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 4

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 6, 8

[12] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE Computer Society, 2019. 2

[13] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2020. 2

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations*, 2015. 5

[15] 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. 2021. 2, 3

[16] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 2

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

[18] 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. 1, 2, 6

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020. 2, 5

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004. 1, 2

[21] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 2, 6, 8

[22] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018. 2

[23] 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. 1, 2, 3, 4, 5, 6, 7, 8

[24] 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. 1, 2, 4, 6

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 4, 5

[26] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360° layout estimation via differentiable depth rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12956–12965, 2021. 1, 2, 3, 6, 7, 8

[27] 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. 2021. 3

[28] 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. 1, 2, 6

[29] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. 2021. 3

[30] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European Conference on Computer Vision*, pages 668–686. Springer, 2014. 1, 5, 6

[31] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pages 519–535. Springer, 2020. 2

[32] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2051–2059, 2018. 1, 2, 5, 6

[33] 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. *International Journal of Computer Vision*, 129(5):1410–1431, 2021. 2, 5, 6, 7, 8