PanoContext-Former: Panoramic Total Scene Understanding with a Transformer

Yuan Dong[1]*, Chuan Fang[2]*, Liefeng Bo[1], Zilong Dong[1], Ping Tan[2]†

[1] Alibaba Group. [2] Hong Kong University of Science and Technology

[1]{dy283090, liefeng.bo, list.dzl}@alibaba-inc.com, [2]cfangac@connect.ust.hk [2]pingtan@ust.hk

Abstract

Panoramic images enable deeper understanding and more holistic perception of 360° surrounding environment, which can naturally encode enriched scene context information compared to standard perspective image. Previous work has made lots of effort to solve the scene understanding task in a hybrid solution based on 2D-3D geometric reasoning, thus each sub-task is processed separately and few correlations are explored in this procedure. In this paper, we propose a fully 3D method for holistic indoor scene understanding which recovers the objects' shapes, oriented bounding boxes and the 3D room layout simultaneously from a single panorama. To maximize the exploration of the rich context information, we design a transformer-based context module to predict the representation and relationship among each component of the scene. In addition, we introduce a new dataset for scene understanding, including photo-realistic panoramas, high-fidelity depth images, accurately annotated room layouts, oriented object bounding boxes and shapes. Experiments on the synthetic and new datasets demonstrate that our method outperforms previous panoramic scene understanding methods in terms of both layout estimation and 3D object detection.

1. Introduction

Single-view indoor scene understanding from a single RGB image is an essential yet challenging problem and has important applications such as augmented reality and service robotics. Most of the existing works solve room layout estimation, object detection, and reconstruction separately. Some recent works, including Factor3D [50], CoOp [19], Total3D [35], and IM3D[63], show that learning these tasks jointly helps to improve the performance on each subtask by exploiting context information. In addition, panoramic image with a 360° field-of-view (FOV) contains much richer

*Equal Contribution.
†Corresponding author.

Figure 1. Given a single RGB panorama, we simultaneously estimate the room layout, oriented object bounding boxes (left), and full scene meshes (right). The first and second rows are examples from the iGibson-Synthetic [62] and ReplicaPano datasets.

information than a regular perspective image, whose FOV is normally around 60°. PanoContext [64] and DeepPanoContext [62] prove that the context becomes significantly more robust and powerful with a larger FOV, which further improves the performance and enables accurate holistic scene understanding. Despite recent progress, the indoor scene understanding problem remains challenging since predicting object pose and shape from a single RGB image can be ambiguous without any 3D prior information in a real indoor environment with occlusion and clutter.

This paper proposes a new method for end-to-end total 3D scene understanding from a single panorama (Fig. 1). Our approach has two important features. Firstly, we lift the 2D panorama into 3D space to exploit 3D information to facilitate indoor scene understanding tasks. In this way, our method can optimize room layout estimation, objects detection and reconstruction simultaneously in an end-to-end fashion. Our experiments show that converting the 2D panorama as a 3D representation in a scene understanding framework can boost performance remarkably. Additionally, we observe that the object features used to estimate boxes also contain information on geometries; therefore,

different from previous methods [35, 62, 63], we learn a shape code to represent each object using an MLP without training a separate network.

Secondly, in order to better capture the 3D context in the scene, we unify different tasks together and propose a novel transformer-based context module for simultaneously predicting object shapes, oriented bounding boxes and 3D room layout. The key insight of this model is to take various scene components as input to discover spatial relationships among object-object and object-layout. Specifically, as the 2D image can provide more high-level semantic information, we blend not only the extracted features in 3D space but also the global image features from panorama. In addition, the random token masking strategy is employed to strengthen interactions across objects and room layout. We verify the benefits of using global image features and token masking strategy in ablation study.

When it comes to the panoramic datasets for holistic scene understanding, more efforts should be put into this area. Existing panoramic datasets are either for single application [52, 56, 59, 66] or missing critic 3D ground truth such as object boxes [1, 3] and object shapes [64, 65]. Compared with annotating the oriented object boxes and 3D shapes which is extremely labor-costing, it could be easier to generate ground truth from a simulator. Zhang et al. [62] release the first holistic panoramic scene dataset with complete ground truth, rendering from synthetic scenes, while the panoramas lack realism and may set the barrier to deploy the algorithm into real-world. To minimize the domain gap between synthetic and real data, we render photo-realistic panoramas and depth images based on high-fidelity scene scan [44], then label room layout, 3D object boxes and shapes semi-automatically. Qualitative evaluations indicate that this new dataset offers accurate and comprehensive ground truths.

In general, the main contributions of our work can be summarized as follows:

- We propose a new fully 3D method for total scene understanding from a single RGB panorama in an end-to-end fashion, which will better preserve the intrinsic structure of the indoor scene.
- We introduce a novel transformer-based context module with global image feature and token masking mechanism that takes the environmental context into account in a single step.
- We introduce ReplicaPano, a new dataset comprising photo-realistic images, real-world object arrangement, room layout, and object meshes for panoramic 3D scene understanding.
- The proposed method achieves state-of-the-art performance on both synthetic and real-world datasets.

2. Related work

Single-View Scene Understanding Scene understanding from a single image is highly ill-posed and ambiguous because of the unknown scale and severe occlusion in the scene. Many works have been proposed to study room layout estimation, 3D object detection and pose estimation, and 3D object reconstruction. Early room layout estimation works often make cuboid assumption [9, 17, 26, 33] or Manhattan assumption [45, 46, 53, 59, 60, 67], while Pintore et al. [38] model room structure as a 3D mesh to exploit the possibility of estimating arbitrary room layout. Object detection works [6, 13, 19, 49] aim to infer 3D bounding boxes and object poses from 2D representation, with a 2D object detection [8, 16] stage. In terms of object reconstruction, CAD models are selected from a large dataset to match the 2D object proposals in [20, 21, 24], while [7, 15, 34, 36, 37] demonstrated that implicit neural representations outperform grid, point, and mesh-based representations in parameterizing geometry and seamlessly allow for learning priors over shapes.

Some recent works start to solve multiple tasks together to exploit context information. CooP [19] introduces the target parameterizing and cooperative training scheme to solve for object poses and the layout of the indoor scene, but object shapes are absent. Total3D [35] is the first work to solve layout, 3D object detection and pose estimation, and object reconstruction jointly. Zhang et al. [63] proposes to improve the performance of all three tasks via implicit neural functions and graph convolutional networks. Liu et al. [29] further improves the visual quality of indoor scene reconstruction using implicit representation. All these aforementioned methods only work on the perspective images, which lack enough information to better parse the entire scene. Zhang et al. [64] first introduced to parse indoor scenes using 360° full-view panorama. Then, the follow-up work [62] utilizes a hybrid framework that leverages image information and scene context to estimate objects' shapes, 3D poses and the room layout from a single panorama. Instead, we propose to lift the 2D input to 3D representation and design a transformer-based context module for the panoramic scene understanding task, which can fully explore spatial context information among different components of an indoor scene.

Transformer Transformer [10, 30, 51] has been the dominant network in the field of NLP for a few years. Inspired by ViT [12], researchers have designed many efficient networks [2, 31, 48, 55, 61] to combine the advantages of both CNNs and transformers.

The review [57] shows that the transformer structure can better learn context information among multi-modal input data. CLIP [39] jointly trains the image encoder and text encoder at the pretraining stage and converts an image classification task as a text retrieval task at test time. Hu
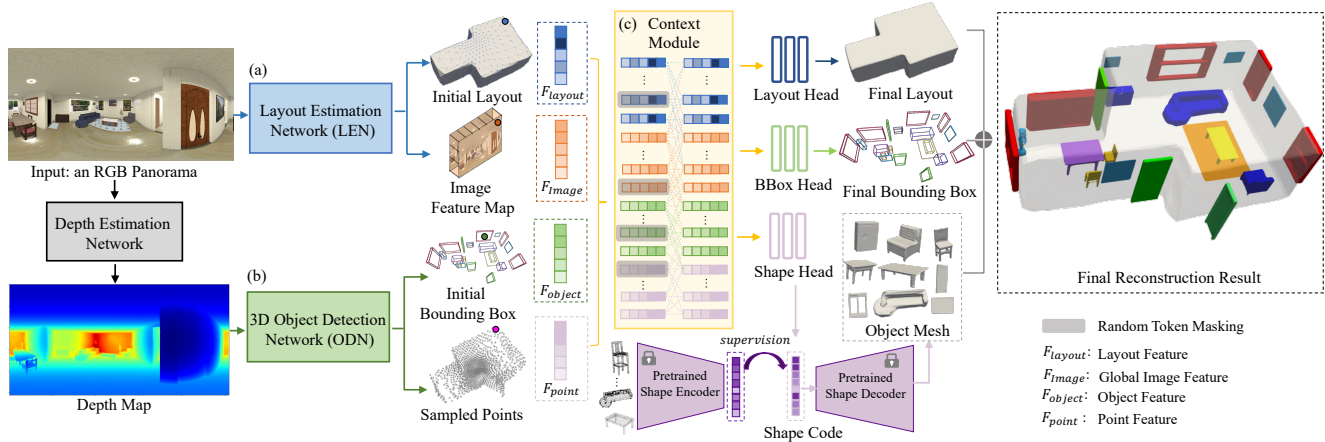
Figure 2. The framework of the proposed holistic scene understanding pipeline. (a) The LEN module maps an panorama to a watertight 3D mesh of the room layout. (b) The ODN module jointly solves the oriented object bounding box and shape based on the estimated point cloud of the indoor scene. (c) The Context Module integrates various embeddings from LEN and ODN modules to fully explore the relationship among each component of the scene. Finally, refined features go through different heads, and the layout, oriented object bounding boxes, and shapes are recovered to reconstruct the full scene.

and Singh [18] combined image and text to conduct multi-modal multi-task training and achieved good results in 7 visual and text tasks. Group-Free [32] and Mask3D [40] utilize the attention mechanism in transformers to fuse the object features and point features iteratively to improve 3D object detection and 3D semantic segmentation respectively. PQ-Transformer [5] and AnchorRec [11] further predicts 3D objects and layouts simultaneously from point clouds, but has limitations in parsing complex and inclined walls. Similarly, conditional object query was used in [54] to fuse point cloud and image features to obtain better results on the 3D detection task. There is a notable advantage of transformers for multi-modal tasks, in this paper, we introduce a transformer-based context module to facilitate holistic indoor scene understanding.

Panoramic Dataset SUN360 [56] is the first real-world panoramic dataset used for place recognition, then it is annotated by Zhang et al. [64] for indoor scene understanding, but only room layout and objects' axis aligned bounding box are provided. 2D-3D-S [1] and Matterport3D [3] are published concurrently with real-world panoramas, but oriented object boxes and meshes are absent. In addition, there are some datasets [52, 59, 66] published recently for the purpose of depth estimation or layout estimation on panorama. Zheng et al. [65] propose a large photo-realistic panoramic dataset for structured 3D modeling, namely Structured3D, but both mesh ground truths of scenes and objects are not published. To tackle that there is no panorama dataset with complete ground truths, author in [62] uses iGibson [41] to synthesize 1500 panoramas with detailed 3D shapes, poses, semantics as well as room layout. However, the real-world indoor scene

dataset containing all ground truth is still missing. To minimize the gap, we introduce a new dataset rendered from real-scan [44], containing 2,700 photo-realistic panorama, real-world room layouts, and object bounding boxes and shapes.

3. Our Method

The proposed pipeline simultaneously predicts the room layout, oriented 3D object bounding boxes and shapes. As shown in Fig. 2, we use a depth estimator to lift the 2D panorama into the point cloud, which is then fed into the Object Detection Network (ODN) to jointly predict 3D object boxes and shape codes. In the meantime, the layout is recovered as a triangle mesh from the input through the Layout Estimation Network (LEN). In this paper, we exploit the transformer's intrinsic advantages and scalability in modeling different modalities and tasks, making learning appropriate spatial relationships among objects and layout easier. Features from layout, image, and 3D objects are fed into the context module to better estimate representations and relations among objects and layout. Finally, the room layout and object shapes are recovered as mesh, then scaled and placed into appropriate locations to reconstruct the full scene. We elaborate on the details of each module in this section.

3.1. Layout Estimation

We model the room layout as a 3D mesh from the input panorama, which also relaxes the geometrical constraints (e.g., forcing vertical walls and/or planar walls and ceilings). Similar to Deep3DLayout [38], we map panoramic

image to a triangle mesh representation (V, E, F), where $V(n, 3)$ is the set of $n = 642$ vertices, $E(m, 2)$ is the set of m edges, each connecting two vertices, and $F(n, d)$ are the image feature vectors of dimension $d = 288$ associated to vertices, denoted as $\mathbf{F_{layout}}$ in the following. Two Graph Convolution Network (GCN) blocks deform an initial tessellated sphere by offsetting its vertices, driven by associating image features to mesh vertices in a coarse-to-fine form. Unlike Deep3dlayout [38] which only extracts features from equirectangular view, we additionally extract features from perspective views (e.g., ceiling and flooring views) through Equirectangular-to-Perspective (E2P) conversion. Then, E2P-based feature fusion [59] is employed to fuse two types of features and get gravity aligned features. Specifically, we use ResNet-18 as the architecture for both equirectangular view and perspective views, the input dimension of image $\mathbf{I}$ is $3 \times 512 \times 1024$, the output dimension of fused global image feature $\mathbf{F_{image}}$ is $512 \times 16 \times 32$. The ablation experiment in Sec. 4.3 shows that the accuracy of room layout benefits from perspective features.

Drawing on the previous multi-modal transformer models [27, 28], in order to fully utilize the layout information and the high-level image information, we inject the global image feature $\mathbf{F_{image}}$ and layout features $\mathbf{F_{layout}}$ from the first GCN block into the context module. Then the refined layout representation is sent into the layout head (the second GCN block). As a result, the second block returns the final deformed vertices $V^*(4n - 6, 3)$ for the layout mesh.

3.2. 3D Object Detection and Mesh Generation

After lifting the panorama into 3D space as a point cloud, our ODN can accurately detect the oriented bounding box and regress shape code for each object. We adopt a similar strategy as Group-Free [32], first downsampling the point cloud through Fabinacci sampling and then feeding the downsampled point cloud $S \in \mathbb{R}^{N \times 3}$ into the backbone network and the Initial Object Sampling module to get point cloud features and K initial object candidates denoted as $\mathbf{F_{point}} \in \mathbb{R}^{d_o \times M}$ and $\mathbf{F_{object}} \in \mathbb{R}^{d_o \times K}$ respectively, where $K = 256, M = 1024$ and the feature dimension $d_o = 288$. To automatically learn the contribution of all points to each object, these intermediate results will serve as point tokens and object tokens in the next subsection.

Inspired by [22, 23], we observe that shape information is embedded in the object feature in the process of 3D object detection. Thus, in addition to the existing object prediction head, we add a shape prediction head to jointly predict the shape latent code and bounding box of the candidate object. The shape latent code is supervised by a pretrained autoencoder of object mesh, here we choose ONet [34] to serve as the autoencoder, because of its computation-friendly size of object shape latent code (1D vector of size 512), which can be easily used to construct the shape loss during the train-

ing. The ONet is pre-trained on ShapeNet [4] and refined on iGibson-Synthetic [62] with data augmentation.

3.3. Transformer-based Context Module

Given a single panorama, our goal is to further explore the intrinsic relationships among different components of the indoor scene. We designed the transformer-based context module with a multi-layer encoder structure to extract better representations of objects and room layouts from different features. As shown in Fig. 3, the position embeddings of point, object, layout, and global image are computed by applying independent linear layers on the parameterization vector of point (x, y, z), 3D box (x, y, z, l, h, w), layout vertice (x, y, z), and unit spherical coordinate $(\cos\phi\sin\theta, \sin\phi, \cos\phi\cos\theta)$, respectively. The global image feature $\mathbf{F_{image}}$ along with point feature $\mathbf{F_{point}}$, object feature $\mathbf{F_{object}}$, and layout feature $\mathbf{F_{layout}}$ are point-wise summed with their position embeddings and then are concatenated together and act as the input for the context module:

$$\mathbf{Z} = [\mathbf{F_{image}}, \mathbf{F_{layout}}, \mathbf{F_{point}}, \mathbf{F_{object}}]. \quad (1)$$

The context module is composed of 6 stacked transformer encoder layers, each layer includes a multi-head self-attention (MHSA) layer and a feed-forward network (FFN). MHSA is the foundation of a transformer, allowing the model to jointly attend to information from different representation subspaces. In a self-attention module, embedding $\mathbf{Z}$ will go through three projection matrices ($\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V$) to generate three embedding $\mathbf{Q}$(query), $\mathbf{K}$(key) and $\mathbf{V}$(value):

$$\mathbf{Q} = \mathbf{ZW^Q}, \mathbf{K} = \mathbf{ZW^K}, \mathbf{V} = \mathbf{ZW^V}. \quad (2)$$

The output of self-attention is the aggregation of the values that are weighted by the attention weights. In our case, we propose a token random masking scheme to help the encoder to be robust and effective in handling situations with heavy occlusions, formulated as:

$$\text{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \odot \mathbf{M} \right) \mathbf{V}, \quad (3)$$

where d is the dimension of query embedding and $\mathbf{M}$ is the specific masking matrix. Multiple self-attention layers are stacked and their concatenated outputs are fused by weighting matrix $\mathbf{W^h}$, to form MHSA:

$$\text{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sum_{h=1}^{H} \text{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})\mathbf{W^h}. \quad (4)$$

After iterative refinement of MHSA, the resulting embedding of different stages are fed into different prediction heads to generate the results of each task, which will be ensemble to produce superior results.
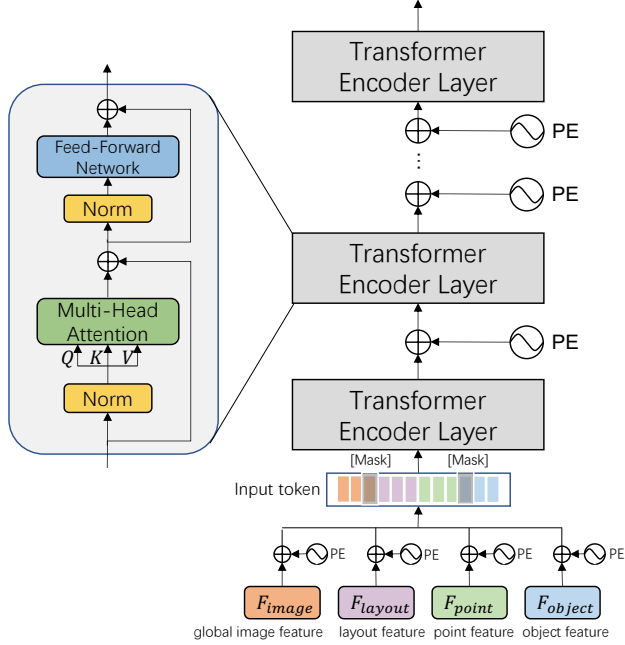
Figure 3. Architecture of the transformer-based context module.

3.4. Loss Function

In this section, we summarize the learning objectives alongside their corresponding loss functions, and we describe our composite loss for end-to-end training.

Layout Loss At first, we adopt the loss function from Pintore et al. [38] to define layout loss which measures the prediction with respect to the ground truth layout:

$$\mathcal{L}_{layout} = \lambda_p \cdot \mathcal{L}_{pos} + \lambda_n \cdot \mathcal{L}_{norm} + \lambda_e \cdot \mathcal{L}_{sharp}. \quad (5)$$

where $\mathcal{L}_*$ and λ_* are the losses and coefficients for vertex position, surface normal, and edge sharpness, respectively.

Object Loss The loss for ODN is similar to [32], including sampling loss $\mathcal{L}_{samp}$, objectness loss $\mathcal{L}_{objness}$, classification loss $\mathcal{L}_{cls}$, center offset loss $\mathcal{L}_{cen}$, size classification loss $\mathcal{L}_{size\_cls}$, and size offset loss $\mathcal{L}_{size\_off}$. Additionally, 1) since we aim to estimate the oriented bounding box of the object, the box's heading prediction with a cross-entropy loss $\mathcal{L}_{head\_cls}$ and a smooth-L1 loss $\mathcal{L}_{head\_off}$ is included. 2) the shape code prediction loss $\mathcal{L}_{shape}$. Let θ denote the estimated shape codes, we use a smooth-L1 loss to minimize the errors between predictions and ground truth:

$$\mathcal{L}_{shape} = \frac{1}{K} \sum_{k=1}^{K} \ell_1(\theta - \bar{\theta}), \quad (6)$$

where ground truths $\bar{\theta}$ are given from pre-trained autoencoder. For the sake of brevity, these losses will be referred to as a set $\{\mathcal{L}_{object\_loss}\}$.
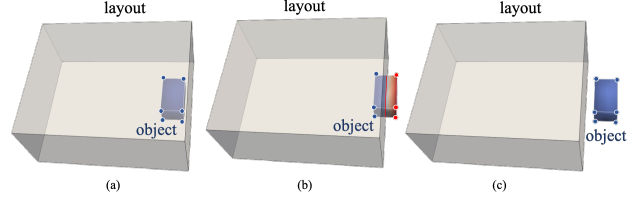

Figure 4. Object-Layout physical violation example. The physical violation loss is calculated only when the object intersection with layout (b). There is no physical constraint when the object is completely inside (a) or outside the layout (c).

We define the object estimation losses on all encoder layers in the context module, which are averaged to form the final loss:

$$\mathcal{L}_{object} = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_{obj}^{l},$$

$$\mathcal{L}_{obj}^{l} = \sum_{x \in \{\mathcal{L}_{object\_loss}\}} \beta_x * \mathcal{L}_x. \quad (7)$$

Each β_x is the loss weight corresponding to the specific object loss.

Physical Violation Loss In order to produce a physically plausible scene and regularize the relationships between objects and layout, we add physical violation loss as a part of the joint loss. As shown in Fig. 4, when the bounding box of an object intersects with the layout (i.e., walls, ceiling, or floor), the physical violation loss is calculated with the Manhattan distance to the layout. Some types of objects do intersect with the layout, such as windows and doors. So the physical constraints are only applied for categories that should never intersect with the layout. The physical violation loss is defined as:

$$\mathcal{L}_{physic} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{ins} \mathcal{L}_{3d\_violation}$$

$$\mathcal{L}_{3d\_violation} = \sum_{i=1}^{8} (relu(x_i^k - \max(X^L)) + relu(\min(X^L) - x_i^k)) \quad (8)$$

where x_i^k is corner of the kth object bounding box, X^L is a set of vertices of layout mesh. The $relu$ is applied to consider only the outside corners. $\mathbb{1}_{ins}$ has a value of 1 if the bounding box is not completely outside of the layout, and a value of 0 otherwise.

All the loss functions in joint training can be defined as :

$$\mathcal{L} = \sigma_l * \mathcal{L}_{layout} + \sigma_o * \mathcal{L}_{object} + \sigma_p * \mathcal{L}_{physic}. \quad (9)$$

3.5. Panoramic Dataset

There currently exists no publicly available realistic panoramic dataset with object shape and orientation anno-

tations. We publish **ReplicaPano**, a real-world panoramic scene understanding dataset with full ground truth. With the help of the high-fidelity textured mesh provided by Replica [44], we render photo-realistic panorama from 27 rooms diversely furnished by 3D objects. For each room, we randomly render 100 pairs of equirectangular RGB and depth images, all the images are gravity-aligned and the height of the camera center is 1.6m. Given a panorama, we utilize PanoAnnotator [58] to accurately label the room layout. Based on the colored point cloud and semantic segmentation information provided by Replica, we semi-automatically annotate the bounding box for each object in each room. Following the NYU-37 object labels [43], we select 25 categories of objects that are commonly seen in indoor scenes. Due to the lack of complete object mesh in Replica, we manually annotate posed CAD models on the 3D scene scan from multiple large-scale 3D shape datasets [4, 14, 47]. Finally, we get 2,700 photo-realistic panoramas with depth images, room layouts, 3D object bounding boxes, and object meshes. More quantitative evaluations and samples of ReplicaPano can be found in the supplementary files.

4. Experiment

In this section, we compare our model with both holistic scene understanding and single-task methods and perform ablation studies to analyze the effectiveness of the designs.

4.1. Experiment Setup

Dataset. We use two datasets in our experiments. **1) iGibson-Synthetic.** The data is synthesized by the iGibson simulator [41]. Same as the setting in DeepPanoContext [62], we use 10 scenes for training and 5 scenes for testing. **2) ReplicaPano.** To demonstrate our work's effectiveness in real-world scenes, among 27 rooms, we use 16 for training, 4 for validation, and 7 for testing.
Metrics. The results of each sub-task are evaluated with the metrics used in previous works [35, 62, 63]. Object detection is measured using mean average precision (mAP) with the threshold of 3D bounding box IoU set at 0.15. The room layout estimation error is tested by standard metrics for indoor layout reconstruction (i.e., 2D-IoU and 3D-IoU) followed by Pintore et.al [38, 45, 46]. Since the object mesh generation in our method is significantly different from other scene understanding work, we only compare the result with that of others qualitatively.
Implementation. The borrowed monocular depth estimation network (i.e., Unifuse [25]) and 3D auto-encoder network (i.e., ONet [34]) are finetuned individually on each dataset from the weights pretrained on Matterport3D and ShapeNet, respectively. The input point cloud for the object detection network is sampled to 50K by Fibonacci sampling from the estimated depth. The auto-encoder network takes

300 points from the surface of each watertight model as input and embeds each sample as a vector of size 512. In the context module, 10% of tokens are randomly masked. We trained object detection, layout estimation, and mesh generation jointly from scratch on a single NVIDIA V100 GPU. More training details are given in the supplementary files.

4.2. Comparisons with State-of-the-art Methods

Object Detection We compare our 3D object detection results with previous state-of-the-art holistic scene understanding and single-task learning methods. DeepPanoContext [62] is the only method to achieve total 3D scene understanding directly on panoramic images. Total3D [35] and IM3D [63], which work with perspective images, are extended to panorama for comparison on iGibson-Synthetic dataset. In order to show the effectiveness of lifting 2D panorama into the 3D space, we extend DeepPanoContext (named DeepPanoContext-3D) as follows: we use Point-Net++ to extract the object geometry feature from point cloud and then concatenate this feature with other appearance features to estimate 3D bounding boxes. As for the single task comparison, the point-based object detection method Group-Free [32] is chosen as baseline. The results of each method on iGibson-Synthetic are shown in Tab. 1. Since DeepPanoContext shows higher performance than Total3D and IM3D, we only compare it and its extension on ReplicaPano, results can be found in Tab. 2.

As shown in Tab. 1 and Tab. 2, our proposed method consistently outperforms both holistic understanding methods and the point-based detection baseline on most categories and the average mAP. We can see that DeepPanoContext has been significantly improved by using the recovered point cloud, which indicates the 3D methodical approach is absolutely necessary. The Tab. 2 shows our method gains better results for categories that are closely related to room layout, such as door and rug, since the transformer-based context model encourages rational spatial relationships among objects and room layout. For a few categories, such as floor lamp and chair, DeepPanoContext-3D performs better. The gap in performance for these categories is due to two factors: 1) The depth estimation model failed to recover tiny structure, for example, the pole of a floor lamp, which deteriorates the performance of our method. 2) DeepPanoContext-3D uses a finetuned 2D detector to initialize the estimation and achieve good performance for heavily occluded objects (e.g., chairs are occluded behind a table and dryers are embedded in cabinets). Improving depth quality and introducing a 2D detector into our method may help to improve the accuracy further.
Layout Estimation Previous panoramic scene understanding work does not give quantitative analysis in terms of the layout estimation, thus we only compare our method with recent state-of-the-art layout estimation methods [38, 45,

Method	chair	soft	table	fridge	sink	door	floor lamp	bottom cabinet	top cabinet	sofa chair	dryer	mAP↑
Total-Pano [35]	20.84	69.65	31.79	43.13	68.42	10.27	16.42	34.42	20.83	62.38	33.78	37.45
Im3D-Pano [63]	33.08	72.15	37.43	70.45	75.20	11.58	6.06	43.28	18.99	78.46	41.02	44.34
DeepPanoContext [62]	27.78	73.96	46.85	74.22	75.29	21.43	20.69	52.03	50.39	77.09	59.91	52.69
DeepPanoContext-3D	**39.41**	78.03	51.44	75.24	81.46	51.97	**60.01**	55.56	42.58	79.99	**60.07**	61.43
Group-Free [32]	27.83	96.04	61.57	**84.69**	87.69	82.20	27.20	56.46	77.99	79.21	8.29	62.65
Ours	38.47	**98.15**	**66.61**	82.77	**89.55**	**87.49**	40.31	**59.53**	**80.71**	**83.42**	13.83	**67.35**

Table 1. Comparisons of object detection on iGibson-Synthetic with state-of-the-art. We use mean average precisions with 3D IoU threshold 0.15 and evaluate 11 common object categories following [35, 62, 63]. DeepPanoContext-3D is the extended version of Deep-PanoContext [62] with point cloud.

Method	cabinet	door	chair	curtain	lamp	rug	sofa	table	trash	tv	mAP↑
DeepPanoContext [62]	35.33	6.78	47.04	13.6	12.15	4.49	26.87	73.34	39.59	4.86	26.41
DeepPanoContext-3D	52.49	11.42	**70.39**	32.38	20.02	9.10	30.13	**82.24**	**63.22**	12.19	38.36
Group-Free [32]	59.56	42.21	52.83	**34.07**	19.65	32.90	80.59	51.47	44.64	**52.76**	47.07
Ours	**63.69**	**46.74**	54.02	30.41	**20.04**	**48.53**	**80.96**	46.42	51.53	47.82	**49.02**

Table 2. Comparisons of object detection on ReplicPano.

Method	iGibson-Synthetic		ReplicaPano	
	2D-IoU↑	3D-IoU↑	2D-IoU↑	3D-IoU↑
HorizonNet [45]	89.22	89.18	84.56	83.59
HoHoNet [46]	90.13	89.97	84.76	84.05
Led2Net [53]	90.39	90.30	84.62	83.91
Deep3dLayout [38]	90.65	90.40	84.87	83.50
Ours	**92.24**	**92.04**	**85.98**	**84.58**

Table 3. Comparisons of layout estimation on iGibson-Synthetic and ReplicaPano. Evaluation metrics include 2D and 3D intersection-over-union (IoU) following [38, 45, 46].

Method	depth metric		detection metric
	Abs.Rel.↓	RMSE↓	mAP↑
Panoformer-pretrain [42]	0.0774	0.2105	54.77
Unifuse-finetune [25]	0.0328	0.1107	67.35
Panoformer-finetune [42]	0.0214	0.0997	69.09
GT-depth	-	-	79.46

Table 4. The impact of depth accuracy. Evaluation metrics include absolute relative error (Abs. Rel.) and root mean square error (RMSE) for depth and mAP for object detection. Panoformer-pretrain is pre-trained on Matterport3D [3], while *-finetune means the depth estimator gets finetuned on iGibson-Synthetic.

46, 53]. As shown in Tab. 3, our method achieves the best performance among other baselines, indicating joint training with the context model helps to improve the layout estimation from a single panorama.

Holistic Scene Reconstruction Qualitative comparison with DeepPanoContext and DeepPanoContext-3D are demonstrated in Fig. 5, our method obtains the best indoor scene reconstruction, including the object pose, room layout, and object shape reconstruction.

4.3. Ablation Study

In this section, we conduct some ablation studies on iGibson-Synthetic to clarify the importance of each component in our method.

Impact of depth quality We first investigate how the accuracy of the depth map impacts the final 3D object detection. Two depth estimation networks, Unifuse [25] and PanoFormer [42] are involved in Tab. 4, which reveal that object detection results benefit from higher depth quality. In addition, we observe that even if the proposed method uses

a depth estimator without finetuning, the performance still slightly outpasses that of DeepPanoContext (Tab. 1), which employed a 2D fine-tuned detector for initialization.

Effect of architecture and loss To figure out the effect of each module, we provide detailed ablation experiments in terms of object detection and layout estimation. The results are summarized in Tab. 5. The first 2 rows show the room layout benefit from perspective features. The third row indicates that introducing joint training and physical violation loss consistently improves the results of object detection and layout estimation. As for the fourth and fifth rows, we can conclude that our method can generate better representation and relationships among objects and the room layout, with the help of global image tokens and the token masking strategy, thus obtaining better results on each task.

5. Conclusion

In this paper, we propose a fully 3D method for end-to-end indoor scene understanding from a single RGB panorama
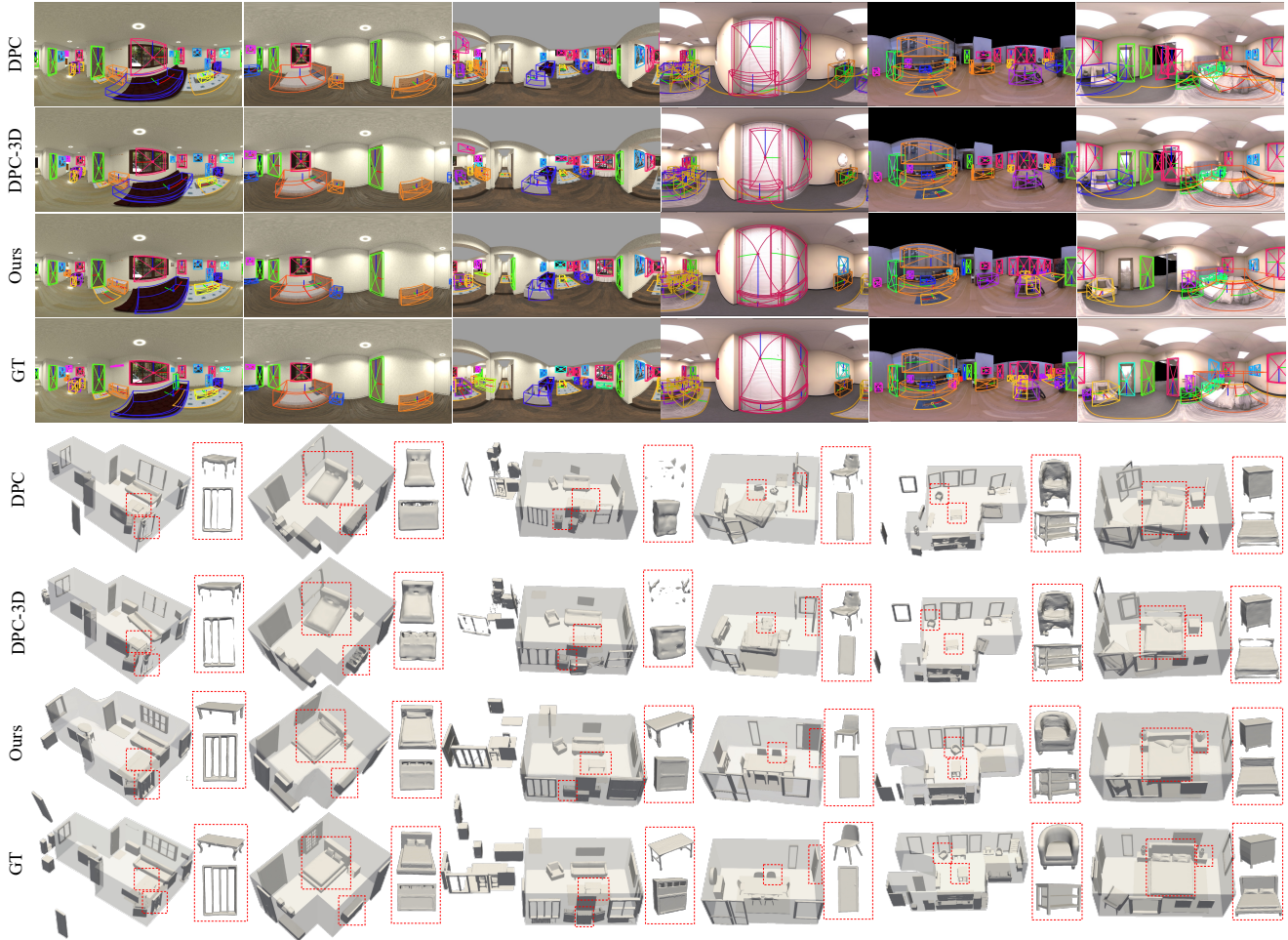
Figure 5. Qualitative comparisons on 3D object detection and scene reconstruction. In the top four rows, we compare our object detection results with DeepPanoContext (DPC), DeepPanoContext-3D (DPC-3D), and ground truth in the panoramic view. The color of the bounding boxes represents their categories. The bottom four rows show the scene reconstruction results, with two magnified object reconstruction results presented on the right-hand side. Note that the first three columns are the results on iGibson-Synthetic, and the last three columns are the results on ReplicaPano.

Perspective Feature	Joint Training	Physical Violation Loss	Token Masking	Global Image Token	mAP↑ (11 categories)	mAP↑ (57 categories)	2D-IoU↑	3D-IoU↑
✗	✗	✗	✗	✗	62.59	40.44	90.65	90.40
✓	✗	✗	✗	✗	-	-	90.98	90.73
✓	✓	✓	✗	✗	65.68	41.50	91.41	91.20
✓	✓	✓	✓	✗	66.27	42.22	92.14	91.78
✓	✓	✓	✗	✓	66.78	41.97	91.77	91.56
✓	✓	✓	✓	✓	**67.35**	**43.55**	**92.24**	**92.04**

Table 5. The ablation studies on iGibson-Synthetic dataset, demonstrates how our proposed designs improve the accuracy on object detection and layout estimation. We show in the last row the full architecture setup.

image. To better learn the context information in the panorama, we use a Transformer-based context module to learn the relationship between objects and room layout. In addition, we introduce a new dataset for real-world panoramic holistic scene understanding. Experiments demonstrate that our method achieves state-of-the-art performance on both synthetic and real-world datasets.

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, 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 3, 7

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4, 6

[5] Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Pq-transformer: Jointly parsing 3d objects and layouts from point clouds. *IEEE Robotics and Automation Letters*, 7(2): 2519–2526, 2022. 3

[6] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 2

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[9] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 616–624, 2016. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[11] Mingyue Dong, Linxi Huan, Hanjiang Xiong, Shuhan Shen, and Xianwei Zheng. Shape anchor guided holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21916–21926, 2023. 3

[12] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2

[13] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[14] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 6

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[17] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *2009 IEEE 12th international conference on computer vision*, pages 1849–1856. IEEE, 2009. 2

[18] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 3

[19] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 2

[20] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 187–203, 2018. 2

[21] Moos Hueting, Pradyumna Reddy, Vladimir Kim, Ersin Yumer, Nathan Carr, and Niloy Mitra. Seethrough: finding chairs in heavily occluded indoor scene images. *arXiv preprint arXiv:1710.10473*, 2017. 2

[22] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10632–10640. IEEE, 2022. 4

[23] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 275–292. Springer, 2022. 4

[24] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5134–5143, 2017. 2

[25] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama

depth estimation. *IEEE Robotics and Automation Letters*, 6 (2):1519–1526, 2021. 6, 7

[26] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2136–2143. IEEE, 2009. 2

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 4

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 4

[29] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 429–446. Springer, 2022. 2

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[31] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[32] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 3, 4, 5, 6, 7

[33] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 936–944, 2015. 2

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2, 4, 6

[35] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 2, 6, 7

[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[37] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 2

[38] Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. Deep3dlayout: 3d reconstruction of an indoor layout from a spherical panoramic image. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 2, 3, 4, 5, 6, 7

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[40] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 3

[41] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021. 3, 6

[42] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 {\deg} depth estimation. *arXiv preprint arXiv:2203.09283*, 2022. 7

[43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 6

[44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 3, 6

[45] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 2, 6, 7

[46] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 2, 6, 7

[47] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[49] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018. 2

[50] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 302–310, 2018. 1

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[52] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360 {\deg} videos. *arXiv preprint arXiv:1811.05304*, 2018. 2, 3

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

[54] Yikai Wang, TengQi Ye, Lele Cao, Wenbing Huang, Fuchun Sun, Fengxiang He, and Dacheng Tao. Bridged transformer for vision and point cloud 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12114–12123, 2022. 3

[55] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 2

[56] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012. 2, 3

[57] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022. 2

[58] Shang-Ta Yang, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. Panoannotator: A semi-automatic tool for indoor panorama layout annotation. In *SIGGRAPH Asia 2018 posters*, pages 1–2. 2018. 6

[59] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3363–3372, 2019. 2, 3, 4

[60] Haozheng Yu, Lu He, Bing Jian, Weiwei Feng, and Shan Liu. Panelnet: Understanding 360 indoor environment via panel representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 878–887, 2023. 2

[61] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021. 2

[62] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021. 1, 2, 3, 4, 6, 7

[63] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 1, 2, 6, 7

[64] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 668–686. Springer, 2014. 1, 2, 3

[65] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2, 3

[66] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 2, 3

[67] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2051–2059, 2018. 2