Holistic Pose Graph: Modeling Geometric Structure among Objects in a Scene using Graph Inference for 3D Object Prediction

Jiwei Xiao[1,2], Ruiping Wang[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
jiwei.xiao@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

Abstract

*Due to the missing depth cues, it is essentially ambiguous to detect 3D objects from a single RGB image. Existing methods predict the 3D pose for each object independently or merely by combining local relationships within limited surroundings, but rarely explore the inherent geometric relationships from a global perspective. To address this issue, we argue that modeling geometric structure among objects in a scene is very crucial, and thus elaborately devise the **H**olistic **P**ose **G**raph (HPG) that explicitly integrates all geometric poses including the object pose treated as nodes and the relative pose treated as edges. The inference of the HPG uses GRU to encode the pose features from their corresponding regions in a single RGB image, and passes messages along the graph structure iteratively to improve the predicted poses. To further enhance the correspondence between the object pose and the relative pose, we propose a novel consistency loss to explicitly measure the deviations between them. Finally, we apply **H**olistic **P**ose **E**stimation (HPE) to jointly evaluate both the independent object pose and the relative pose. Our experiments on the SUN RGB-D dataset demonstrate that the proposed method provides a significant improvement on 3D object prediction.*

1. Introduction

3D object prediction from a single RGB image is extremely challenging, which estimates 3D bounding boxes for each object in a scene. The main difficulty of this task is to predict the depth information that a single RGB image loses during the projection from 3D real world to the 2D image. How are humans capable of making precise estimations when looking at an image? Humans not only have rich prior knowledge about the category-specific object, but also can leverage the geometric relationships among different objects in a scene to alleviate the uncertainties of prediction for each object. Existing methods [12,14,16,17,20,37] consider using the prior knowledge to reason about the object pose independently but it is rough sometimes. As a result of the inevitable deviation of prediction per object, there will be a certain amount of inaccurate estimations, as shown in Figure 1 (a), which cause humans' misunderstanding about the scene. These results mainly are caused by the improper relative pose estimation and could have been effectively avoided by using the geometric relationships.

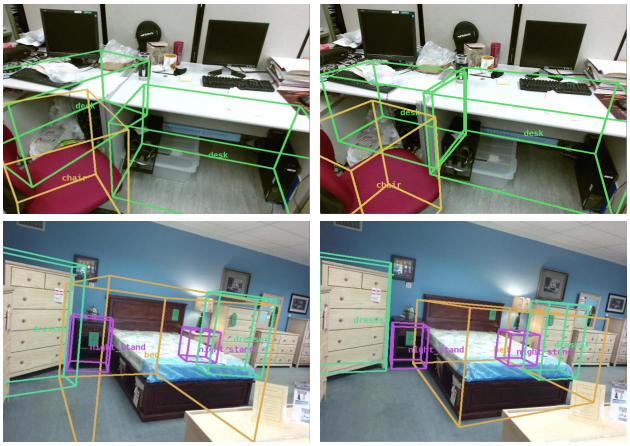

Figure 1. **The effect of using geometric relationships**. (a) some results from our baseline method [27], which make inaccurate estimation about the relative pose. (b) our model makes more reasonable prediction after using HPG.

Inspired by the above observations, we seek to leverage the geometric relationship to add more constraints on each object for more reasonable and precise estimation. Modeling the geometric relationships explicitly will help us exclude many impossible solutions in the 3D space. For example, as shown in top of Figure 1 (a), we should have gotten

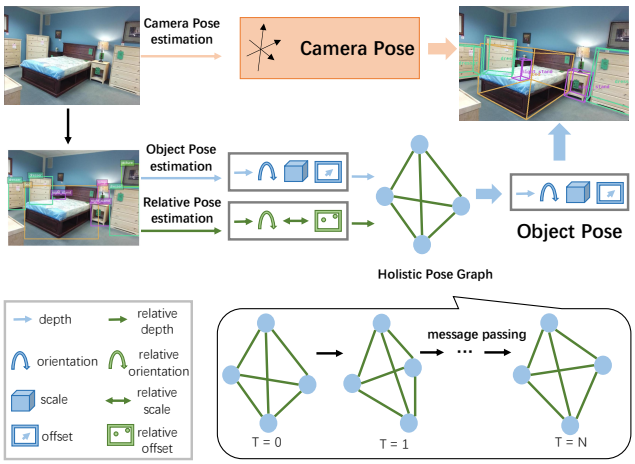

Figure 2. **Overview of our proposed method**. We firstly estimate the 3D camera pose, object pose and relative pose, according to the whole image and 2D detections. As shown above, the blue nodes define the object pose and the green edges define the relative pose. We build the Holistic Pose Graph and pass messages along this graph topology. Combining the camera pose and the final object pose, we can get all the 3D bounding boxes in the image on the basis of the camera system.

the cues that two desks in the image have the same orientation to avoid the wrong estimation. Moreover, we try to utilize the holistic geometric relationships rather than only consider pair-wise relationships, in order to better maintain the consistency between the 2D image and the 3D real world. Consequently, we devise the Holistic Pose Graph (HPG) to efficiently model geometric structure among objects in a scene. As shown in Figure 1 (b), more reasonable predictions are obtained after using HPG.

Figure 2 shows the overview of our method. We firstly estimate the camera pose from the global features of the whole image. After detecting the 2D objects, we crop the corresponding regions of the image to extract the object pose features. Specifically, the relative pose features are composed of the related object pose features, the visual features of the union region in the image, and the 2D relative geometric features computed from the coordinates of 2D bounding boxes. We use these features to initialize the graph nodes that define the object pose and the edges that define the relative pose, and further build the HPG. By passing messages along this graph topology iteratively, both nodes and edges integrate holistic geometric information and update their state simultaneously. Finally, we combine the camera pose and the final object pose to parametrize the 3D bounding boxes. To further estimate reasonable scene layout, we propose the consistency loss based on the inherent constraints between the object pose and the relative pose, which can enhance the correspondence between them.

To evaluate the performance of 3D object prediction, existing metrics only treat each prediction independently by calculating the 3D Intersection over Union (IoU) between the predicted 3D bounding box and ground truth, but fail to treat all objects in the image as a whole. Under these metrics, the results in the top row of Figure 1 (a) and (b) can not be distinguished because all the objects' IoU are above the preset threshold. However, the results of Figure 1 (b) are indeed more rational in human cognition. Consequently, we introduce Holistic Pose Estimation (HPE) that more holistically evaluates both the 3D bounding box of the independent object and the relative pose of each pair of objects.

We evaluate our model [1] on SUN RGB-D dataset [32] to verify the effectiveness of the Holistic Pose Graph. Results show that our proposed method outperforms previous methods both on the existing metrics and HPE.

2. Related Work

3D object prediction from a single RGB image is extremely challenging. Early works [2, 22, 39–41] based on scene geometry utilize the 3D world prior and scene grammar to estimate 3D holistic scene. In recent years, many learning-based methods focus on room layout estimation [3, 26, 30, 42] and object pose estimation [4, 14–16, 27]. These methods combine the category-specific prior knowledge and the visual appearance feature of the independent object in a single RGB image to predict the 3D bounding boxes. To address the 2D-3D ambiguity, state-of-the-art methods propose various techniques to improve the performance of 3D object prediction. Huang *et al*. [15] devise a new parametrization of 3D bounding boxes to enforce the 2D-3D consistency. Huang *et al*. [14] propose an intermediate representation to bridge the 2D-3D gap. While considerable improvements have been achieved, such methods generally make independent predictions per object and ignore the importance of geometric relationships.

Modeling 2D relationships can assist us in better understanding the attributes of objects and the holistic scene. The exploitation of relationships has demonstrated benefits for a bunch of computer vision tasks both in 2D and 3D. 2D relationships mainly include semantic relationships and geometric relationships. For 2D object detection, [13, 24] use semantic relationships to improve the performance on classification tasks. [35, 36, 38] similarly extract semantic relationship features for the task of scene graph generation. For 3D tasks, [5–7, 9, 10, 18, 21, 25] pursue reasoning about geometric relationships. Focusing on the 3D prediction task relevant to this study, Nie *et al*. [27] implicitly encode multilateral relation in each object's surroundings. Kulkarni *et al*. [20] consider the pairwise relations to train the object pose module and the relation module respectively. Different from the above methods, our approach explicitly encodes

[1] Our source codes are available at http://vipl.ict.ac.cn/resources/codes.

the holistic geometric relationships, and models geometric structure for graph inference to enhance the correspondence between the objects and relationships.

Existing metrics on 3D object prediction from a single RGB image mainly come from the SUN RGB-D benchmark [32], with following adjustment of the 3D IoU threshold from 0.25 to 0.15 in Huang *et al*. [15], taking account of the challenge of missing depth information in RGB-based input. [15] further develops 3D Box Estimation to reflect the ability of mapping 2D ground truth bounding boxes to 3D bounding boxes by excluding the influences of the 2D detector. In order to evaluate the performance of both the independent object and the relative pose of each pair of objects, we introduce Holistic Pose Estimation based on the graph structure to compute the accuracy for $\langle$ subject pose, relative pose, object pose $\rangle$ triplets.

3. Approach

In this work, we devise the Holistic Pose Graph for modeling geometric structure among objects in a scene to reason about the object pose and relative pose. We illustrate our framework in Figure 3. The details are introduced in the following sections.

3.1. Holistic Pose Graph

Graph representation. $G = (P_V, P_E)$ represents all geometric poses of the input image. The node $P_v \in P_V$ defines the object pose, and the edge $P_e \in P_E$ defines the relative pose. We parameterize the object pose same as the prior work [15] and the relative pose with similar description. The object pose P_v is described by four parameters (δ, d, ϕ, s). The $\delta \in \mathbb{R}^2$ defines the offset between the 3D projected center and the center of the 2D bounding box, $d \in \mathbb{R}$ is the distance between the camera center and 3D object center, $\phi \in \mathbb{R}^{3\times3}$ represents the orientation of 3D bounding box, and $s \in \mathbb{R}^3$ is the length, width, and height of 3D bounding box. We build the center of the world system located at the camera center with its y-axis perpendicular to the floor and its x-axis toward the camera following [27]. Thus, we can use the pitch and roll angles (α, β) to represent the camera extrinsic parameters $R(\alpha, \beta) \in \mathbb{R}^{3\times3}$. We can formulate the 3D bounding box through the camera intrinsic parameters $K \in \mathbb{R}^{3\times3}$, camera extrinsic parameters $R(\alpha, \beta) \in \mathbb{R}^{3\times3}$, the center of 2D bounding box $c \in \mathbb{R}^2$ and the corresponding object pose. Firstly, the center of 3D bounding box $C \in \mathbb{R}^3$ can be computed as

$$C = dR(\alpha, \beta)^{-1} \frac{K^{-1}[c + \delta, 1]^T}{\|K^{-1}[c + \delta, 1]^T\|_2}. \quad (1)$$

Combining C, ϕ and s can then decide a 3D bounding box $B \in \mathbb{R}^{3\times8}$ in the world system. The relative pose P_e has a similar parametrization $(\delta_{ij}, d_{ij}, \phi_{ij}, s_{ij})$ with P_v, which represents the relative pose between subject i and object j.

Graph initialization. We utilize the 2D detection results to crop the detected regions, and then use ResNet-34 to extract the object pose features P_v. We concatenate each pair of object pose features, the 2D relative geometric features g, and the corresponding union region's visual features u encoded by ResNet-34 to represent the relative pose features P_e^{ij}, which can be formulated as $[P_v^i, P_v^j, g, u]$. The g is computed from the 2D bounding box of subject i and object j as in [19]. Specifically, x_i, y_i, w_i, h_i define a 2D bounding box of object i, B_I, B_U define the intersection region and the union region respectively, and $f()$ is an FC layer to increase the dimension of the 2D relative geometric features.

$$g = f([\frac{x_i - x_j}{\sqrt{w_j h_j}}, \frac{y_i - y_j}{\sqrt{w_j h_j}}, \sqrt{\frac{w_i h_i}{w_j h_j}}, \frac{w_i}{w_j}, \frac{h_i}{h_j}, \frac{B_I}{B_U}]). \quad (2)$$

3.2. Message Passing

We build the HPG as a fully connected graph that models all geometric relationships between each pair of objects. GRU [1] is employed to encode features effectively for its high efficiency as advocated in [24, 35]. To integrate the holistic geometric information, we pass messages along the graph structure to iteratively update the node GRUs and edge GRUs states. The state of the two kinds of GRUs is initialized by the object pose features P_v and the relative pose features P_e, respectively. Each node and edge in HPG maintain its state in its corresponding GRU unit, where all nodes share the same GRU weights, and all edges share the other set of GRU weights. The illustration of the message passing mechanism is shown in Figure 4. During each message passing iteration, each node updates its state with the messages from all related edges, and each edge updates its state with the messages from its subject node and object node. As in [35], we denote the state of node GRU and edge GRU as h_i, h_{ij}, the i-th node message as m_i, and the edge message from the i-th node to j-th node as m_{ij}. Specifically, the m_i and m_{ij} are formulated as:

$$m_i = \sum_{j:i\to j} \sigma(w_{out}[h_i, h_{ij}])h_{ij} + \sum_{j:j\to i} \sigma(w_{in}[h_i, h_{ji}])h_{ji}, \quad (3)$$

$$m_{ij} = \sigma(w_{sub}[h_i, h_{ij}])h_i + \sigma(w_{obj}[h_j, h_{ij}])h_j. \quad (4)$$

In above equations, $[,]$ is a concatenation operation, and σ represents a sigmoid function. $w_{out}, w_{in}, w_{sub}, w_{obj}$ are learnable weights. After the process of message passing, we use four FC layers to individually predict the four parameters (δ, d, ϕ, s) of object pose P_v. Similarly, we also use the other four FC layers to get the four parameters $(\delta_{ij}, d_{ij}, \phi_{ij}, s_{ij})$ of relative pose P_e.

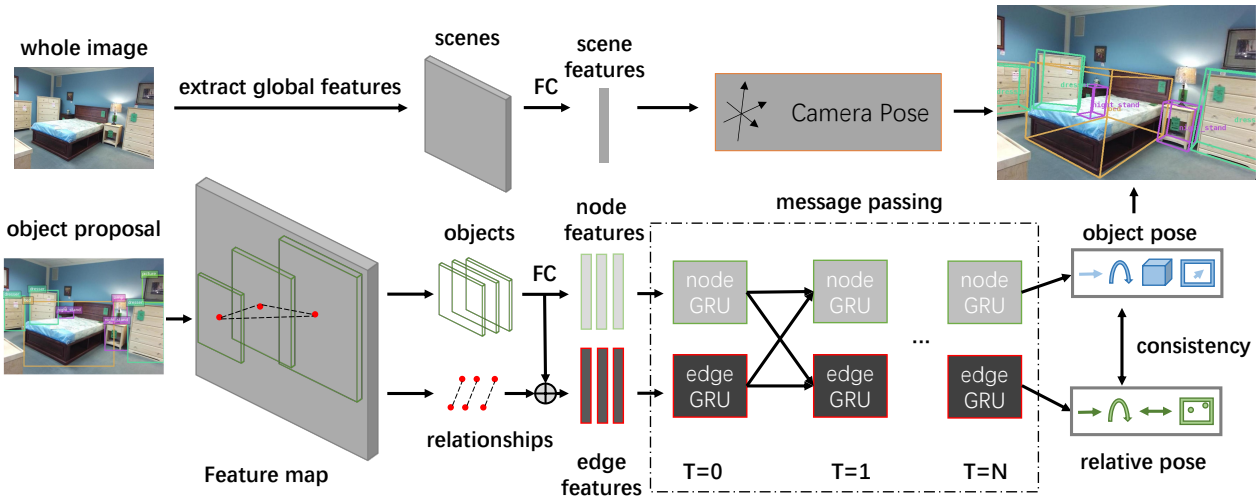

Figure 3. **HPG: The framework of our method**. There are two branches of our network to predict the camera pose and each object pose respectively. Firstly we adopt the ResNet-34 [11] architecture to extract the global features for estimating the camera pose. According to the results of 2D object detection, we crop the detected object proposals to extract object features initializing node GRU, and then concatenate each pair of object features and the union region's visual features to extract relationship features initializing edge GRU for building HPG. With HPG, we then pass messages along the graph structure and update the state of each node and edge iteratively. By imposing direct supervisions on the output object pose and relative pose respectively, a consistency loss is further proposed to build correspondence between the two kinds of poses. Finally, we combine the camera pose and each object pose to estimate the 3D bounding boxes in the scene.

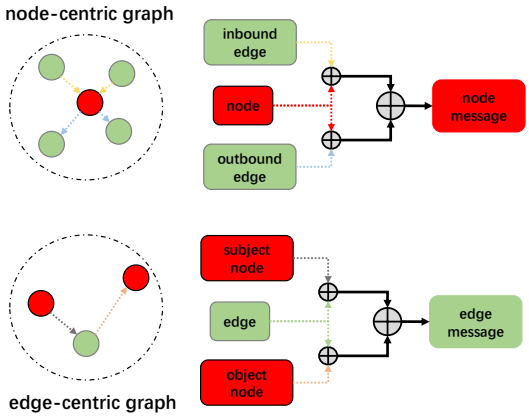

Figure 4. **An illustration of the Message Passing mechanism**. The node message is from the inbound and outbound edge GRUs and its own state. The edge message is from the subject and object node GRUs and its own state.

3.3. Loss Functions

As shown in Figure 3, HPG finally outputs the object pose and relative pose, which we directly use their ground truth to supervise. Besides, we further propose the consistency loss for imposing supervisions on the two pose prediction modules to maintain their correspondence. Consequently, we define three loss functions as L_{obj} (object pose), L_{rel} (relative pose), L_{con} (consistency). As noted in [15, 28], directly regressing absolute angles or depth is error-prone, which magnifies the variance of predicted variables. Thus, we set the learning way of $\{\phi, d\}$ as a combination of classification and regression, but directly regress $\{\delta, s\}$ because of their inherent low variance. L_{obj} and L_{rel} are defined as:

$$L_{obj} = \sum_{x\in\{\delta,s\}} \lambda_x^{reg} L_x^{reg} + \sum_{y\in\{\phi,d\}} (\lambda_y^{reg} L_y^{reg} + \lambda_y^{cls} L_y^{cls}), \tag{5}$$

$$L_{rel} = \sum_{x\in\{\delta_{ij},s_{ij}\}} \lambda_x^{reg} L_x^{reg} + \sum_{y\in\{\phi_{ij},d_{ij}\}} (\lambda_y^{reg} L_y^{reg} + \lambda_y^{cls} L_y^{cls}). \tag{6}$$

L_*^{cls} is softmax loss function and L_*^{reg} is smooth-L1 loss function. λ_*^* are the weights for corresponding loss functions. Due to the two kinds of predicted pose in the same scene, there is inherently correspondence between them. For a pair of object pose $(P_v^i,\ P_v^j)$ and their relative pose P_e^{ij}, they are theoretically equivalent. To explicitly measure the consistency between $(P_v^i,\ P_v^j)$ and P_e^{ij}, we formalize P_e^{ij*} using its related $(P_v^i,\ P_v^j)$, represented as $(\delta_{ij}^*, d_{ij}^*, \phi_{ij}^*, s_{ij}^*) = (\delta_j - \delta_i, d_j - d_i, \phi_j - \phi_i, s_j/s_i)$. For L_{con}, we denote $h(,)$ as its function which calculates the deviation of P_e^* and P_e in the same way as the former loss fuction Equ. 6:

$$L_{con} = h(P_e^*, P_e). \tag{7}$$

Finally, we also adopt the cooperative loss L_{co} from [15] and the global loss L_g from [27], which add physical constraints and enhance the consistency. In summary, the loss functions for the whole network training can be written as:

$$L = \lambda_{obj} L_{obj} + \lambda_{rel} L_{rel} + \lambda_{con} L_{con} + \lambda_{co} L_{co} + \lambda_g L_g. \tag{8}$$

λ_* are the weights of above five loss functions.

4. Holistic Pose Estimation

We propose HPE for evaluating the performance of both the object pose and the relative pose. We evaluate the 3D bounding box for object pose by computing the IoU between predicted boxes and ground truth boxes, and further design the geometric relationships similarity S to evaluate the relative pose. S can be denoted as $\{S^{Loc_\theta}, S^{Loc_l}, S^{Ori}\}$. Specifically, given a pair of 3D boxes for subject i and object j, we denote their relative location as $Loc_{ij} \in \mathbb{R}^3$, which is the difference vector computed from the centers of the two 3D boxes, and denote their relative orientation as $Ori_{ij} \in \mathbb{R}$, which represents the absolute deviation of rotation angle around the y-axis between the two boxes.

S^{Loc_θ} measures the angle similarity between the predicted relative location Loc_{ij}^{pre} and ground truth relative location Loc_{ij}^{GT} as formulated:

$$S_{ij}^{Loc_\theta} = \frac{Loc_{ij}^{pre} \times Loc_{ij}^{GT}}{|Loc_{ij}^{pre}| \times |Loc_{ij}^{GT}|}. \tag{9}$$

$S_{ij}^{Loc_l}$ measures the length similarity between the two relative locations as:

$$S_{ij}^{Loc_l} = \left| \frac{|Loc_{ij}^{pre}| - |Loc_{ij}^{GT}|}{|Loc_{ij}^{GT}|} \right|. \tag{10}$$

S_{ij}^{Ori} describes the similarity between the predicted relative orientation Ori_{ij}^{pre} and ground truth relative orientation Ori_{ij}^{GT} as:

$$S_{ij}^{Ori} = |Ori_{ij}^{pre} - Ori_{ij}^{GT}|. \tag{11}$$

As shown in Figure 5, (a) denotes ground truth, (b) and (c) are two supposed predictions. For human evaluation, (b) is obviously better than (c), but prior metrics only consider the accuracy of each object box and can not distinguish the two predictions. Under HPE, we can distinguish the above two predictions by further computing the accuracy of each geometric relationship. Vividly, we not only limit the predicted boxes on the surrounding region of ground truth, but also add the constraint on the relative pose just like connecting each pair of objects with a flexible rod to maintain the holistic scene layout.

Specifically, we set the relative threshold ε_{rel} as $\{0.5, 0.5, 30°\}$ in the experiments, corresponding to $\{\varepsilon^{Loc_\theta}, \varepsilon^{Loc_l}, \varepsilon^{Ori}\}$ individually. We regard the predicted geometric relationship as true positive when $\{S^{Loc_\theta} \geq \varepsilon^{Loc_\theta}, S^{Loc_l} \leq \varepsilon^{Loc_l}, S^{Ori} \leq \varepsilon^{Ori}\}$. The computation of HPE can be formulated as:

$$Acc = \frac{\sum_{i,j \in P_V} C(P_v^i, P_v^j)}{N}. \tag{12}$$

where $C(,)$ is a discriminant operation, and hence $C(P_v^i, P_v^j)$ describes whether the P_e^{ij} is true. We design

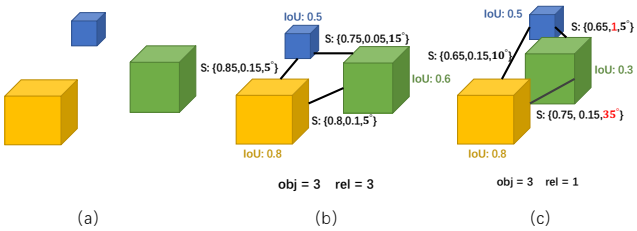

Figure 5. **An illustration of HPE**. (a) shows the 3D ground truth boxes. (b) and (c) are two supposed prediction results. We show the IoU and the similarity for each box or edge. At the bottom of each prediction, we show the number of true predicted objects and relationships.

two types of $C(,)$ for different aims. One is only considering the geometric relationships similarity, and the other further demands the IoU of subject and object both meeting their threshold. For computing Acc among the whole test set, N denotes the number of P_E. Moreover, we can also compute Acc based on each image. Specifically, for counting based on each image, we think it is a true image prediction if the number of true positives discriminated by $C(,)$ reaches half of the number of objects in the image.

5. Experiments

5.1. Experimental Setup

Datasets. We train our model and compare it with other methods [14, 15, 27] on the SUN RGB-D dataset [32] including 5285 training images and 5050 testing images. As [27] we use the same train/test split and the object labels provided in NYU-37 [31] for fair comparison.

Metric. We evaluate 3D object detection by the average precision (AP) on NYU-37 object categories. Similar to 3D object detection, we evaluate 3D box estimation using 2D ground truth boxes as input only to test the ability of mapping 2D to 3D. Besides, we further compare our method with prior works on our developed task of Holistic Pose Estimation (HPE). For all such metrics, we set the threshold of 3D IoU to 0.15 as in [15].

Implementation. We train our 2D detector [29] on the COCO dataset [23] and fine-tune it on SUN RGB-D [32]. The backbone of the image feature extractor for the whole image and object proposal both are ResNet-34 [11]. We jointly train our camera pose module and object pose module with the ground truth of 2D bounding boxes. The annotations of the relative pose are from the ground truth of the object pose, which is calculated same as P_e^* mentioned in Section 3.3. The settings about learning weights of loss functions are introduced in the supplementary materials.

5.2. 3D Box Estimation

The performance of 3D object detection is determined by both 2D object detection and 3D bounding box estima-

Table 1. Comparisons of 3D object detection on SUN RGB-D dataset. The results of [15, 27] are cited from [27], which are trained on NYU-37 object labels.

Method	bed	chair	sofa	table	desk	dresser	nightstand	sink	cabinet	lamp	mAP
CooP [15]	57.71	15.21	36.67	31.16	19.90	15.98	11.36	15.95	10.47	3.28	21.77
Total3D [27]	60.65	17.55	44.90	36.48	27.93	21.19	**17.01**	18.50	14.51	5.04	26.38
Ours (w/o. HPG)	60.52	21.28	52.54	35.71	31.90	23.00	12.22	16.86	14.67	5.24	27.39
Ours	**67.07**	**30.55**	**56.63**	**44.51**	**37.82**	**23.40**	16.93	**25.70**	**17.73**	**7.15**	**32.75**

Table 2. Comparisons of 3D object detection on SUN RGB-D dataset. The results of [14, 16] are cited from their original papers, which are trained with fewer object categories. We provide the results of our method and [27] on the common categories for comparison.

Method	bed	chair	sofa	table	desk	toilet	sink	shelf	lamp	mAP
HoPR [16]	58.29	13.56	28.37	12.12	4.79	16.50	2.18	1.29	2.41	15.50
PerspectiveNet [14]	**71.39**	**34.94**	55.63	34.10	14.23	**73.73**	**34.41**	4.21	**9.54**	36.91
Total3D [27]	60.65	17.55	44.90	36.48	27.93	44.24	18.50	4.93	5.04	28.91
Ours	67.07	30.55	**56.63**	**44.51**	**37.82**	60.97	25.70	**13.32**	7.15	**38.19**

tion. To more directly compare our proposed method with existing methods, focusing on estimating the ability of mapping 2D image patches to 3D bounding boxes, we use 2D ground truth boxes as input following [15]. Based on using mIoU for evaluation, we further compute the accuracy (Acc) by setting the IoU threshold for the true positive. The comparisons using both mIoU and Acc are reported in Table 4, which are trained on SUN RGB-D dataset with NYU-37 object labels. The results show that our model exactly improves the ability of mapping the 2D plane to 3D real world.

5.3. 3D Object Detection

We compare our method with state-of-the-art methods [14–16, 27] using the same metric as [27], where the mean average precision (mAP) is computed with 3D bounding box IoU. The comparison is shown in Table 1, where (w/o. HPG) denotes our full model without HPG. The results demonstrate the advantage of our method over state-of-the-arts and the effectiveness of HPG. Besides, since [14, 16] have used different categories, for fair comparison, we list the common categories in Table 2. PerspectiveNet [14] proposes a more effect 2D-3D mapping way than Total3D. The 2D-3D mapping way of ours is similar to Total3D, but the key to our superior performance is to better utilize geometric relationships. Specifically, our HPG can be embedded in other 3D object detection networks with proper modification (such as CooP). More experiments results are listed in the supplementary materials.

Existing 3D reconstruction methods [20, 34] also predict object pose. We compare our method with them by training our model on NYU v2 dataset [31] with six object categories and using the same metric referring to [20]. The results are reported in Table 3. The prior works [20, 34] pretrain their model on SUNCG [33] dataset with 3D model supervision, but our performance is close to them without these extra supervisions. For fair comparison, we use the NYU v2 dataset and the annotations from SUN RGB-D dataset [32] to retrain 3D-RelNet [20] and Total3D [27] without 3D model supervisions and any pretrained models. To better distinguish our reproduced results and the corresponding results reported in their original papers, we add "*" for our reproduction. Our method significantly outperforms 3D-RelNet [20]* benefited from the utilization of the camera system. Compared with the reproduced [27]* [2], we also reach higher performance on object pose prediction.

5.4. Holistic Pose Estimation

In Section 4, we introduce the details of HPE. As mentioned there, the $C(,)$ and N have two settings respectively for different levels evaluation. Consequently, we summarize four metrics for comparison.

- **Relative Pose Accuracy** ($RelAcc$) : The accuracy of relative pose prediction only considering the geometric relationships similarity.
- **Phrase Accuracy** ($PhrAcc$) : The accuracy of phrase prediction including the geometric relationships similarity and the IoU of both subject and object.
- **Relative Pose Accuracy based on image** ($RelAcc_I$) : The accuracy of relative pose prediction with $RelAcc$ computed based on image.
- **Phrase Accuracy based on image** ($PhrAcc_I$) : The accuracy of phrase prediction with $PhrAcc$ computed based on image.

In the experiments, we use 2D ground truth boxes as input. The thresholds of S^{Loc_θ}, S^{Loc_l} and S^{Ori} are 0.5, 0.5

[2] Note that the original Total3D implementation, as reported in Tab. 3 ("Total3D [27]"), has used the annotations from [8], which differs from our reproduction "Total3D [27]*".

Figure 6. The results of 3D object detections on SUN RGB-D dataset. Each group of qualitative result contains three columns which are the prediction of our baseline (Total3D [27]), our proposed method, and ground truth respectively.

Table 3. Comparisons of object pose estimation with the existing methods on NYUv2 dataset. Specifically, 3D-RelNet [20]* and Total3D [27]* are our reproduced results with the same setting as ours for a fair comparison .

Method	3D Model Supervision	Translation(meters) Median (lower is better)	Mean	(Err$\leq 0.5m$) % (higher is better)	Rotation(degrees) Median (lower is better)	Mean	(Err$\leq 30°$) % (higher is better)	Scale Median (lower is better)	Mean	(Err≤ 0.2) % (higher is better)
Factored 3D [34]	yes	0.49	0.62	51.0	14.6	42.6	63.8	0.37	0.40	18.9
3D-RelNet [20]	yes	0.41	0.54	60.9	14.0	39.6	67.0	0.33	0.38	21.7
Total3D [27]	no	0.48	0.61	51.8	14.4	43.7	66.5	0.22	0.26	43.7
3D-RelNet [20]*	no	0.50	0.67	50.2	21.8	53.4	53.3	0.38	0.42	16.0
Total3D [27]*	no	0.51	0.69	48.9	15.7	37.0	66.9	0.27	0.32	32.8
Ours	no	**0.43**	**0.57**	**57.7**	**14.3**	**36.7**	**68.0**	**0.25**	**0.30**	**37.0**

Table 4. Comparisons of 3D box estimation on SUN RGB-D dataset. For each column, the left and right results denote the Acc and mIoU individually.

Method	Acc/mIoU	bed	chair	sofa	table	desk	dresser	nightstand	sink	cabinet	lamp
CooP [15]	37.77/13.70	74.82/28.6	33.87/13.2	57.59/19.9	50.95/18.0	42.28/14.4	29.13/9.6	20.95/7.9	24.36/9.7	25.95/9.3	17.84/6.4
Total3D [27]	49.42/18.27	80.33/30.5	43.74/17.4	68.54/24.9	61.93/22.8	53.93/19.2	44.34/15.9	39.13/**14.6**	45.89/17.2	36.90/12.6	19.46/7.6
Ours(w/o. HPG)	49.23/18.11	78.86/30.8	44.26/17.2	74.02/26.8	60.80/21.5	55.62/19.6	45.63/**17.0**	34.78/12.7	39.38/15.0	37.62/12.8	21.35/7.7
Ours	**54.27/20.04**	**81.62/33.0**	**50.70/20.5**	**77.62/29.3**	**67.90/24.0**	**60.29/21.6**	**48.54**/16.4	**40.32**/13.7	**49.00/18.2**	**43.57/14.7**	**23.78/9.0**

Table 5. Comparisons on SUN RGB-D dataset under the metrics of HPE.

Method	$RelAcc$	$PhrAcc$	$RelAcc_I$	$PhrAcc_I$
CooP [15]	20.56	2.82	39.33	4.31
Total3D [27]	38.25	6.69	58.28	12.88
Ours	**40.09**	**9.19**	**60.83**	**18.49**

and 30° respectively. The IoU threshold of the 3D bounding box is 0.15. As shown in Table 5, our method has better performance than state-of-the-art methods, benefited from modeling geometric structure explicitly with HPG. It not only justifies the effectiveness of our method, but also demonstrates that HPE further distinguishes the performance of different methods from a holistic perspective to analyze the rationality of scene layout.

5.5. Qualitative Experiments

We show some typical qualitative results of our method and the baseline (Total3D [27]) on SUN RGB-D dataset. The three rows in Figure 6 represent the improvements of 3D object prediction about location, scale, and rotation individually. It can be seen that our method predicts more accurate 3D bounding boxes when there are more objects in the scene. It proves that HPG can integrate more geometric information to maintain a more reasonable scene layout.

5.6. Analysis of Graph Inference

We show some qualitative and quantitative results to analyze the process of graph inference along HPG. In Figure 7, we visualize the process of dynamically updating the object pose. After each message passing process, each object will change its pose gradually to maintain a more reasonable holistic geometric structure. We evaluate the process of graph inference quantitatively in Figure 8, and observe that the performance under each metric goes up with the iteration increasing and becomes stable after about 3 iterations of graph inference, as verified in both the qualitative and quantitative results.

Besides, we further explore the impact of different sample rates of the graph edges by random sampling. The results indicate that more edges, more constraints, lead to

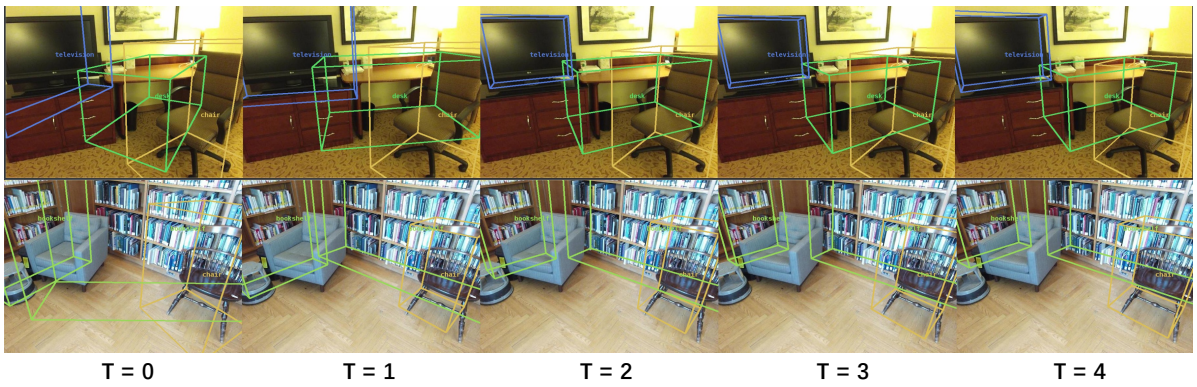

Figure 7. Visualization of the intermediate results during graph inference. T denotes the iterations of the message passing process.

higher performance (details in the supplementary material).

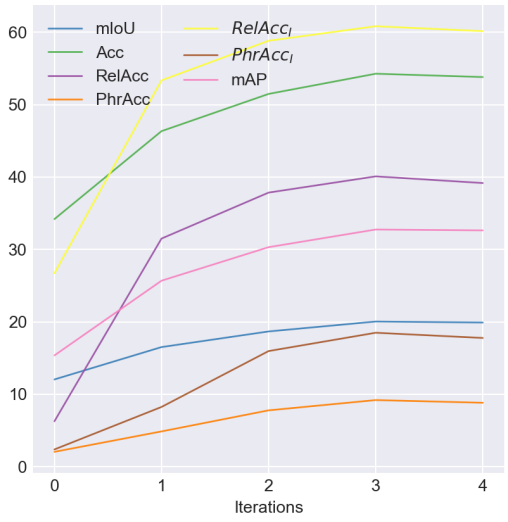

Figure 8. Quantitative intermediate results during graph inference on all metrics.

5.7. Ablation Study

In this section, we mainly analyze the influences of the Holistic Pose Graph and consistency loss. To better understand the contributions of each module, we ablate our method with four settings in Table 6:
S_0: Baseline method [27]
S_1: Final model - HPG
S_2: Final model - consistency loss
S_3: Final model

Comparing S_1 with S_3: Under the setting of **S_1**, we encode the relative pose but do not build HPG for graph inference. The results demonstrate that the key for performance improvement is the HPG rather than only focusing on pairwise relations [20].

Comparing S_2 with S_3 and S_1: The last four rows of **S_2** are better than **S_1** because of the HPG improving the performance about the relationships prediction. Compared with **S_3**, **S_2** discards the consistency loss, and thus the object pose module and the relative pose module are trained respectively, which causes the first three rows about the objects prediction with obvious gap between **S_2** and **S_3**.

Comparing S_0 with S_2: The last four rows of **S_2** are better than **S_0**. It proves that modeling geometric structure explicitly using HPG is a better way to leverage the relationships.

Table 6. Ablation studies of all metrics on SUN RGB-D dataset [32]. mAP is the result of 3D object detection, mIoU and Acc both correspond to 3D box estimation, and the last four rows are evaluated on HPE.

Metric	S_0	S_1	S_2	S_3
mAP	26.38	27.39	27.99	**32.75**
mIoU	18.27	18.11	18.41	**20.04**
Acc	49.42	49.23	49.58	**54.27**
$RelAcc$	38.25	33.79	**40.48**	40.09
$PhrAcc$	6.69	6.27	8.05	**9.19**
$RelAcc_I$	58.28	54.49	60.59	**60.83**
$PhrAcc_I$	12.88	12.17	16.43	**18.49**

6. Conclusion

We propose to model geometric structure among objects in a scene using graph inference for 3D object prediction. The experiments on SUN RGB-D dataset demonstrate that Holistic Pose Graph is a better way to utilize the geometric relationships than existing methods, and the key to improving the performance of the object prediction along with the relationships prediction is our devised consistency loss. Besides, we introduce Holistic Pose Estimation that further evaluates the rationality of the scene layout based on considering the accuracy of independent object.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. U19B2036, 61922080, 61772500, CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009, and National Key R&D Program of China (2020AAA0105200).

References

[1] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014*, pages 103–111, 2014. 3

[2] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Understanding indoor scenes using 3d geometric phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 33–40, 2013. 2

[3] 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

[4] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. In *Advances in Neural Information Processing Systems*, pages 1733–1743, 2018. 2

[5] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas A. Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *ACM Trans. Graph.*, 31(6):135:1–135:11, 2012. 2

[6] Matthew Fisher, Manolis Savva, and Pat Hanrahan. Characterizing structural relationships in scenes using graph kernels. *ACM Trans. Graph.*, 30(4):34, 2011. 2

[7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 2

[8] Ruiqi Guo and Derek Hoiem. Support surface prediction in indoor scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2144–2151. IEEE Computer Society, 2013. 6

[9] Abhinav Gupta, Alexei A Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Proceedings of the European Conference on Computer Vision*, pages 482–496. Springer, 2010. 2

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 5

[12] Tong He and Stefano Soatto. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. In *AAAI*, pages 8409–8416, 2019. 1

[13] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 2

[14] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single RGB image via perspective points. In *Advances in Neural Information Processing Systems*, pages 8903–8915, 2019. 1, 2, 5, 6

[15] 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*, pages 206–217, 2018. 2, 3, 4, 5, 6, 7

[16] 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*, pages 187–203, 2018. 1, 2, 6

[17] Hamid Izadinia, Qi Shan, and Steven M. Seitz. IM2CAD. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2422–2431. IEEE Computer Society, 2017. 1

[18] Chenfanfu Jiang, Siyuan Qi, Yixin Zhu, Siyuan Huang, Jenny Lin, Lap-Fai Yu, Demetri Terzopoulos, and Song-Chun Zhu. Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars. *Int. J. Comput. Vis.*, 126(9):920–941, 2018. 2

[19] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2019. 3

[20] Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, and Abhinav Gupta. 3d-relnet: Joint object and relational network for 3d prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2212–2221, 2019. 1, 2, 6, 7, 8

[21] David C Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. 2010. 2

[22] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1417–1424, 2013. 2

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 5

[24] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018. 2, 3

[25] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision*, pages 852–869. Springer, 2016. 2

[26] 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

[27] 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 (CVPR)*, June 2020. 1, 2, 3, 4, 5, 6, 7, 8

[28] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018. 4

[29] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 5

[30] Yuzhuo Ren, Shangwen Li, Chen Chen, and C-C Jay Kuo. A coarse-to-fine indoor layout estimation (cfile) method. In *Asian Conference on Computer Vision*, pages 36–51. Springer, 2016. 2

[31] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760, 2012. 5, 6

[32] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 2, 3, 5, 6, 8

[33] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 6

[34] 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. 6, 7

[35] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 2, 3

[36] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pages 670–685, 2018. 2

[37] Shunyu Yao, Tzu-Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman, and Josh Tenenbaum. 3d-aware scene manipulation via inverse graphics. In *Advances in Neural Information Processing Systems*, pages 1891–1902, 2018. 1

[38] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 2

[39] 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, 2014. 2

[40] Yibiao Zhao and Song-Chun Zhu. Image parsing with stochastic scene grammar. *Advances in Neural Information Processing Systems*, 24:73–81, 2011. 2

[41] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3119–3126, 2013. 2

[42] 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