BORM: Bayesian Object Relation Model for Indoor Scene Recognition

Liguang Zhou[1,2], Jun Cen[2,3], Xingchao Wang[1,2], Zhenglong Sun[1,2], Tin Lun Lam[1,2,†], Yangsheng Xu[1,2]

***Abstract*— Scene recognition is a fundamental task in robotics perception. For human beings, scene recognition is reasonable because they have abundant object knowledge of the real-world. The idea of transferring object knowledge from humans to scene recognition is significant but still less exploited. In this paper, we propose to utilize meaningful object representations for indoor scene representation. First, we utilize an improved object model (IOM) as a baseline that enriches the object knowledge by introducing a scene parsing algorithm pretrained on the ADE20K dataset with rich object categories related to the indoor scene. To analyze the object co-occurrences and pairwise object relations, we formulate the IOM from a Bayesian perspective as the Bayesian object relation model (BORM). Meanwhile, we incorporate the proposed BORM with the PlacesCNN model as the combined Bayesian object relation model (CBORM) for scene recognition and significantly outperforms the state-of-the-art methods on the reduced Places365 dataset, and SUN RGB-D dataset without retraining, showing the excellent generalization ability of the proposed method. Code can be found at https://github.com/FreeformRobotics/BORM.**

I. Introduction

Scene understanding is a fundamental cognition ability for the robot to conduct the task in an unknown environment. To make the robot the part of family life, such as family members like dogs or cats, the most fundamental problem is to know where the robot is. There is an increasing need to equip the robot with the scene recognition ability. e.g., the ability to locate itself semantically, to known which room it is located, such as the living room or a bedroom, or a kitchen.

Knowing the information about the current environment is quite essential for robots to make more intelligent actions. Fortunately, scene recognition has been an important research area among computer vision and robotics for decades. There have been many scene recognition algorithms proposed that using visual information for scene understanding. Liu et al. [1] [2] propose a color and geometric features based adaptive descriptor for scene recognition. Also, indoor scenes can be characterized by global spatial features [3]. However, these methods are based on the traditional lower-level features, which are not accurate enough, and not interpretable at the semantic level.

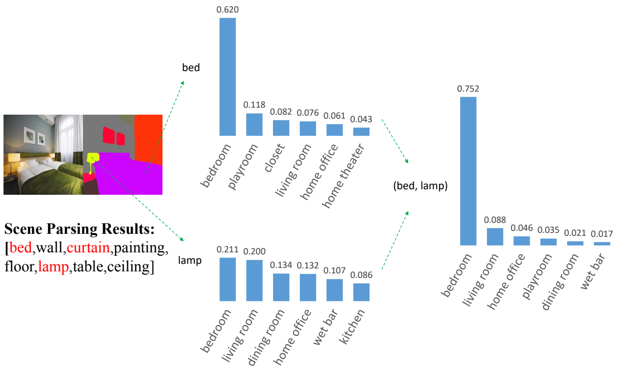


Fig. 1. The left part of the figure shows the input image and the corresponding scene parsing result. We have conducted a Bayesian probabilistic analysis using the BORM on the Places365-14 dataset. The middle part shows the conditional probability of $P(scene|object)$, and the top 6 scene given the object "bed" and "lamp" are presented, respectively. The right part shows the joint conditional probability of $P(scene|(object, object))$, and the top 6 scene given the object pair $(bed, lamp)$. As observed from the figure, the proposed BORM suppresses the probability of scene-common object pairs while highlights the probability of the scene-specific object pairs.

Recently, many deep learning based methods have demonstrated impressive performance over various computer vision tasks. For example, ResNet has been shown a convincing ability in image classification tasks for years [4]. However, the classification of the image using ResNet is like using a black box to interpret the image, which is not really in an interpretable manner. Chen et al. [5] consider the word-embedding model to reformulate the scene understanding problem in a more semantic meaningful perspective. The output of ResNet module, object detection module, and scene parsing module have been encoded in the one global vector for scene understanding. However, the performance of the word embedding method is just slightly higher than ResNet, as reported. Besides, there are three streams used for scene recognition, which is relatively inefficient. The utilization of semantic information for scene recognition is important. To model the descriptors probabilities, a Bayesian filtering method is proposed [6], but the object information is not explicitly used. To use the object information, a object classifier is used to classify low-level visual features to objects [7], while the relation among objects is not exploited. To utilize the object relation for scene recognition, spatial object-to-object relation is studied for RGB-D scene recognition [8]. Besides, a Long Short-Term Memory modeling method is proposed to investigate the object relation with ROI selection [9]. To utilize the object information in the scene, the object model [10] is proposed as complementary

[†] Corresponding Author: Tin Lun Lam (tllam@cuhk.edu.cn)

This work was supported in part by the funding AC01202101025 and 2019-INT007 from the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

[1] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen

[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen.

[3] Hong Kong University of Science and Technology

semantic information of the scene combined with ResNet to better interpret the given scene. However, using only the object vector might not be discriminative enough for scene understanding [11]. A Object-to-Scene model is proposed, where the object features and object relation are learned by object feature aggregation module and object attention module [12], respectively. Overall, these above-mentioned method lack the statistically analysis of object distribution in the scene. Specifically, there are some common objects across various scene and specific objects only appear in the specific scene. Therefore, the Bayesian object relation model is proposed for improving the accuracy of scene recognition by highlighting the non-discriminative objects while preserving the discriminative objects for scene understanding [13]. However, all of these methods neglect the probabilistic relation among object pairs, which contains an essential message for scene recognition.

In this paper, we consider the probabilistic relation between the object pairs as shown in Fig. 1. We can observed that there is a high probability of the object pair appear in the specific scene, e.g., the bed and lamp is most likely (75.2%) appeared in the bedroom, and can be regarded as a scene-specific object pair.

To utilize this conditional object pair relation w.r.t the various scene, we propose the Bayesian object relation model (BORM) that derives the joint conditional probability of the object pairs given the various scene. With the BORM, the joint conditional probability of the scene-specific object pairs will be enhanced and the joint conditional probability of scene-common object pairs will be suppressed.

In summary, our main contributions of this paper are as follows:

- We utilize an IOM as baseline, based on scene parsing algorithm pretrained on ADE20K dataset that contains more relevant object classes for indoor scene representation. Surprisingly, the IOM surpasses the object model over **20.5%** accuracy on average.
- Meanwhile, To conduct an in-depth study of object pair relations, we propose a Bayesian object relation model (BORM) that enhances the scene-specific object pair relation and suppresses the scene-common object pair relation in a Bayesian probability manner for indoor scene representation.
- We combine the proposed BORM and the PlacesCNN model as CBORM, which significantly outperforms the state-of-the-art methods on the Places365-7 and Places365-14 dataset over 2.0% and 2.1% accuracy, and SUN RGB-D dataset over 2.0% without retraining the model, showing the excellent generalization ability of the proposed method.

The rest of the paper is organized as follows. Section II introduces the related work of scene recognition and object knowledge for indoor scene representation. Section III describes IOM that pretrained on ADE20K with 150 categories based on the scene parsing algorithm, and the BORM that conduct an in-depth study of object relation in a Bayesian perspective. In Section IV, we discuss the PlacesCNN model for scene recognition. Moreover, we combine the proposed BORM and the PlacesCNN model as the CBORM for scene recognition. Section V shows the experimental settings and results, and numerical experiments have been conducted to verify the effectiveness of our proposed method. Finally, the conclusions and future directions are pointed out in Section VI.

II. Related Work

In this section, we review the research works related to our paper in two aspects: scene recognition, and object knowledge from object detection or scene parsing algorithm for indoor scene representation. We also discuss the differences and connections between these related works and our method.

A. Scene Recognition

Scene recognition has been an important research problem in robotics area for decades, which can be utilized for topological map construction and mobile robot localization [1] [2] [14]. Moreover, it has the potential to be applied for robot to perform efficient recognition of functional areas [15], identification of the person [16], and execute task accordingly. The early methods, mainly focus on extraction of local features like color descriptors [2] [17]. To better recognize indoor images, the combination of local and global features are utilized [3]. However, these methods only captures lower-level features of the scene while the high-level semantic structures are difficult to capture [18].

To utilize the high-level semantic information, some methods propose to leverage the mid-level concepts. e.g., Zhou et al. [19] [20] propose a CNN based classifier to learn deep features for scene recognition on the Places Dataset. Liao et al. [21] use deep learning with a multi-task training method that incorporate both semantic segmentation and scene recognition tasks. Zhu et al. [18] propose a discriminative multimodal feature fusion framework for scene recognition. However, these methods neglect the critical object information for scene recognition.

To incorporate object information, Li et al. [22] represent images by using objects appearing in them as object bank (OB) method. Brucker et al. [23] counts the co-occurrence frequencies as potentials in conditional random field for scene labeling. In DEDUCE [10], one hot object vector is used as complementary information for scene recognition, where the object information is separately considered. In context based Word Embeddings [5], there are three streams for scene recognition, one stream is the scene parsing model pretrained on ADE20K, the other is ResNet50 pretrained on reduced Places365 dataset, and a Word Vectors Module computes the content of two modules and refines the results. Song et al. [24] considers spatial object-to-object relations with the intermediate (object) representations. In Semantic-aware method [25], the image representation and context information are combined, where context information consists of scene objects and stuff, and their relative locations.

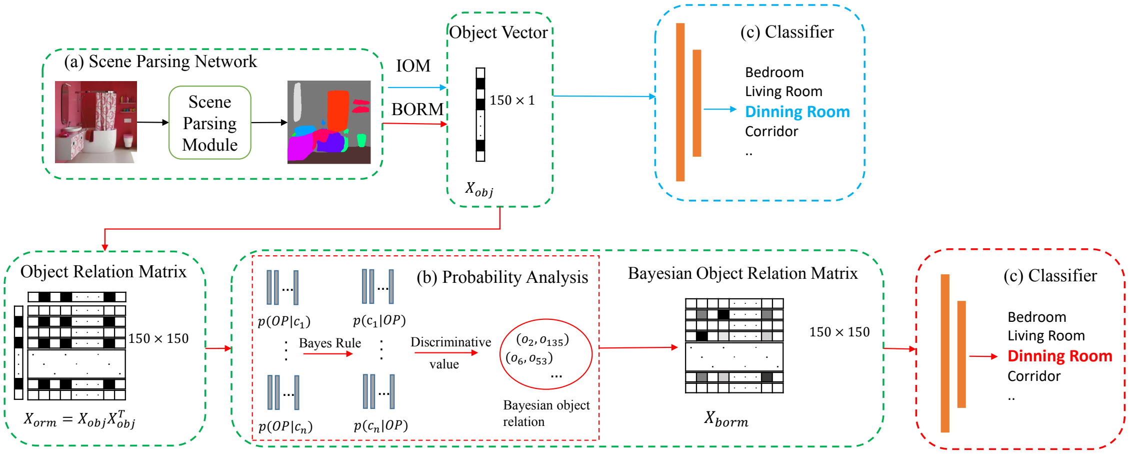


Fig. 2. A variations of improvements over the object model, where blue and red arrow shows the flow of the IOM and BORM, respectively.

However, all of these methods have not considered the probabilistic relation of the object pairs given the various scene. In this work, we consider the object pair co-occurrences given various scene in a Bayesian Perspective with proposed Bayesian object relation model (BORM). To the best of our knowledge, in the scene recognition task, the object relation modeled in a probabilistic perspective is less exploited.

B. Object Knowledge for Indoor Scene Representation

Recently, object detection is a popular research area in computer vision, thanks to the emerging of the large-scale labeled data and advanced GPUs. Many excellent algorithms have been developed for object detection like the one-stage object detector, YoloV3 [26], SSD [27] and two-stage object detector, like Joint SSD [28], Faster R-CNN [29], Mask RCNN [30], Cascaded RCNN [31]. The one-stage detector has a higher speed while maintain the similar performance compared with two-stage detector, therefore, the YoloV3 is adopted as the part of object model in DEDUCE [10]. To incorporate the object knowledge for scene understanding, the object detection algorithm is first pretrained on the public available dataset. There are several mainstream dataset for object detection. PASCAL VOC [32] contains 20 categories of objects, which is very limited, and most labels are like car, bus, bicycle, airplane, and all of these are not relevant to indoor scene representation. To better represent indoor scene, DEDUCE use the YoloV3 pretrained on the MS COCO dataset [33], which contains 80 object categories. However, the half of objects in MS COCO are outdoor objects like giraffe, elephant, which are not relevant to indoor scene representations.

Scene parsing algorithm is used to segment objects and stuff in the still image, which has demonstrated surprisingly performance recently [34]. Zhou et al. [35] propose a scene parsing algorithm to detect a wide range of object and stuff in the pixel level with ADE20K Dataset, which contains 150 classes of object knowledge in pixel level. To utilize the rich object knowledge of the ADE20K Dataset, we adapt the scene parsing model pretrained on the ADE20K dataset as the improved object model (IOM), which shows a significantly improvements over the OM pretrained on the MS COCO dataset.

III. Bayesian Object Relation Model

In Section III, we present the object models IOM and BORM, as shown in Fig. 2. Since the object model proposed in Deduce [10] only contains information about few categories of objects pretrained on the MS COCO dataset, which is very limited for indoor scene recognition because most of the object categories are not relevant to indoor scene representation. e.g., the object categories of elephant and giraffe from the super-category of animal, airplane and bus from the super-category of vehicle, traffic light and stop sign from the super-category of outdoor, and so on. In total, there are half of the objects can be categorized as outdoor objects. We believe that if the object model possesses the more rich and relevant object knowledge about the scene, the better performance of scene recognition can be obtained. Therefore, we propose the IOM pretrained on the ADE20K dataset with rich and relevant object categories for indoor scene representation. To consider the probabilistic relation of object pairs, we assume some object pairs are scene-specific, and some object pairs are scene-common. To this end, we propose a novel BORM that highlights the scene-specific object pairs while suppresses the scene-common object pairs from a Bayesian probabilistic perspective.

A. Improved Object Model (IOM)

Different from the basic OM that only have 80 objects information of environment based on the YoloV3 pretrained on the MS COCO dataset, we propose the IOM based on the scene parsing algorithm pretrained on ADE20K [34] dataset that convert the output the scene parsing algorithm to an object vector $\mathbf{X_{iom-150}}$ with 150 dimensions where the 1 means the detected objects, while 0 means the objects are not in the given image. Compared with the OM, we now have a

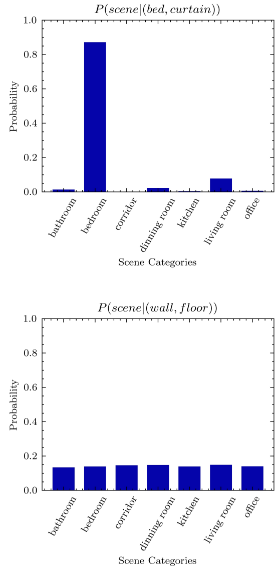


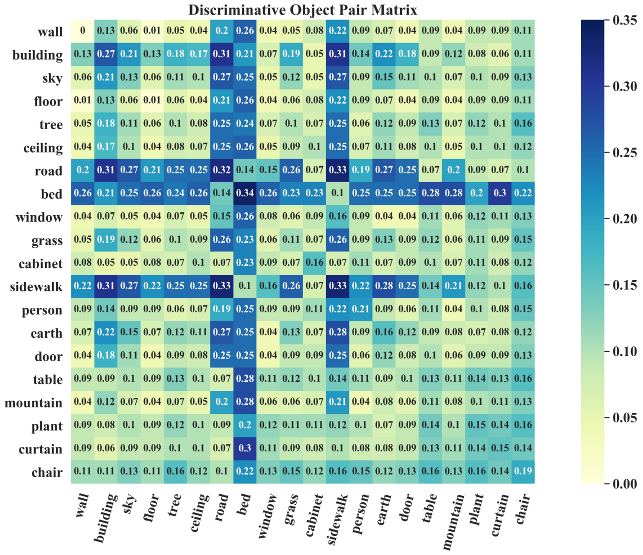
Fig. 3. The BORM contains the discriminative value of object pairs. We have conducted statistical analysis on the Places365-7 dataset. The top figure shows the $p(scene|(bed, curtain))$, the probability distribution of object pair (bed, lamp) over the seven indoor scenes, and the standard deviation is 0.30, which means the $(bed, curtain)$ is a discriminative object pair for indoor scene recognition. The bottom figure shows the $p(scene|(wall, floor))$, the probability distribution of object pair $(wall, floor)$ among seven indoor scenes, which are almost the same. The standard deviation is 0.01, which indicates the $(wall, floor)$ is a non-discriminative object pair for indoor scene recognition, because this object pair appears in every scene equally.

new scene representation of 150 dimension $\mathbf{X_{iom-150}}$. The new scene representation $\mathbf{X_{iom-150}}$ will be fed to a two-layer fully connected network for classification with the size of 32 and the number of scenes, respectively.

B. Bayesian Object Relation Model (BORM)

Instead of using one hot object vector as the feature representations for the scene, which lacks the probabilistic relation between different object pairs, we propose a novel BORM method that measures the probabilistic relation of object pairs in a Bayesian perspective. The scene recognition problem is shaped by the fact that a few object pairs are scene-common, but most object pairs are scene-specific. As observed in Fig. 3, the probability of scene-specific object pair $(bed, curtain)$ and scene-common object pair $(wall, floor)$ are presented.

The scene-specific object pairs indicate the object pairs that have a high probability at a particular scene and have low probability at others. e.g., the object pairs like $(bed, curtain)$ tend to be scene-specific, which means their probability $p(scene|(bed, curtain))$ can be quite distinctive among the various scene categories and has the highest probability that appears in the bedroom. Therefore, they have a large standard deviation (0.31). In contrast, scene-common object pairs distributed at various scenes with a similar probability. e.g., the common object pairs like $(wall, floor)$ frequently appear in many different scenes, which means their joint conditional probability $p(scene|(wall, floor))$ is quite similar across various scenes, i.e., they have an extremely small standard deviation (0.01) compared with those scene-specific object pairs.



Discriminative Object Pair Matrix

	wall	building	sky	floor	tree	ceiling	road	bed	window	grass	cabinet	sidewalk	person	earth	door	table	mountain	plant	curtain	chair
wall	0	0.13	0.06	0.01	0.05	0.04	0.2	0.26	0.04	0.05	0.08	0.22	0.09	0.07	0.04	0.09	0.04	0.09	0.09	0.11
building	0.13	0.27	0.21	0.13	0.18	0.17	0.31	0.21	0.07	0.19	0.05	0.31	0.14	0.22	0.18	0.09	0.12	0.08	0.06	0.11
sky	0.06	0.21	0.13	0.06	0.11	0.1	0.27	0.25	0.05	0.12	0.05	0.27	0.09	0.15	0.11	0.1	0.07	0.1	0.09	0.13
floor	0.01	0.13	0.06	0.01	0.06	0.04	0.21	0.26	0.04	0.06	0.08	0.22	0.09	0.07	0.04	0.09	0.04	0.09	0.09	0.11
tree	0.05	0.18	0.11	0.06	0.1	0.08	0.25	0.24	0.07	0.1	0.07	0.25	0.06	0.12	0.09	0.13	0.07	0.12	0.1	0.16
ceiling	0.04	0.17	0.1	0.04	0.08	0.07	0.25	0.26	0.05	0.09	0.1	0.25	0.07	0.11	0.08	0.1	0.05	0.1	0.1	0.12
road	0.2	0.31	0.27	0.21	0.25	0.25	0.32	0.14	0.15	0.26	0.07	0.33	0.19	0.27	0.25	0.07	0.2	0.09	0.07	0.1
bed	0.26	0.21	0.25	0.26	0.24	0.26	0.14	0.34	0.26	0.23	0.23	0.1	0.25	0.25	0.25	0.28	0.28	0.2	0.3	0.22
window	0.04	0.07	0.05	0.04	0.07	0.05	0.15	0.26	0.08	0.06	0.09	0.16	0.09	0.04	0.04	0.11	0.06	0.12	0.11	0.13
grass	0.05	0.19	0.12	0.06	0.1	0.09	0.26	0.23	0.06	0.11	0.07	0.26	0.09	0.13	0.09	0.12	0.06	0.11	0.09	0.15
cabinet	0.08	0.05	0.05	0.08	0.07	0.1	0.07	0.23	0.09	0.07	0.16	0.07	0.11	0.07	0.09	0.1	0.07	0.11	0.08	0.12
sidewalk	0.22	0.31	0.27	0.22	0.25	0.25	0.33	0.1	0.16	0.26	0.07	0.33	0.22	0.28	0.25	0.14	0.21	0.12	0.1	0.16
person	0.09	0.14	0.09	0.09	0.06	0.07	0.19	0.25	0.09	0.09	0.11	0.22	0.21	0.09	0.06	0.11	0.04	0.1	0.08	0.15
earth	0.07	0.22	0.15	0.07	0.12	0.11	0.27	0.25	0.04	0.13	0.07	0.28	0.09	0.16	0.12	0.09	0.08	0.07	0.08	0.12
door	0.04	0.18	0.11	0.04	0.09	0.08	0.25	0.25	0.04	0.09	0.09	0.25	0.06	0.12	0.08	0.1	0.06	0.09	0.09	0.13
table	0.09	0.09	0.1	0.09	0.13	0.1	0.07	0.28	0.11	0.12	0.1	0.14	0.11	0.09	0.1	0.13	0.11	0.14	0.13	0.16
mountain	0.04	0.12	0.07	0.04	0.07	0.05	0.2	0.28	0.06	0.06	0.07	0.21	0.04	0.08	0.06	0.11	0.08	0.1	0.11	0.13
plant	0.09	0.08	0.1	0.09	0.12	0.1	0.09	0.2	0.12	0.11	0.11	0.12	0.1	0.07	0.09	0.14	0.1	0.15	0.14	0.16
curtain	0.09	0.06	0.09	0.09	0.1	0.1	0.07	0.3	0.11	0.09	0.08	0.1	0.08	0.08	0.09	0.13	0.11	0.14	0.15	0.14
chair	0.11	0.11	0.13	0.11	0.16	0.12	0.1	0.22	0.13	0.15	0.12	0.16	0.15	0.12	0.13	0.16	0.13	0.16	0.14	0.19

Fig. 4. The Bayesian object relation matrix of first 20 objects. e.g., the $(bed, curtain)$ is the scene-specific object pair, in which will mostly leading to the bedroom, while $(wall, floor)$ forms the scene-common object pair because it is appear in everywhere with almost same probability.

We adopt a pipeline to estimate the posterior probability $P(c_j|o_h, o_i)$. First, we use the Places365-7 and Places365-14 dataset for learning the BORM statistically with only the training set, while testing on the SUN RGB-D dataset, where the probability distribution is regarded similar to the Places365-7 dataset.

Given a set of images I_{c_j} from a scene category c_j, the conditional probability of object o_i appears on the scene c_j is:

$$\begin{gathered} P\left(o_i|c_j\right) = N_{o_i}/N_{I_{c_j}} \\ i \in [1, N_{objs}], j \in [1, N_{scenes}] \end{gathered} \tag{1}$$

where N_{o_i} is the total number of i-th object o_i appears in I_{c_j} and $N_{I_{c_j}}$ is total number of images of I_{c_j}. Meanwhile, where N_{objs}, and N_{scenes} represent the number of objects in the pretrained model and number of scene categories we have in the dataset respectively. Assume the statistical independence of each object, we obtain the joint conditional probability $P\left(o_h, o_i|c_j\right)$ of o_h and o_i appear in the scene c_j:

$$\begin{gathered} P\left(o_h, o_i|c_j\right) = P\left(o_h|c_j\right) P\left(o_i|c_j\right) \\ h, i \in [1, N_{objs}], j \in [1, N_{scenes}] \end{gathered} \tag{2}$$

The $P(c_j|o_h, o_i)$, the posterior probability of scene class c_j given an object pair (o_h, o_i), can be derived by the Bayes Rule and Law of Total Probability.

$$\begin{aligned} P\left(c_j|o_h, o_i\right) &= \frac{P\left(o_h, o_i|c_j\right) P\left(c_j\right)}{P(o_h, o_i)} \\ &= \frac{P\left(o_h, o_i|c_j\right) P\left(c_j\right)}{\sum_j P\left(o_h, o_i|c_j\right) P\left(c_j\right)} \end{aligned} \tag{3}$$

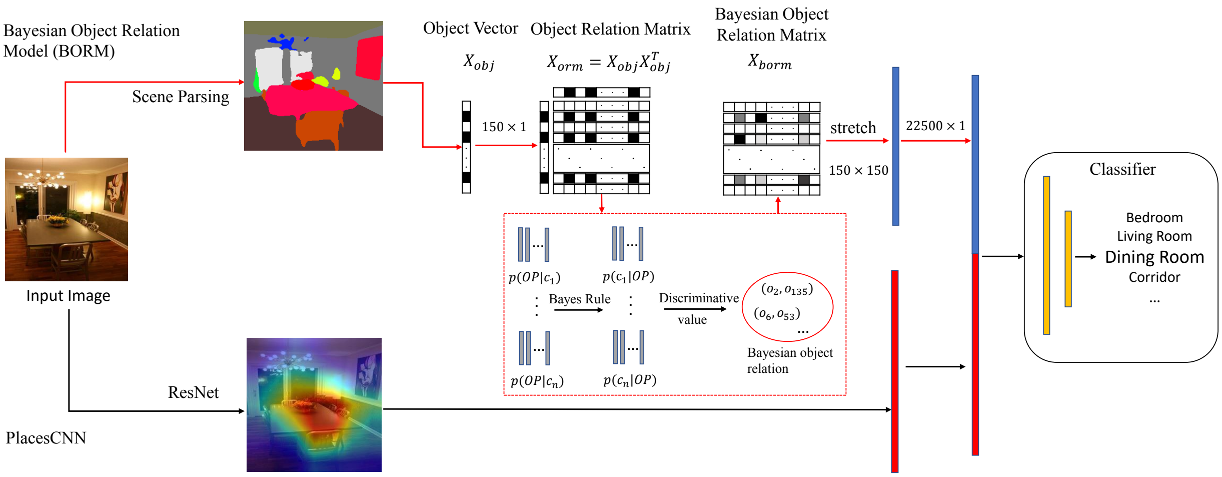


Fig. 5. The proposed combined Bayesian object relation model (CBORM) contains two streams. The first stream with the red arrow is the proposed BORM that utilizes the scene parsing algorithm for scene segmentation. The second stream with the black arrow is the PlacesCNN model for feature extraction with the ResNet backbone network. The output of the BORM will first be converted to a discriminative object feature. Meanwhile, the results of the BORM and feature extraction result of PlacesCNN will be concatenated as one combined scene feature and fed to the two-layer FC network for scene classification.

where the $P(c_j)$ is the probability of scene c_j in the dataset, and $\sum_{j=1}^{N_{scenes}} P(c_j) = 1$. We construct the posterior probability matrix $P(c_j|o,o)$ by calculating relation of each object pairs:

$$P\left(c_j|o,o\right) = \begin{bmatrix} P\left(c_j|o_1,o_1\right) & \cdots & P\left(c_j|o_1,o_k\right) & \cdots & P\left(c_j|o_1,o_n\right) \\ \vdots & & \vdots & & \vdots \\ P\left(c_j|o_j,o_1\right) & \cdots & P\left(c_j|o_j,o_k\right) & \cdots & P\left(c_j|o_j,o_n\right) \\ \vdots & & & & \vdots \\ P\left(c_j|o_n,o_1\right) & \cdots & P\left(c_j|o_n,o_k\right) & \cdots & P\left(c_j|o_n,o_n\right) \end{bmatrix} \tag{4}$$

First, the posterior probability $P(c_j|o_h, o_i)$ is calculated. To calculate the discriminative value of the object pairs, the standard deviation is applied to posterior probabilities among scene categories, denoted as $std(P(c_j|o_h, o_i))$. The discriminative value of the posterior probabilities among scene categories is defined as $dis(o_h, o_i)$:

$$\text{dis}\left(o_h, o_i\right) = std_{c\in 1,\ldots,N_{scenes}}(P(c_j|o_h, o_i)) \tag{5}$$

The $dis(o_h, o_i)$ is the discriminative value for measuring the discriminalibility of the object pair to the scene categories. Instead of using object vector as the feature representations for the scene, which dismiss the relationship between different objects, we first construct an object relation matrix $\mathbf{X_{orm}} = \mathbf{X_{iom}}\mathbf{X_{iom}^T}$ for representing the given scene. The value of $\mathbf{X_{orm}}$ will be replaced by the discriminative value $dis(o_h, o_i)$ at the same position. Therefore, the Bayesian object relation matrix is obtained by element-wise matrix multiplication, $\mathbf{X_{borm}} = \mathbf{X_{orm}} * \mathbf{dis(o_h, o_i)}$, with the size of 150x150. After that, the matrix will be stretched to a one dimensional vector with the size of 22500x1, and will be fed into the three-layer fully connected network with the size of 8192, 2048, and number of scene classes.

As shown in Fig. 4, the discriminative matrix of the first 20 object pairs are displayed. The $(bed, curtain)$ is a scene-specific object pair that mostly appear in the bedroom, and they form a discriminative object pair with a discriminative value of 0.30. The $(wall, floor)$ is a common object pair appear everywhere and have an extremely small discriminative value of 0.01, which means they form a non-discriminative object pair.

IV. CBORM Model

A. PlacesCNN Model

In order to obtain the scene representation, we use the PlacesCNN model [20] with the base architecture ResNet [4] as a backbone network, which is pretrained on the ImageNet [36] dataset. There are two versions of ResNet according to the settings in DEDUCE [10] and Word2Vec [5]. ResNet18 is used for Places365-7 and SUN RGB-D dataset, while ResNet50 is used for Places365-14 dataset. Specifically, we preserve the output of ResNet18 with 512 dimensions or ResNet50 with 2048 dimensions, namely F_{scene}, the feature of the PlacesCNN model for the scene representation.

B. Combined Model

The Bayesian object relation matrix of the BORM is with the size of 150x150, which is first stretched to a vector of size 22500x1. The vector of BORM will be fed into the two fully connected layers and an discriminative object feature F_{BORM} with 512 dimensions (for Places365-7) or 2048 (For Places365-14) will be the feature representation of BORM.

Meanwhile, we combine the PlacesCNN model of the ResNet backbone with BORM, as the combined Bayesian object relation model (CBORM).

TABLE I

DATASET SPLIT SETTING, WHERE THE NUMBER OF TRAINING SET AND NUMBER OF TEST SET ARE LISTED BELOW.

Dataset	Training	Test
Places365-7	35000	701
Places365-14	75000	1500
SUN RGB-D	35000(from Places365-7)	2077

C. Combined Classifier

For the Places365-7 dataset, there are two streams. To fair compare with the object model [10], one stream is based on ResNet18 pretrained on the Places365 and finetuned on the Places365-7 dataset. The other stream is BORM. The feature representations of ResNet18 and BORM will be concatenated and fed into a two-layer FC network with a dimension of 512, and 7 respectively.

Similarly, for the Places365-14 dataset, there are two streams. To fair compare with the word-embedding [5] model, while one stream is ResNet50, and the other is BORM. The feature representations of ResNet50 and BORM will be fed into a two-layer FC network with a dimension of 512, and 14 respectively.

V. EXPERIMENTAL RESULTS

A. Experimental Settings

We evaluated the proposed models on reduced SUN RGB-D and Places365 dataset. To be noticed, aim to investigate the generalizability of our model, we evaluate our model pretrained on Places365-7 dataset on the SUN RGB-D dataset without retraining it. We'll introduce the implementation details and training procedure and different experiment settings.

1) Implementation Details: For the PlacesCNN model, ResNet18 or ResNet50 architecture is adopted in our experiment for ablation study. The optimizer used is the Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, the momentum of 0.9, and the weight decay of 0.0001. We decrease the learning rate 10 times every 10 epoch, and every time when updating the learning rate, we reload the parameters which have the best accuracy before this timestamp. The total number of epoch during training is 40. To be noticed, we use the training sets of Places365-7 and Places365-14 dataset for learning the BORM statistically, while testing on the SUN RGB-D dataset, where the BORM is the same as the Places365-7 dataset.

2) Dataset Settings:

Places365 Dataset: In this paper, we use the reduced Places365 [20] dataset to test our methods, since it is the most largest and challenging scene classification dataset yet, and it contains broad categories in the indoor environment. In the experiment, we only consider the indoor scene recognition. There are two different settings on the reduced Places365 dataset. The one is Places365 with 7 classes includes Corridor, Dinning Room, Kitchen, Living Room, Bedroom, Office, and Bathroom, denoted as Places365-7. The test set setting is the same as the official dataset and described in [10]. In addition, we use the reduced Places365 with 14 indoor scenes in Home environment includes Wet bar, Home theater, Balcony, Closet, Kitchen, Bedroom, Playroom, Laundromat, Bathroom, Living Room, Home office, Dining room, Staircase, and Garage denoted as Places365-14. The dataset splitting follows the same setting as described in [5]. The dataset splition can be seen from Table. I.

SUN RGB-D Dataset: SUN RGB-D dataset [37] is a challenging dataset for scene understanding that contains not only RGB images but also depth information of each image. It contains 3784 images collected by Kinect V2 and 1159 collected by Intel RealSense. Moreover, it incorporates 1449 images from the NYUDepth V2 [38], and 554 images from the Berkeley B3DO Dataset [39], both captured by Kinect V1. Finally, it takes 3389 manually selected distinguished frames without significant motion blur from the SUN3D videos [40] captured by Asus Xtion.

In our experiment, we mainly consider the indoor environment understanding. Therefore, we use the reduced SUN RGB-D dataset includes Office, Kitchen, Bedroom, Corridor, Bathroom, Living room, and Dining room, where the test set split is the same as the official dataset. There are 3741 RGB images in total for testing. And we test our model pretrained on the Place365-7 on SUN RGB-D dataset without retraining.

B. Experimental Results

1) Effect of Object Knowledge: As illustrated in Fig. 6, a group of ablation studies have been conducted for evaluating the effect of object knowledge to indoor scene recognition on the Places365-14, Places365-7, and SUN RGB-D dataset, respectively. The x-axis represent the number of object information IOM have about the indoor scene and is added by 20 from 90 to 150 sequentially selected from the vector. Plus, IOM-80 is the baseline accuracy. Obviously, as the number of object information increases, the much better scene recognition accuracy is achieved, e.g., on the Places365-14 dataset, the IOM-150 reaches 74.1% accuracy, which is **10.0%** higher than IOM-80. This improvement shows the number of object information is proportional to scene recognition accuracy. Moreover, the comparison experiments between the IOM and OM on three datasets, shows an average of **20.5%** improvements can be achieved. After analyzing the object categories of OM pretrained on the MS COCO, we observed there are only half of the object categories are related to indoor scenes. In contrast, the other half is related to outdoor scenes. Therefore, the relevance of object categories of object model with the scenes is essential for scene recognition, e.g., the information of elephant and giraffe in OM will not be valuable for indoor scene recognition. Similarly, the bus and train are not beneficial to indoor scene recognition.

2) Analysis of BORM: We conduct experiments for BORM and IOM in the reduced Places365-7 dataset, and results are displayed in Table II. The experiment results show the BORM and IOM model has an advantage over the OM and yields an average accuracy of 83.1% and 82.4%, surpassing the OM model about **20%** accuracy. The experiment result proves that with more object knowledge about the sur-

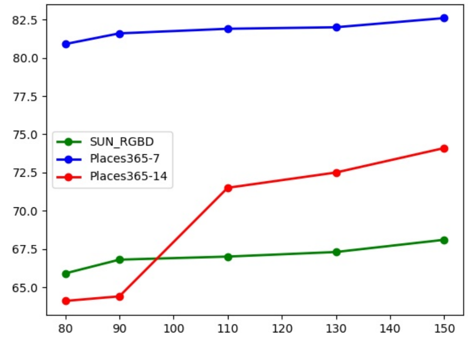


Fig. 6. Ablation study of improved object model (IOM) with different number of object knowledge ranging from 80 to 150 categories as shown in horizontal axis. The vertical axis shows the accuracy on percentage.

rounding environment, the greater scene recognition accuracy can be reached. Then, we test the model pretrained on the Places365-7 dataset on the SUN RGB-D test set, and similar conclusion can be drawn. Moreover, the BORM outperforms the IOM with 0.7% and 1.1% accuracy on the Places365-7 and SUN RGB-D dataset, respectively, which validates the knowledge of the co-occurrences between object pairs and their probabilistic relation forms an important indoor scene representation.

Similarly, in the Table III, we have conducted experiments on the reduced Places365-14 dataset, and experiment results show the IOM and BORM tremendously improves the performance over OM with **27%** accuracy. The results suggest the effectiveness of BORM and IOM over the OM especially when the number of scene classes of dataset is large.

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE REDUCED PLACES365-7 DATASET AND SUN DATASET OF SCENE RECOGNITION ACCURACY

Method	Config	Acc(Places365-7)	Acc(SUN)
PlacesCNN [20]	ResNet18	80.4	63.3
	ResNet50	82.7	67.2
Deduce [10]	Φ_{obj} (OM)	62.6	53.6
	Φ_{scene}	87.3	66.8
	$\Phi_{comb.}$	88.1	70.1
Ours	IOM	82.4	68.1
	BORM	83.1	69.2
	CBORM	**90.1**	**72.1%**

TABLE III

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE REDUCED PLACES365-14 DATASET OF SCENE RECOGNITION ACCURACY, THE * INDICATES THE RE-IMPLEMENT OF THE METHOD.

Method	Config	Acc
PlacesCNN [20]	ResNet18	76.0
	ResNet50	80.0
Word2Vec [5]	ResNet50+Word2Vec	83.7
*Deduce [10]	Φ_{obj} (OM)	47.0
Ours	IOM	74.1
	BORM	74.9
	CBORM	**85.8**

3) Performance Comparison: As shown in Table II, we conduct the ablation study of using only the BORM model and the CBORM model. We found the CBORM yield an average accuracy of **90.1%**, which greatly outperforms the ResNet18 and ResNet50 of PlacesCNN baselines about **10%** and **8%** respectively. Meanwhile, CBORM outperforms the BORM with 7% accuracy and 4% on the Places365-7 and SUN RGB-D dataset, respectively.

In comparison with the state of the art, Table II shows that the combined improves the scene recognition by **2.0%** in the reduce Places365-7 dataset. Also, as shown in Table III, the CBORM improve the recognition accuracy by **2.1%** in the reduced Places365-14 dataset. Both results show the combined model achieves comparable results to some recent approaches that use the word-embedding method to extract the semantic meaning of the environment, or use the combination of scene and object representations for better scene understanding. Surprisingly, Table II shows the performance of our method over the method in [10] with **2%**, showing the excellent generalization ability of CBORM over other methods on the Reduced SUN RGB-D dataset.

These results demonstrate that CBORM is successful in recognizing the scene images with a competitive accuracy. This improved effectiveness of CBORM over the state-of-the-art justifies our reasonable assumption that relation of object pairs is an essential complementary information for indoor scene recognition.

VI. CONCLUSION AND FUTURE WORK

In this paper, we aim to transfer object knowledge from humans to indoor scene recognition. First, we propose the IOM that with rich and relevant object categories for indoor scene representation. Besides, inspired by the nature of scene-common and scene-specific object pairs, we establish a BORM that obtains the probabilistic relations among object pairs given various scene, where the probability of scene-specific object pairs will be enhanced and the probability of scene-common object pairs will be suppressed. Moreover, we utilize the PlacesCNN model with ResNet as a backbone network for classification, which demonstrates a substantial result on the scene understanding. Hence, we combine the PlacesCNN and proposed BORM as CBORM for a more interpretable scene recognition algorithm, and experiment results show our proposed method significantly outperforms the state-of-the-art methods.

In the future, we plan to integrate our algorithm to real robots like mobile robots, and flying robots for the construction of the semantic map includes the label of the scene and detailed semantic information of the environment. The construction of a semantic map using the proposed scene recognition algorithm would be useful for navigation in unknown places for robots and humans because both the semantic label of the region and semantic meaning of the environment are provided. Furthermore, our system could be applied to autonomous robots and enabling them to assist humans in safety and rescue missions inside a house or a building.

References

[1] M. Liu, D. Scaramuzza, C. Pradalier, R. Siegwart, and Q. Chen, "Scene recognition with omnidirectional vision for topological map using lightweight adaptive descriptors," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 116–121.

[2] M. Liu and R. Siegwart, "Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 310–324, 2013.

[3] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] B. X. Chen, R. Sahdev, D. Wu, X. Zhao, M. Papagelis, and J. K. Tsotsos, "Scene classification in indoor environments for robots using context based word embeddings," in *2018 IEEE International Conference on Robotics and Automation (ICRA) Workshop*, 2018.

[6] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 4763–4770.

[7] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 1406–1413.

[8] X. Song, S. Jiang, B. Wang, C. Chen, and G. Chen, "Image representations with spatial object-to-object relations for rgb-d scene recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 525–537, 2019.

[9] C. Laranjeira, A. Lacerda, and E. R. Nascimento, "On modeling context from objects with a long short-term memory for indoor scene recognition," in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2019, pp. 249–256.

[10] A. Pal, C. Nieto-Granda, and H. I. Christensen, "Deduce: Diverse scene detection methods in unseen challenging environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4198–4204.

[11] H. Zeng, X. Song, G. Chen, and S. Jiang, "Learning scene attribute for scene recognition," *IEEE Transactions on Multimedia*, 2019.

[12] B. Miao, L. Zhou, A. Mian, T. L. Lam, and Y. Xu, "Object-to-scene: Learning to transfer object knowledge to indoor scene recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021.

[13] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," *Pattern Recognition*, vol. 74, pp. 474–487, 2018.

[14] H.-Y. Lin, C.-W. Yao, K.-S. Cheng *et al.*, "Topological map construction and scene recognition for vehicle localization," *Autonomous Robots*, vol. 42, no. 1, pp. 65–81, 2018.

[15] C. Ye, Y. Yang, R. Mao, C. Fermüller, and Y. Aloimonos, "What can i do around here? deep functional scene understanding for cognitive robots," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4604–4611.

[16] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person reidentification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 618–626.

[17] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1582–1596, 2009.

[18] H. Zhu, J.-B. Weibel, and S. Lu, "Discriminative multi-modal feature fusion for rgbd indoor scene recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2969–2976.

[19] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[20] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[21] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 2318–2325.

[22] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *International journal of computer vision*, vol. 107, no. 1, pp. 20–39, 2014.

[23] M. Brucker, M. Durner, R. Ambruş, Z. C. Márton, A. Wendt, P. Jensfelt, K. O. Arras, and R. Triebel, "Semantic labeling of indoor environments from 3d rgb maps," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1871–1878.

[24] X. Song, C. Chen, and S. Jiang, "Rgb-d scene recognition with object-to-object relation," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 600–608.

[25] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and Á. García-Martín, "Semantic-aware scene recognition," *Pattern Recognition*, vol. 102, p. 107256, 2020.

[26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[28] C. Yi, L. Zhou, Z. Wang, Z. Sun, and C. Tan, "Long-range hand gesture recognition with joint ssd network," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2018, pp. 1959–1963.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[31] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[32] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[34] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[35] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[37] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[38] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[39] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3d object dataset: Putting the kinect to work," in *Consumer depth cameras for computer vision*. Springer, 2013, pp. 141–165.

[40] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1625–1632.