Image Representations with Spatial Object-to-Object Relations for RGB-D Scene Recognition

Xinhang Song, Shuqiang Jiang* *IEEE Senior Member*, Bohan Wang, Chengpeng Chen, Gongwei Chen

***Abstract*—Scene recognition is challenging due to the intra-class diversity and inter-class similarity. Previous works recognize scenes either with global representations or with the intermediate representations of objects. In contrast, we investigate more discriminative image representations of object-to-object relations for scene recognition, which are based on the triplets of <object, relation, object> obtained with detection techniques. Particularly, two types of representations, including co-occurring frequency of object-to-object relation (denoted as COOR) and sequential representation of object-to-object relation (denoted as SOOR), are proposed to describe objects and their relative relations in different forms. COOR is represented as the intermediate representation of co-occurring frequency of objects and their relations, with a three order tensor that can be fed to scene classifier without further embedding. SOOR is represented in a more explicit and freer form that sequentially describe image contents with local captions. And a sequence encoding model (e.g., recurrent neural network (RNN)) is implemented to encode SOOR to the features for feeding the classifiers. In order to better capture the spatial information, the proposed COOR and SOOR are adapted to RGB-D data, where a RGB-D proposal fusion method is proposed for RGB-D object detection. With the proposed approaches COOR and SOOR, we obtain the state-of-the-art results of RGB-D scene recognition on SUN RGB-D and NYUD2 datasets.**

***Index Terms*—Scene recognition, object-to-object relation, sequential representations, RGB-D, object detection**

I. Introduction

THE goal of scene recognition is to annotate images with scene categories. Humans have innate talent to recognize those abstract scenes without hard training, while it is typically challenging for computer. For humans, some scenes (e.g., coast, mountain, and forest) can be directly distinguished by taking a glance [1], while some other scenes (e.g., bedroom, living room, dining room and classroom) may require to be distinguished from some local points of view such as objects and their relations. For computers, former scenes may be recognized by directly training the deep learning models with massive data (e.g., Places [2], [3]) under the supervision of global scene labels. While recognizing latter scenes usually requires more complete understanding of image contents (from the local point of view). The main components of scenes are objects, such as tree, rock, car, bed etc. [4]. Recognizing scenes by extracting object based representations is an intuitive way. However, the diversity of spatial layouts (of objects) and the object co-occurrences (between scenes) may lead to the intra-class difference and inter-class similarity of scenes, which limit the accuracy of scene recognition.

Previous works integrate object information for scene recognition from various aspects. Some works [5], [6], [7], [8] directly use object presence as the intermediate representations, which are fed to scene classifiers after feature quantification. Wang *et al.* [9] implement object detection to locate the positions (bounding boxes) of objects, where the local features are extracted inside the bounding boxes of objects. George *et al.* [10] propose to detect objects for active learning of fine-grained scene recognition. Some other works [11], [12], [4] separately train convolutional neural network (CNN) models on object-centric database (ImageNet [13]) with the supervision of object labels and on scene-centric database (Places [2]) with the supervision of scene labels, then features extracted from different types (object and scene) of CNN models are combined for scene recognition. These works mainly focus on the explicit object presence or hidden features (e.g., CNN activation trained in supervision of object labels) of objects. However, only representing images with object based features (as in those previous works) still cannot tackle the ambiguity caused by the object co-occurrences between scenes.

An alternative way is to integrate more discriminative components, such as object relations, into the representations of scenes. Also obtaining spatial relations between objects can somehow represent the spatial layout of the scenes, which is helpful to distinguish some particular scenes that are confused by object based representations. For instance in Fig. 1, different scenes such as *dining room* and *classroom* may contain similar objects such as table and chair. When only using the object presence to represent images, the object based features of those two categories are difficult to be distinguished. However, when considering the intermediate representations with spatial layout, those scenes can be distinguished, i.e., in *dining room*, the table is usually *surrounded* by the chairs, while in *classroom* the chairs are mostly *behind* the table.

One intuitive way to integrate spatial relations in intermediate representations is to implement object detection techniques [14], [15], [16] to simultaneously obtain object labels and the positions of their bounding boxes. Some previous works [17], [18], [9] have implemented object detection techniques for scene recognition. However, these works either only use the object labels without spatial information (in [17], [18]), or only locate the position of bounding boxes to extract local features (in [9]). None of these works have attempted to detect

Xinhang Song, Shuqiang Jiang, Chengpeng Chen, Gongwei Chen and Bohan Wang are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS) Institute of Computer Technology, CAS, Beijing, 100190, China
xinhang.song@vipl.ict.ac.cn, sqjiang@ict.ac.cn, {bohan.wang, chengpeng.chen, gongwei.chen}@vipl.ict.ac.cn.
*Corresponding author: Shuqiang Jiang.

(a) dining room

(b) classroom

Fig. 1. Scenes *dining room* and *classroom* with similar object presence of *tables* and *chairs*, but different spatial layout. Intermediate representations such as object presence may not be powerful enough to distinguish these two scenes, more discriminative information such as spatial relations are desired.

spatial relations between objects with both object labels and the bounding boxes, which are more discriminative to abstract scenes.

Moreover, RGB-D data is helpful to locate the objects. The low cost depth sensors, such as Microsoft Kinect, can capture RGB-D data, which extends traditional RGB recognition by including depth information. Depth camera can provide spatial information to detect object boundaries and understand the spatial layout of objects in scenes. Combining RGB with depth images to recognize scenes usually achieves better performance than only using RGB or depth images. In previous works, depth information is modeled using handcrafted features. Although automatic feature learning from the data with CNN can provide more discriminative representations, the lack of large enough RGB-D databases limits the complex CNN models to be suitably trained with RGB-D data. Recently, Song *et al.* [19] proposed SUN RGB-D, a larger scene RGB-D dataset that can be used to train more complex models for scene recognition [9], [20].

In this paper, we propose to detect object and their relations on RGB-D data for scene recognition. Particularly, the framework of two types object-to-object relation based representations are illustrated in Fig. 2. First, an intermediate representation of the co-occurrences of objects and their relations is proposed to exploit the spatial layouts of scenes, which is shown in the top row of Fig. 2. Particularly, the proposed intermediate representation is based on the co-occurring frequency of object-to-object relations (denoted as COOR), which is represented as a three order tensor ($\langle object, relation, object \rangle$)[1]. Compared to previous intermediate representations of object presence, COOR is more discriminative to those scenes with similar object distributions (see Fig. 1). And we show that COOR extracted from multimodal RGB-D data is helpful to obtaining powerful representations and complementing the global features.

Second, in contrast to previous intermediate representations (e.g., object presence and co-occurrences), a more discriminative and explicit type of representation, in a freer form of sentences of local captions, is proposed to describe images with richer semantic information, such as objects and their relations. Object-to-object relations are sequentially represented in our proposed explicit representations, denoted as SOOR. Without the limitation of fixed data structure in COOR[2], richer types of object-to-object relations are proposed for SOOR . The framework of scene recognition with SOOR is illustrated in the bottom row of Fig. 2. First, sentences of local captions (SOOR) are generated by a template which is typically designed for scene recognition. Then SOORs are encoded with sequential model, i.e., recurrent neural network (RNN). Finally, the encoded features are pooled into global features to feed the scene classifier (implemented as multi-layer perceptron, MLP). In order to obtain more accurate object labels and bounding boxes, object detection is implemented on the RGB-D data, where a multi-modal proposal fusion method is proposed for RGB-D object detection.

Compared to the preliminary conference paper of this work [21], the contributions of this paper can be summarized as: 1) we propose SOOR to represent objects and relations in a freer form of representations (sentences of local captions) to avoid the limitation of fixed data structure; 2) we propose richer types of relations, such as extended directional relations, distance and area, to provide more discriminative information ; 3) we propose a multimodal Faster-RCNN-RNN framework to detect and sequentially encode object-to-object relation for RGB-D scene recognition; 4) the proposed method obtains gains of performance on widely used RGB-D datasets, such as SUN RGB-D and NYUD2, outperforming the state-of-the-art.

II. Related work

A. Intermediate representation

Previous works mainly extract intermediate representations of semantic concepts, such as objects, attributes, or some types of hidden patterns, to describe images. Vogel and Schiele [22] proposed to represent natural scenes with regional intermediate representation of local concepts such as *water*, *rocks* or *foliage*. Similarly, Object bank [7] trains classifiers with multi-scale images from ImageNet to obtain a more descriptive representation. The *classemes* representation [8] is based on

[1]Representing images with COOR is first proposed in our preliminary conference paper [21] (referred to OOR in that work).

[2]For instance, obtaining COOR with 19 categories of objects and 16 types of relations results in the three order tensors in size of $19 \times 16 \times 19$ (5776 dimension after flattening), while only a few elements are nonzero (meaningful). Also the data structure (three order tensors) of COOR only supports the fixed types of relations, which limits the flexibility of COOR.

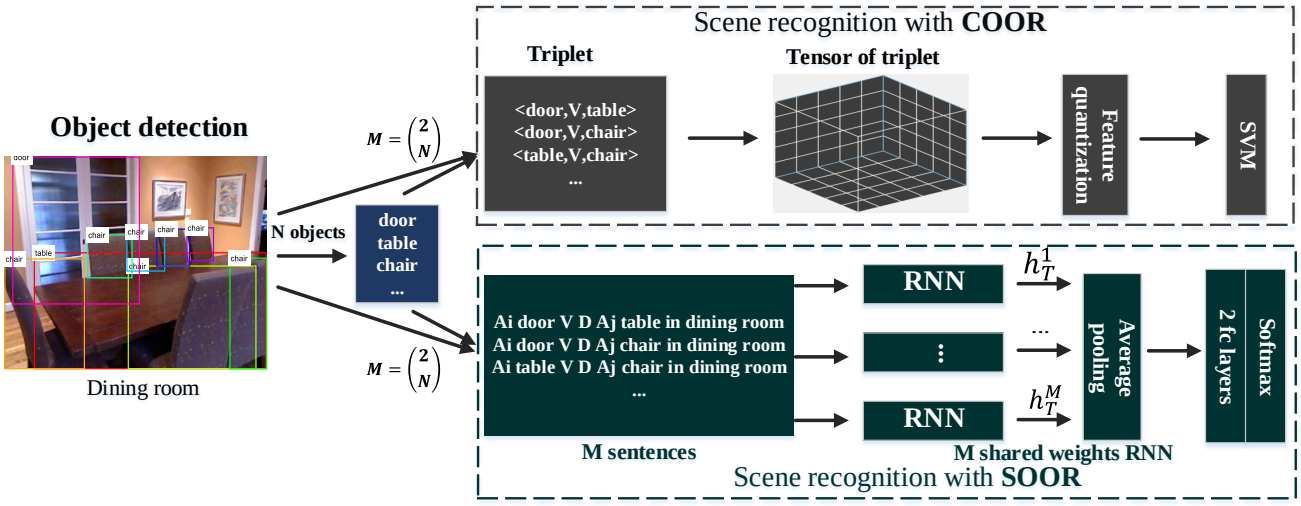


Fig. 2. Framework of scene recognition with COOR and SOOR. N objects are detected to generate $M = \binom{2}{N}$ triplets and sentences. The triplets are converted into COOR, and then fed to classifiers. All the sentences of (SOOR) are first encoded with RNN, and then pooled to the global vectors, which are finally fed to the scene classifier. Different types of relations, such as V, D, A, are introduced in subsection IV-A.

a set of fixed basis classes. Attributes [23], [24] follow a similar idea, where classifiers are trained to detect whether certain attributes are present or not. Attributes can be modeled at both local and global levels, and defined for both objects [23] and scenes [25]. Different types of CNN models are used to extract features, where different types of features are combined with different strategies, including scale-specific networks [4] or Fisher Vector (FV) encoding [11], [12]. Those types of intermediate representations are either extracted from local patches in dense grid or from global images. However, without precisely locating the object regions (e.g., with object detection), the extracted features may not be reliable enough for scene recognition.

B. Scene recognition with object detection

Some works [17], [18], [9] have implemented object detection technique for scene recognition. George *et al.* [10] propose to represent scene images by object distributions based on object detection, which is then optimized to distinguish fine-grained scenes by semantic clustering. Bappy *et al.* [17] combine object detection with manual annotation for active learning of scene recognition. Only representing images with object distributions lacks of spatial information. Wang *et al.* [9] extract local features located by the object detection, and local features are embedded with Fisher Vector.

In this paper, we adapt Faster R-CNN to RGB-D data by proposing multi-modal proposal fusion methods, where the accuracy of locating objects is obviously improved than only using RGB data. In addition to object detection, we further detect spatial relations to represent images for scene recognition. If the object presence (obtained by object detection technique) is regarded as first order representations of image content (e.g., objects), the proposed object-to-object relation based representations can be regarded as a mixture of first and second order representations, which are more discriminative than the scenes with object co-occurrences.

C. RGB-D recognition

Besides widely used CNN models, earlier works design handcrafted features for depth data, which captures depth-specific properties and avoids the requirements of large scale data. Gupta *et al.* [26] propose to detect contours for segmentation of depth images then the outputs of segmentation are quantified (as local features) into global features for scene classification. Banica *et al.* [27] first quantify local features with a second order pooling method, which are used for segmentation and scene classification. More recently, some architectures of multi-layered networks can be trained directly from large amounts of unlabeled data. Socher *et al.*[28] use a single layer CNN trained unsupervisedly on patches, and combined with a recurrent neural network (RNN).

More recent works rely on transferring and fine tuning the pretrained RGB CNN model (e.g., Places-CNN pretrained on RGB database, Places [2]) to depth data [29], [9], [20], [19]. Some approaches[20], [29] propose to incorporate CNN architectures by fine tuning with CNN architectures of two streams. Zhu *et al.* [20] fine tune the depth CNN models from the RGB CNN models by including a multi-model fusion layer, which simultaneously considers inter- and intra-modality correlations, meanwhile regularizing the learned features to be compact and discriminative. Rather than transferring or fine tuning, Song *et al.* [30] propose to train depth-specific CNN models from scratch in weak supervision. More than just using CNN models, Wang *et al.* [9] first implement object detection technique to obtain bounding boxes, and local CNN features are extracted inside the bounding boxes. Also, CNN features are extracted from images. Then CNN features of multiple modalities (RGB and depth) and multiple regions (local regions and images) are combined with a component aware fusion method. Rather than hidden features of CNN models, we also integrate object-to-object based representations, such as COOR and SOOR in the proposed framework.

D. Image captioning

The goal of image captioning is to describe images with natural languages. Generally, there are two branches of methods of generating image captions. One branch is the integrated models [31], [32], where words or concepts are first detected, and then integrated into sentences with some fixed templates. Another branch is the generated models [33], [34] that follow an similar CNN-RNN architecture. In generated models, CNN is regarded as "encoder", where the visual information are encoded to some feature vectors, and RNN is acted as "decoder", where the encoded features are sequentially decoded to the natural languages. More recently, dense captioning frameworks [35], [36] are proposed to generate local captions, which focus on the local regions of images. The frameworks of dense captioning are mainly inspired by Faster R-CNN [16], which first implement region proposal network (RPN) to obtain dense proposals. Then a multi-loss layer, consisting of a loss layer of the coordinates of bounding boxes and a loss layer of local captions implemented by RNN model.

These generated models require large amount of annotations of captions, while our proposed method requires annotation of objects, which are easier to be collected than captions. Also, previous works (such as [31], [32]) focus on generating (local) captions that are friendly to human understanding, while our proposed approach is designed to generate sentences of local captions (i.e., SOOR) to describe objects and their spatial layouts in the scenes, which benefits scene recognition by integrating richer and more discriminative components into the representations of scenes.

III. Intermediate Representations of Object-to-Object Relation

A. Co-occurring frequency of object-to-object relation

Common intermediate representations are mostly about object presence, which can be represented as $P_O^I = \left[p_1^I, p_2^I, \ldots, p_{|O|}^I\right]$ (see Fig. 4 (b)), where p_i^I is the appearance frequency of object i observed in the image I, and O is the object vocabulary. With such type of representation, the object co-occurrences between scenes may reduce the discrimination between those scenes, leading to the ambiguity to the classifying models. An alternative (more discriminative) way is representing images with statistical matrix of object-to-object co-occurrences, $P_{OO}^I = \begin{bmatrix} p_{11}^I & \cdots & p_{1|O|}^I \\ \vdots & \ddots & \vdots \\ p_{|O|1}^I & \cdots & p_{|O||O|}^I \end{bmatrix}$ (also see Fig. 4 (c)), where each element p_{ij}^I represents the appearance frequency of co-occurrence between object i and j.

However, only representing images with objects information (e.g., object presence or object co-occurrence) lacks considering of the spatial layouts of scenes, which may still bring ambiguity to the scenes with object co-occurrences. Thus, the proposed intermediate representation integrates objects and their spatial relations, which is represented as a type of triplet $\langle object,\ relation,\ object \rangle$, also denoted as co-occurring frequency of object-to-object relation (COOR) representation. The proposed COOR can be formulated as a three order tensor

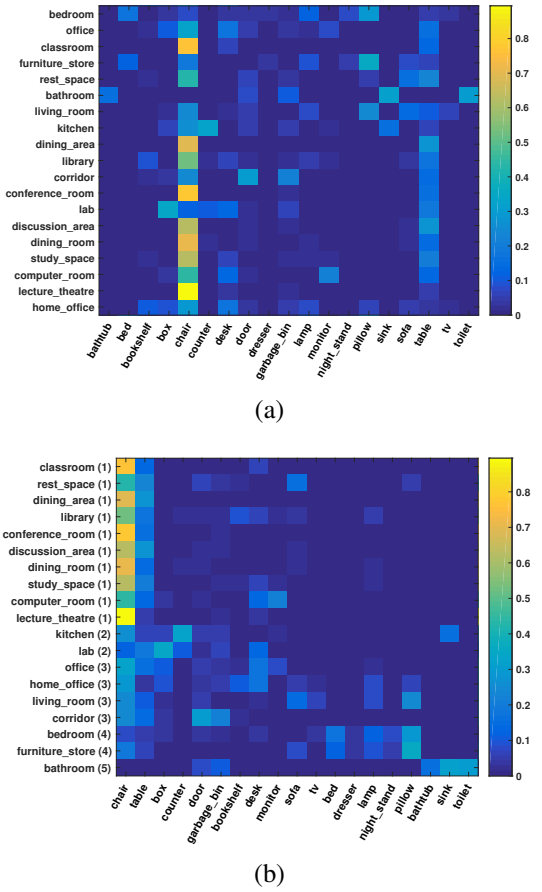


Fig. 3. Analysis of objects and scenes co-occurring, (a) correlation matrix, (b) correlation matrix after reordering with clustering, where the "(#)" represents the cluster ID.

$P_{COOR}^I \in R^{|O|\times|V_\alpha|\times|O|}$ (see Fig. 4d), where V_α is the vocabulary of directional relations between objects. Particularly, we define the relative relations based on the coordinates of object bounding boxes $b = [x_1, y_1, x_2, y_2]$, where $[x_1, y_1]$ are the coordinates of left-top corner, and $[x_2, y_2]$ represent the right-down corner. The relation between object i and j are represented as

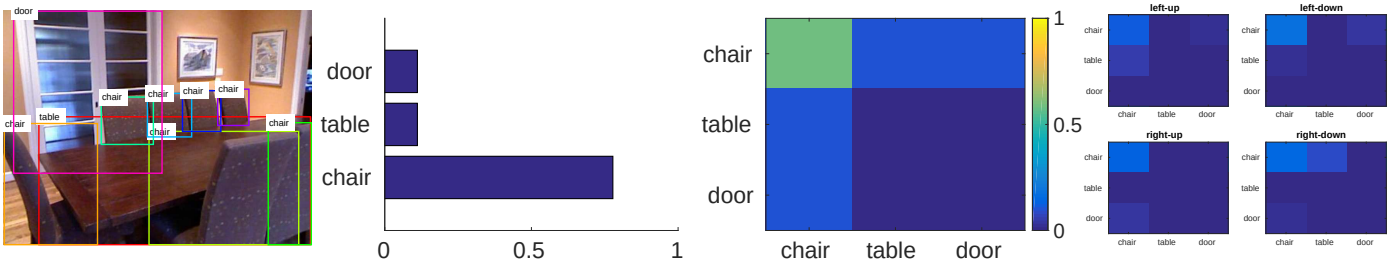
$$V_\alpha^{(i,j)} = \left[g\left(b^i - b^j\right)\right] = \left[g\left(x_1^i - x_1^j\right), g\left(y_1^i - y_1^j\right), g\left(x_2^i - x_2^j\right), g\left(y_2^i - y_2^j\right)\right] \quad (1)$$

,

where $g(x) = \begin{cases} 0, & if \quad x <= 0 \\ 1, & if \quad x > 0 \end{cases}$. There are $|V| = 2^4 = 16$ different types of spatial relations between objects.

B. Insights of using object-to-object relation

The key limitation of previous object based intermediate representations (e.g., object presence and object co-occurrences) is that the object co-occurrences between scenes may confuse the classifying models. In order to obtain some insights of the proposed COOR, we first analyze the affects of object co-occurring (between scenes) for scene recognition.



(a) "dining room" (b) P_O^I (c) P_{OO}^I (d) P_{COOR}^I

Fig. 4. Feature visualization of a toy example, (a) one image of "dining room" from SUN RGB-D, and the object annotations are from the ground truth, (b) the visualization of P_O^I, which is normalized by the counting of appearances, we select three object categories for this toy example, (c) the visualization of P_{OO}^I, the frequency of object co-occurring, (d) the visualization of P_{COOR}^I, which should be represented as 3D tensor. Since 3D tensor is hard to visualize in (2D) space, we slice it in dimension of "$relation$", obtaining feature maps of four types of relative relations (only using less types of relations for better visualization) , such as "left-up", "left-down", "right-up", "right-down". For instance, "left-up" represents the co-occurring frequency between objects in such relation. Note that (c) and (d) are the visualizations of feature, and for training classifiers, the features are stretched to vectors.

And then we compare the proposed COOR with other types of intermediate representations on the SUN RGB-D [19] database. This database contains 40 categories with 10335 RGB-D images. Following the publicly available split in [19], [9], 19 most common categories are selected, consisting of 4,845/4659 images for training/test. Also, 19 popular object categories are selected in [19] for the object detection, whose annotations are given with the bounding box coordinates and object labels.

1) Scene recognition with intermediate representations: We illustrate the statistics of the co-occurring (appearing in the same image) between objects and scenes in a correlated matrix W, which is visualized in Fig. 3 (a). Each element w_{so} in the matrix W is the co-occurring frequency between scene s and object o. With such matrix W and the object distributions P_O^I, the scene probability distribution can be obtained as:

$$P_S^I = P_O^I W^\intercal \tag{2}$$

where P_S^I can be used for predicting the scene label by finding the scene category with maximum probability. Also, the scene label can be predicted by training SVM [37] classifiers based on the representations of P_O^I, P_{OO}^I, and P_{COOR}^I.

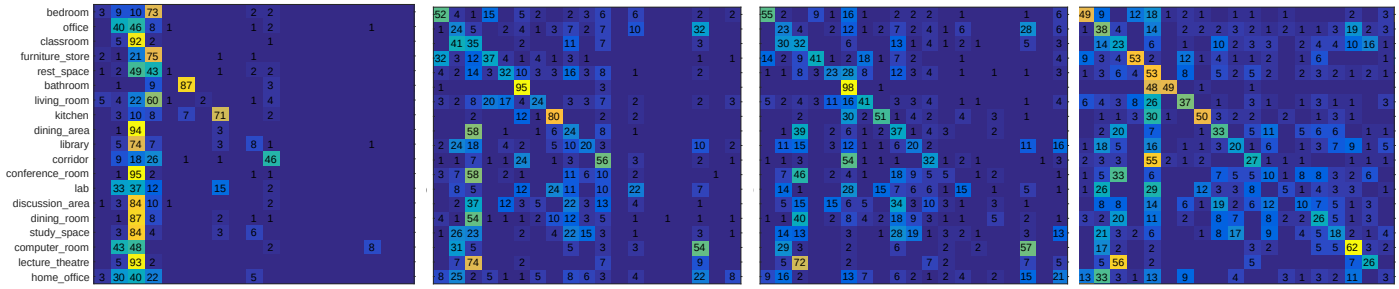
2) Comparison of different types of co-occurring based representations: Confusion matrices of different representations are compared in the Fig. 5, where the overall classes accuracy is calculated by the mean of diagonal values. Note that the accuracy of P_S^I (in Fig. 5 (a)) is much lower than others. Particularly, many scenes such as "dining area", "conference room" and "dining room" are misclassified to "classroom". This confused problem is mainly caused by similar object co-occurrences among these scenes. In order to better visualize such confusion between scenes, the correlation matrix in Fig. 3 (a) is re-organized with the spectral clustering [38] algorithm. Thus the closed categories (in the same cluster) can visualized in a closer way. The Fig. 3 (b) visualizes the correlation matrix after re-ordering the rows and columns. In practice, the number of cluster is $k = 5$, the scenes (rows) are reordered by the cluster ID, and the objects (columns) are reordered according to the clustered results of spectral clustering. It can be observed that many scene categories such as "classroom", "dining area", "dining room", and "conference room" are clustered together because of the co-occurring of objects like "chairs" and "tables" (see the first two columns in Fig. 3 (b)), this is also the reason of the confusion between scenes in Fig. 5 (a).

3) Insights of removing confusion with COOR: In addition to using Eq. 2 to predict scene labels, we also recognize scenes with SVM classifier with different types of intermediate representations of P_O^I, P_{OO}^I, and P_{COOR}^I. The confusion matrixes of using different types of intermediate representations are visualized in Fig. 5 (b)-(d). Note that the results of both Fig. 5 (a) and (b) are based on the same object based representation P_O^I, but using different classifiers, i.e., maximum probability (use Eq. 2) and SVM. The comparisons of them illustrate the efficiency of more discriminative classifier SVM. Moreover, using P_{COOR}^I with SVM (in Fig. 5 (d)) obtains even better performance than P_O^I and P_{OO}^I, including better overall classes accuracy and less confused problems in the confusion matrix. For instance, the confusion between "classroom" and "dining room" in Fig. 5 (d) is much better than that in Fig. 5 (a), where the rate of misclassifying from "dining room" to "classroom" gets lower from 87% to 20%. This also supports our hypothesis (in Fig. 1) that the scenes with object co-occurrences also can be distinguished by integrating the relative relations into the intermediate representation.

The above evaluations are based on the ground truth of annotated objects. Since the annotations of object are not available for the scene recognition in real world, the object detection technique is implemented to obtain the labels and bounding boxes of objects, which will be introduced in the Section V.

IV. Sequentially Encoding Object-to-object Relation for Scene Recognition

The insights of COOR illustrate the efficiency of integrating object-to-object relations into intermediate representations. One benefit of using intermediate representation is not requiring further encoding, COOR can be fed to the classifiers after flatting the tensors to vectors. However, using such data structure of COOR with a fixed size tensor to represent the co-occurring frequency of triplet $\langle object, relation, object \rangle$ also limits the extensibility of COOR. With the increasing



(a) Accuracy: 22.8% (b) Accuracy: 28.1% (c) Accuracy: 28.8% (d) Accuracy: 31.1%

Fig. 5. Confusion matrices, evaluated with annotated objects from ground truth, (a) P_S^I in Eq. 2, (b) P_O^I with SVM, (c) P_{OO}^I with SVM, (d) P_{COOR}^I with SVM. The horizontal and vertical coordinates are symmetric.

types of relations, the size of COOR tensor sharply expands, which limits COOR from exploring richer types of spatial relations for more complex scenes. Alternatively, the triplet of $\langle object,\ relation,\ object\rangle$ can be represented as a freer form of phrases or sentences, such as, "Object (A) Relation Object (B)". With such type of representation, external elements such as new type relations can be directly insert into the representations. For instance, more types of relations can be inserted as "Object (A) Relation (a), (b), ..., (x) Object (B)", where [Relation (a), (b), ...(x)] represents new types of relations, while it requires higher order or new tensors to insert new elements within COOR, while is cost consuming. Details of using more types of relations to generate representations of SOOR are introduced in the following section. In contrast to representing objects and their relations with the co-occurring frequency tensors in COOR, we propose to explicitly represent objects and their relations with sequential representations, such as phrases and sentences of captions. Thus, more types of relations can be sequentially listed in those sentences (captions).

In previous works, sequential representations such as image captions and local captions are generated to describe images, which are more friendly to human understanding. Particularly for scene recognition, we propose to generate sequential representations with the detected objects and relations using a fixed template, which are in form of the sentences of local captions. The generated local captions are then encoded with sequential model (e.g., RNN) to obtain sentence-level scene distributions (or hidden features). Since our task is to predict scene labels for the global images, those hidden features of local captions are then pooled to the global features for scene recognition. In our implementation, SOOR encoding and pooling are implemented in an end-to-end framework (see Fig. 2).

A. Richer types of spatial relations

In addition to the directional relations, more types of relations (such as extended directional relations, distances and area size of objects) are integrated in SOOR without explicitly increasing the size of features. Particularly, we focus on the spatial relations, which can somehow reflect the spatial layout of the whole scenes (after aggregation of each SOOR). We do not consider semantic relations, such as "riding on" and "playing against", since those types of relations require external detectors, which requires external annotations for training. Also the accuracy of those detectors of semantic relations may not reliable enough to improve the scene recognition.

1) Extended directional relations: With co-occurring based representations, some types of directional relations are defined in Eq. 1. Although those relations consider the relative directions between objects, while they do not consider another particular type of relations, i.e., overlaps between objects. Note that overlapping is also an meaningful type of relation between objects. In order to detect the relations of overlapping, we extend Eq. 1 as follows:

$$V_\beta^{(i,j)} = \left[g\left(x_1^i - x_2^j\right), g\left(y_1^i - y_2^j\right), g\left(x_2^i - x_1^j\right), g\left(y_2^i - y_1^j\right)\right] \tag{3}$$

where $V_\beta^{(i,j)}$ represents the cross relation between objects i and j. For instance, $V_\beta^{(i,j)} = [1,1,0,0]$ represents that there exists overlap between object i and j. By combining $V_\alpha^{(i,j)}$ and $V_\beta^{(i,j)}$, we obtain the extended relations $V^{(i,j)} = \left[V_\alpha^{(i,j)}, V_\beta^{(i,j)}\right]$. For instance, with $V_\alpha^{(i,j)} = [0,0,0,0]$ and $V_\beta^{(i,j)} = [0,0,1,1]$, obtaining $V^{(i,j)} = \left[V_\alpha^{(i,j)}, V_\beta^{(i,j)}\right] = [0,0,0,0,0,0,1,1]$, which means object i is on the left-up of object j and has overlap (at the right-bottom corner of object i and left-up corner of object j).

2) Distance : Besides the directional relations, we also consider the distance between objects as another type of relation. Distance is a relevant factor to reflect relation. The smaller distance (between objects) usually results closer relations. Particularly, we define 3 types of distance, including central distance, minimum distance and maximum distance.

Central distance measures the distance between the centers of the objects, which can be formalized as follows:

$$D_C^{(i,j)} = dist\left(center\left(b^i\right), center\left(b^j\right)\right) \tag{4}$$

where $dist\,(x,y)$ represents Euclidean distance between x and y, and $center\,(x)$ means the central coordinates of object bounding boxes, i.e., $center\left(b^i\right) =$

$\left[\left(x_1^i + x_2^i\right)/2, \left(y_1^i + y_2^j\right)/2\right]$. In addition to central distance, minimum distance $D_{\min}^{(i,j)}$ and maximum distance $D_{\max}^{(i,j)}$ are also considered. Similar to the directional relations, all these types of distance are quantified into binary codes by $f(x) = \begin{cases} 0, & if \quad x <= thres \\ 1, & if \quad x > thres \end{cases}$, resulting vectors $D = \left[D_C^{(i,j)}, D_{\min}^{(i,j)}, D_{\max}^{(i,j)}\right]$ of $|D| = 3$ dimensions.

3) Area : Besides relative relations, the attribute of the object area is also included in the generation of local captions. Similar to the above relations, the area of the objects $S\left(b^i\right)$ are quantified into binary codes with the length of two bits.

B. Generation of SOOR

Sequential representations (SOOR) are generated by the objects, their attributes and relations in a template, which are in a form of local captions. Particularly, the template of our SOOR is represented as follows: "Attribute(i) Object(i) Relation (V) Relation (D) Attribute(j) Object(j) in Scene". For instance, "*01* (small) *chair 00000011* (left-up, overlapped) *001* (closed distance) *11* (large) *table in dining room.* "

C. Sequentially encoding model

Since the proposed SOOR is generated in the form of sentences of local captions that represented in a grammatical order (it can be ensured because we generate local captions in a fixed template), it's intuitive to sequentially predict scene labels with RNN model (i.e., a sequential model). During the training, the first T words $\mathbf{x} = [x_1, \ldots, x_T]$, where $T+1$ is the length of words in each caption (T is the length of captions, except for the last word, scene label) are sequentially input to RNN model (we implement GRU unit [39]) to obtain the hidden activation $\mathbf{h} = [h_1, \ldots, h_T]$, RNN is formalized as follows:

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right) \tag{5}$$
$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \tag{6}$$
$$\bar{h}_t = \tanh\left(W x_t + U\left(r_t \odot h_{t-1}\right)\right) \tag{7}$$
$$h_t = \left(1 - z_t\right) h_{t-1} + z_t \bar{h}_t \tag{8}$$
$$y = \arg\max\left(\Phi\left(h_T\right)\right) \tag{9}$$

where σ is a logistic sigmoid function, $\odot$ is an element-wise multiplication, Φ is two fully connected layered neural network. The last element h_T of the hidden activations is then passed through two fully connected layers to predicate scene category y of input sequence (see Eq. 9).

Note that the sequential model can not only predict scene labels of local captions, but also obtain the probability vectors of scenes after softmax normalization of $\Phi\left(h_T\right)$. Particularly the predicted scene labels are used for the evaluation of sentence-level recognition, while the probability vectors of scenes are pooled into global vectors for the image-level recognition.

D. Local to global model of scene recognition

In order to predict scene labels for the global images, local captions (SOOR) are encoded into hidden activations, which are then pooled into the global features and fed to scene classifier in an end-to-end architecture. The bottom row of Fig. 2 illustrates the framework of proposed end-to-end architecture. For each image, a fixed number N of objects are detected, so a fixed number $M = \binom{2}{N}$ of local captions are correspondingly generated. All M local captions are used as input to RNN model to obtain M last hidden activations $[h_T^1, \ldots, h_T^M]$. After averagely pooling of M activations, the resulted vector $\hat{h}_T$ is fed to the MLP (consisting of two fully connected layers and one softmax layer) for classification of scene labels.

$$\hat{h}_T = \left(\sum_{i=1}^{M} h_T^i\right)/M \tag{10}$$
$$y = \arg\max\left(\Phi\left(\hat{h}_T\right)\right) \tag{11}$$

V. Multi-modal Object Detection and Features

In order to obtain spatial information in a more accurate way, the object detection technique is implemented on the RGB-D data, where depth data provides extra spatial (distance) information to complement RGB-D data. In particular, we adapt Faster R-CNN model [16] to the RGB-D data, and the RGB and depth modalities are combined by fusing the proposals of each modality.

A. RGB-D object detection

Being a region based method for object detection, Faster R-CNN includes a branch, named region proposal network (RPN), to generate candidates of bounding box. On each candidate region, the CNN hidden features are first extracted by region of interest (RoI) pooling layer, and then fed to the classifying module. The classifying module usually consist of two types of regressors, a softmax classifier for object labels and a regressor for the coordinates of bounding boxes.

In this work, we separately train two Faster R-CNN models on RGB and depth data, which are denoted as FRCN-RGB and FRCN-Depth, respectively, where ZF net [40] is used as pre-trained model in the training process. In addition to train separate models, both RGB and depth models are combined by connecting two branches of CNNs (i.e., RGB and depth branches). Particularly, RGB and depth are combined through merging of modality-wise proposals. First, two branches separately generate proposals by using RPN models (requiring FRCN-RGB and FRCN-Depth train RPNs separately). Then the region proposals of different modalities are merged together with our proposed RGB-D proposal fusion method, which is introduced in following subsection. The proposed RGB-D object detection model is denoted as FRCN-RGBD

B. RGB-D proposal fusion

With the RPN branches, the proposals $B_{rgb} = \left\{B_{rgb}^{(1)}, \ldots, B_{rgb}^{(n)}\right\}$ and $B_{depth} = \left\{B_{depth}^{(1)}, \ldots, B_{depth}^{(n)}\right\}$ are

obtained on each pair of RGB and depth images, where each $B^{(i)}_{rgb}$ and $B^{(i)}_{depth}$, $i = [1, \ldots, n]$, contain the proposal information in form of bounding box, including the coordinates $b^{(i)}_{rgb}, b^{(i)}_{depth}$ and confidence score $C^{(i)}_{rgb}$ and $C^{(i)}_{depth}$. Let $N^{(i)}_{rgb}$ and $N^{(i)}_{depth}$ be the number of proposals contained in $B^{(i)}_{rgb}$ and $B^{(i)}_{depth}$, respectively. Besides, we set

$$N^{(i)}_{rgb} = min\left\{\left|C^{(i)}_{rgb} > \alpha_{rgb}\right|, \lambda_{rgb}\right\}$$

$$N^{(i)}_{depth} = min\left\{\left|C^{(i)}_{depth} > \alpha_{depth}\right|, \lambda_{depth}\right\}$$

Where $\left|C^{(i)}_{rgb} > \alpha_{rgb}\right|$ denotes the number of proposals with value $C^{(i)}_{rgb}$ larger than α_{rgb}, which is considered as the confidence threshold, and λ_{rgb} is an empirical value of the maximal number of proposals, ensuring enough proposals. We consider $\alpha = a_{rgb} = \alpha_{depth}$ and $\lambda = \lambda_{rgb} = \lambda_{depth}$, since we do not obtain other prior of RGB and depth proposals. Note that these two types of hype-parameters α and λ decide the number of top $N^{(i)}_{rgb}$ /$N^{(i)}_{depth}$ proposals that are selected from B_{rgb}/B_{depth} according to C_{rgb}/C_{depth}. The selected proposals are denoted as B^S_{rgb}/B^S_{depth}.

After merging the proposals of B^S_{rgb} and B^S_{depth}, we obtain $B^S_{rgbd} = \left[B^S_{rgb} \cup B^S_{depth}\right]$. Note that it is unavoidable to lead to the overlapping between the proposals of B^S_{rgbd}, since B^S_{rgb} and B^S_{depth} are separately generated without process of avoiding overlap. However, the overlap between proposals may lead to heavily redundant features for the scene recognition. In order to avoid overlapping, the process of NMS is implemented to obtain the filtered proposals , which is formulated as follows:

$$B^S_{pooled} = NMS\left\{B^S_{rgb} \cup B^S_{depth}, \beta\right\}$$

where β is the threshold of Intersection-over-Union (IoU), which impacts the overlap between the resulted proposals. During NMS, the merged proposals B^S_{rgbd} are first ordered according to the confidence score of $C^S_{rgbd} = \left[C^S_{rgb}, C^S_{depth}\right]$, where C^S_{rgb}/C^S_{rgb} are the corresponding confidence scores of the selected proposals. Then for each pair of proposals with overlap (i.e., IoU between the proposals is larger than β), the proposal with smaller confidence score will be removed from the B^S_{pooled}.

After NMS, the hidden features of RGB and depth CNN (of layer conv5) are concatenated for the region of interest (ROI) pooling. After ROI pooling, the resulted features are fed to a module that consists of 2 fully connected layers, and two types of regressors, including object class score and coordinates (of bounding boxes) regressors.

C. Multi-modal COOR and SOOR

RGB-D fusion of COOR and SOOR are implemented with RGB-D object detection. First, multi-modal object proposals are obtained with RGB-D proposal fusion, then triplets of $\langle object, relation, object\rangle$ are obtained with RGB-D fused proposals, and RGB-D hidden features are concatenated to feed the Region-of-Interest (RoI) pooling within the bounding boxes of the object proposals, which are finally used to generate COOR and SOOR. Compared to single-modal, RGB-D multi-modal COOR and SOOR mainly rely on the proposed multi-modal proposal fusion.

VI. Experiments

A. Setting

1) Dataset: Our approaches are evaluated on two datasets: NYU Depth Dataset version 2 (NYUD2) [41] and SUN RGB-D[19]. The former consists of 27 indoor categories. Following the original training/test split of images 795/654 in [41], all 27 categories are reorganized into 10 categories, where some of categories with few images are combined into a joint category "other". The latter contains 10335 RGB-D images in 40 categories,. Following the public split in [19], [9], the 19 most common categories are selected, consisting of 4,845/4659 images for training/test. The split is provided in the toolbox of SUN RGB-D dataset. For the object detection evaluation, we follow the same split of scene recognition, since the object detection further serves for the scene recognition. All depth images are encoded to HHA images using the code in [42].

2) Evaluation metric: Following [19], [9], we report the average class accuracy for the scene recognition (i.e., mean accuracy overall classes). We follow the evaluation method of [43] to report the average precision (AP) for object detection. Detected results are considered to be true or false positives according to the overlap area with ground truth bounding boxes. To be considered a correct detection, the overlap area a_o between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 50% (i.e., $IoU = 0.5$) by the following evaluation metric:

$$a_o = \frac{area\left(B_p \cap B_{gt}\right)}{area\left(B_p \cup B_{gt}\right)}$$

B. Object detection

1) Implementation : In training of object detection model, both FRCN-RGB and FRCN-Depth use the ZF net [40] as the pre-trained model, and follow the empirical parameter setting of Faster RCNN.

In the implementation of object detection, we set $\lambda = 200$ for both FRCN-RGB and FRCN-Depth models, and set $\alpha = 0.3$, $\beta = 0.6$ following the empirical setting of Faster RCNN work. Note that this setting is only used for object detection. For various types of scene recognition tasks, such as extracting COOR and generating SOOR, we practically evaluate the parameters to obtain better performances. Depth data is represented as raw depth image, which is further encoded to HHA image. The detected spatial relations are in two-dimensional space, although implemented on multi-modal data.

.

TABLE I
OBJECT DETECTION AP (%) OF SUN RGB-D

Model	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage_bin
FRCN-RGB	34.4	63.2	39.8	12.5	43.9	42.2	20.3	30.7	30.0	40.0
FRCN-Depth	54.5	71.6	25.5	5.0	45.4	39.5	22.2	10.5	18.0	34.2
FRCN-RGBD	**57.5**	**75.6**	**44.2**	**17.7**	**49.6**	**48.9**	**25.4**	**33.6**	**40.2**	**49.2**
Model	lamp	monitor	night_stand	pillow	sink	sofa	table	tv	toilet	mAP
FRCN-RGB	38.5	34.3	39.2	33.0	46.9	39.5	34.6	23.2	74.5	37.9
FRCN-Depth	40.0	18.8	34.8	40.2	49.2	44.9	41.2	14.3	70.0	35.8
FRCN-RGBD	**53.0**	**44.0**	**47.6**	**48.6**	**61.1**	**50.3**	**43.2**	**35.2**	**81.7**	**47.7**

2) Category-wise evaluation: We evaluate the performance of object detection on SUN RGB-D dataset, and the comparisons between different models and different modalities are illustrated in Table I with the setting of $\lambda = 200$. Comparing between RGB and depth modalities, depth model works better on some object categories such as "bathtub", "bed", "chair", "pillow" and "table", that contain enough depth information in shapes, while works much worse than RGB model on that objects such as "door", "dresser", "monitor", "tv", that barely have depth in shapes (thin and flat in shape). However, with the RGB and depth fusion, the performances of all category are improved. And the overall result (mAP) of FRCN-RGBD outperforms RGB model with a large margin about 10%, which illustrates the effective of using depth data for object detection.

C. Scene recognition with COOR and SOOR

1) Evaluation of COOR:

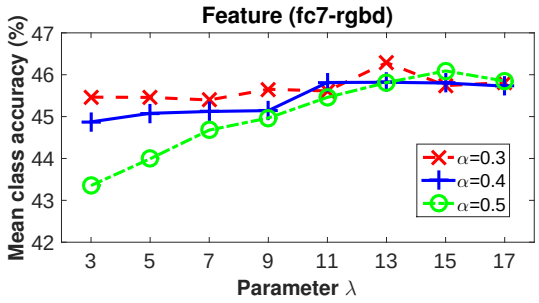


Fig. 6. Parameter evaluation of using object detection for scene recognition, evaluated on the SUN RGB-D dataset, with fc7 activation of proposed FRCN-RGBD.

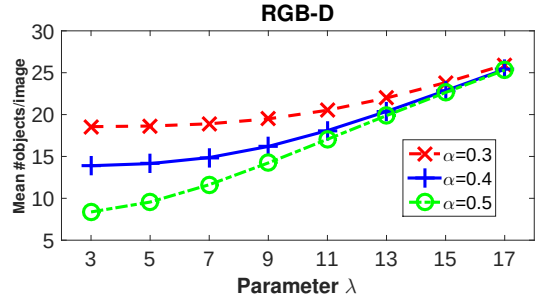


Fig. 7. Average number of objects detected from each image with different parameters α and λ on SUN RGB-D.

a) Parameter evaluation of COOR for scene recognition: For the scene recognition with COOR, the detected information of objects are used for generating COOR. We first obtain region proposals with RPN, then extract fc7 (last but one fully-connected layer) activation for each proposal, and finally the regional fc7 activation are combined to a global feature vector by max pooling on all proposals from one image. In order to speed up classification model training, the general dimension reduction method PCA is perform on $F_{fc-concat}$ to project the features to 512-dimensions. Note that, this dimension reduction process speeds up the classifier training barely with accuracy loss.

The evaluations of parameter α and λ are illustrated in Fig. 6. Note that, for the scene recognition, the larger λ results more overlapped area between proposals, redundant local features and proposals with lower confidence. Thus, we select much smaller λ for scene recognition. In Fig. 6, the smaller α and larger λ lead to better results, and best results are obtained when $\alpha = 0.3$ and $\lambda = 13$. The average number of detected objects of each image are illustrated in Fig. 7. For the larger α (e.g., $\alpha = 0.4$ or $\alpha = 0.5$), it shows that the average number of detected objects obviously increases with λ, which is also the reason of accuracy improvement in Fig. 6, particularly for $\alpha = 0.5$. Based on the selection of best results, we set $\alpha = 0.3$ and $\lambda = 13$ of object detection model for the rest experiments of scene recognition.

TABLE II
SCENE RECOGNITION ACCURACY (%) WITH INTERMEDIATE REPRESENTATION

Intermediate representations	RGB	Depth	RGB-D
P_S^I	16.8	13.9	17.8
P_O^I	31.4	26.5	31.9
P_{OO}^I	32.7	28.7	33.4
P_{COOR}^I	**33.5**	**30.0**	**36.3**

b) Comparisons of using COOR: We build the intermediate representations of different modalities based on the object detection results. The comparisons between different intermediate representations are illustrated in Table II. Note that, the main difference between the results in this table and Fig 5 is the source of object labels and their bounding boxes. The intermediate representations of RGB, depth and RGB-D in this Table II are based on the detected results of FRCN-RGB, FRCN-Depth and FRCN-RGBD, while the results in Fig. 5 are relied on ground truth annotations. Compared with the intermediate representations in Fig. 5, the P_S^I are much worse due to the lack of confidence for the object detection

results. However, the other results of P^I_O, P^I_{OO} and P^I_{COOR} outperform the ones in Fig. 5, especially the ones built on the detected results after RGB-D fusion, obtaining the best result with 36.3%. The advantage of the proposed COOR (with detected bounding boxes) is covering (and extracting) more information (object-to-object relations) in scenes, compared to COOR with annotated bounding boxes. Comparing with different representations, the P^I_{COOR} achieves best results in each modality, which outperforms P^I_{OO} from 0.8% (RGB) to 2.9% (RGB-D), where the main improvement benefits from the depth data.

TABLE III
COMPARISON RESULTS OF GENERATING LOCAL CAPTIONS WITH DIFFERENT TYPES OF RELATIONS ON SUN RGB-D IN ACCURACY (%)

OOR	Relations	Accuracy (%)			
		G.T.	Detected		
			RGB	Depth	RGB-D
SOOR	V	37.0	36.0	33.2	37.8
	VD	**38.5**	37.8	**35.9**	**39.9**
	VDA	37.8	**37.9**	**35.9**	39.5
TF/IDF	VDA	31.3	33.5	31.4	35.4
Word2Vec[44]	VDA	30.7	32.0	29.4	33.3
COOR	V	31.1	33.5	30.0	36.3
COOR+SOOR	V	**39.1**	**38.7**	35.8	**40.9**

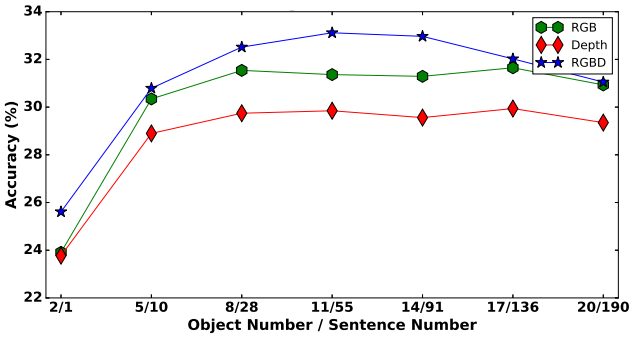


Fig. 8. Scene recognition accuracy with different number of sentences (in x axis, the left number of objects N, and the right one is number of sentences M, i.e., N/M) in each image.

2) Evaluation of SOOR:

a) Evaluation of using SOOR: We evaluate the factor of object and sentence amount of SOOR to the scene recognition, the results are illustrated in Fig. 8. The x axis indicates the number of #(detected objects) /#(generated sentences) of SOOR, represented as N/M, where $M = \binom{2}{N} = N \times (N-1)/2$.

The accuracy of recognition is illustrated in Fig. 8. The best result is obtained with the setting of ($N = 11/M = 55$), suggesting that too few objects may lack of information and too many objects may bring noisy. Thus, a suitable setting of objects leads to the best results, which is used for the rest of experiments.

b) Comparison of using local captions: The comparison results of generating local captions with different types of relations are illustrated in Table III. Particularly, the sentences of local captions (of SOOR) generated by the ground truth (including scene labels, object labels and bounding boxes) are also included in comparison. Recognizing with VD relations (directional relation and distance) based on SOOR obtains best results with ground truth. Also object detection technique is implemented on different modal (RGB, depth and RGB-D) data to generate SOOR, and using RGB-D data achieves best performance in general. Comparing with different types of relations, VD obtains best results with depth and RGB-D data, while VDA obtains best results at RGB data. It's interesting that the SOOR generated by the detected results (particularly on RGB-D data) obtain even better performance (about 1.4%) than local captions generated by ground truth. Note that the ground truth is more reliable, however the amount of ground truth annotations is limited (about $N = 5/M = 10$), which may not represent enough information of the scenes comparing to the detected results $N = 11/M = 55$. Comparing to COOR, recognizing with the SOOR obtains a gain of 3.6% (RGB-D) in accuracy. In addition to RNN encoding, we also include different types encoding methods of SOOR for comparison in Table III, including term frequency, inverse document frequency (TF/IDF) and word to vector (Word2Vec). It can be observed that encoding SOOR with RNN outperforms other encoding methods with a large margin in accuracy.

D. Multi-feature fusion

We evaluate different types of feature fusion to improve the performances of scene recognition, the fusions are categorized in the several aspects.

1) Local and Global features: In addition to COOR and SOOR, we also directly extract CNN hidden features of region (Local) and images (Global). The local features are extracted with fine tuned FRCN models (output of last fc layers). The global features are extracted with CNN models, such as D-CNN [30] for depth and Places-CNN [2] for RGB.

TABLE IV
SCENE RECOGNITION ACCURACY (%) WITH CNN ACTIVATION

Feature	RGB	Depth	RGB-D
Local	43.5	40.0	46.3
Global	41.4	41.1	51.5
Global+Local	46.6	41.1	52.6
Local+COOR	45.8	41.8	50.3
Local+SOOR	46.5	42.8	51.6
Global+COOR	44.9	42.0	52.8
Global+SOOR	46.8	43.5	53.5

2) Comparison of different types of feature fusion: Different types of feature fusion are compared in Table IV. When concatenating with local features, COOR and SOOR outperform local features with 2.3%/3.0% on RGB and 1.8%/2.8% on depth. Note that "Local", COOR and SOOR can all be regarded as local features, which consist of object based information. Thus, it suggests that the gains of concatenating "Local" with COOR and SOOR mainly benefit from the higher order relations. When concatenating global features with COOR and SOOR, we obtain the gains of 3.5%/5.4% on RGB and 0.9%/2.4% on depth, and obtain the best result with Global+SOOR. Compared with Local, SOOR is more suitable

to be combined with global features. When combining RGB-D models, each type of feature (single or fused) improves with large margins.

TABLE V
COMPARISON RESULTS ON SUN RGB-D DATABASE

Models		Accuracy (%)		
RGB	Depth	RGB	Depth	RGB-D
Baseline				
Places-CNN	Places-CNN	41.4	38.7	46.9
-	D-CNN	-	42.4	-
Proposed				
COOR	COOR	33.5	30.0	36.3
G+L+COOR	G+L+COOR	48.3	42.6	**54.0**
SOOR	SOOR	37.9	36.0	39.9
G+L+SOOR	G+L+SOOR	**50.5**	**44.1**	**55.5**
State-of-the-art				
DMFF [20]		36.5	40.4	41.5
Places-CNN + R-CNN [9]		36.5	40.4	48.1
MSMM [45]		41.5	40.1	52.3
RGB-D-CNN [30]		41.5	42.4	52.4

G: global, L: local

TABLE VI
COMPARISONS ON NYUD2 IN ACCURACY (%)

Models		Accuracy		
RGB	Depth	RGB	Depth	RGB-D
Baseline				
Places-CNN	Places-CNN	53.4	51.8	59.5
-	D-CNN	-	56.4	-
Proposed				
COOR	COOR	45.1	40.9	48.6
G+L+COOR	G+L+COOR	62.6	59.8	66.9
SOOR	SOOR	45.9	47.4	50.6
G+L+SOOR	G+L+SOOR	**64.2**	**62.3**	**67.4**
State-of-the-art				
Places-CNN + R-CNN [9]		-	-	63.9
MSMM [45]		-	56.4	65.8
RGB-D-CNN [30]		-	-	66.7

G: global, L: local

E. Comparison to the state-of-the-art

1) SUN RGB-D: Different types of features are combined to compare with other state-of-the-art works [19], [20], [9], [30]. The comparison results are illustrated in Table V. Some works [19], [20] only extract global features, while Wang *et al.* [9] propose to extract both global features and local features for scene recognition. Although the work of [30] does not explicitly extract local features, the depth model (D-CNN) of that work is trained based on local patches with weak-supervision. Thus, local information is implicitly included in that work [30]. Based on the object detection results, by concatenating the local feature and COOR representation, our method Local+COOR obtains 50.3% and outperforms the work of [9] more than 2%. Note that the proposed Local+COOR only extracts the local features (COOR is also a kind of local features), while [9] combines both local and global features. Although the single global feature in Table IV does not work as well as that in [30], by combining local and global features and concatenating with COOR, our Global+Local+COOR achieves the state-of-the-art result with 54.0%, outperforming [30] with 1.6%. Note that we take fc7 (with 4096 dimension) activation as the global feature, in contrast to the fc8 (with 19 dimension) activation in [30]. Moreover, with more types of relations, combining our proposed SOOR and global features (i.e., Global+OOR) obtains the accuracy of 55.5%, which is the state-of-the-art result of SUN RGB-D database to the best of our knowledge.

2) NUYD2 : We do not train particular CNN models for NYUD2 dataset, since this dataset contains much less data than SUN RGB-D. We mainly fine tune the pretrained models of SUN RGB-D to the NYUD2. Particularly, the Faster RCNN models of object detection of SUN RGB-D is directly applied to this NYUD2 dataset. Since using RGB-D features obtains much better performance, we mainly report results of RGB-D features of COOR, SOOR, Global+Local+COOR and Global+Local+SOOR. Comparing to other works [46], [9], [30], [45], our proposed Global+Local+COOR obtains better performance of 66.9%. In addition to COOR, our proposed method Global+Local+SOOR obtains the best results on all three modalities. Among all the modalities, we obtain the state-of-the-art result of 67.4% with RGB-D data.

VII. CONCLUSION

Object co-occurrences are unavoidably appeared between different scenes, representing images with object based intermediate representation may result in ambiguity for scene recognition. By analyzing the limitation of the previous intermediate representations of objects, we propose two types of more discriminative representations, including co-occurring frequency of object-to-object relation (COOR) and sequential representation of object-to-object relation (SOOR) consisting of objects and their spatial relations. First, COOR is proposed as a novel type of more discriminative intermediate representation, which is represented as three order tensors that calculate the co-occurring frequency of the triplets of $\langle object, relation, object \rangle$. Then, SOOR is generated in a form of sequential representations. Without the limitation of fixed data structure, richer types of relations such as extended directional relations, distance, and area are detected to be represented in SOOR. In order to better model the spatial information, both COOR and SOOR are built on RGB-D data. And the depth data is shown to be helpful for both object detection and scene recognition tasks, especially for the objects with depth in shape.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals, in part by the National Postdoctoral Program for Innovative Talents under Grant BX201700255, and in part by China Postdoctoral Science Foundation under Grant 2018M631583.

REFERENCES

[1] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *Journal of Vision*, vol. 7, no. 1, p. 10, 2007.

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., 2014, pp. 487–495.

[3] B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2723009

[4] L. Herranz, S. Jiang, and X. Li, "Scene recognition with cnns: Objects, scales and dataset bias," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 571–579.

[5] L. Li, H.Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *NIPS*, 2010.

[6] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*, 2010.

[7] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vision*, vol. 107, no. 1, pp. 20–39, 2014.

[8] A. Bergamo and L. Torresani, "Classemes and other classifier-based features for efficient object categorization," in *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2014.

[9] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Modality and component aware feature fusion for rgb-d scene classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] M. George, M. Dixit, G. Zogg, and N. Vasconcelos, "Semantic clustering for robust fine-grained scene recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 2016, pp. 783–798. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_47

[11] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos, "Scene classification with semantic fisher vectors," in *CVPR*, 2015.

[12] M. D. Dixit and N. Vasconcelos, "Object based scene representations using fisher scores of local subspace projections," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2811–2819.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0816-y

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[15] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] 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 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[17] J. H. Bappy, S. Paul, and A. K. Roy-Chowdhury, "Online adaptation for joint scene and object classification," in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 227–243.

[18] M. George, M. Dixit, G. Zogg, and N. Vasconcelos, "Semantic clustering for robust fine-grained scene recognition," in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 783–798.

[19] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, Jun. 2015, pp. 567–576.

[20] H. Zhu, J.-B. Weibel, and S. Lu, "Discriminative multi-modal feature fusion for rgbd indoor scene recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[21] X. Song, C. Chen, and S. Jiang, "RGB-D scene recognition with object-to-object relation," in *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 2017, pp. 600–608. [Online]. Available: http://doi.acm.org/10.1145/3123266.3123300

[22] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *Int. J. Comput. Vision*, vol. 72, no. 2, pp. 133–157, Apr. 2007.

[23] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. on Image Process.*, vol. 36, no. 3, pp. 453–465, 2014.

[25] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vision*, vol. 108, no. 1-2, pp. 59–81, 2014.

[26] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-014-0777-6

[27] D. Banica and C. Sminchisescu, "Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images," in *CVPR*, 2015.

[28] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *NIPS*, 2012.

[29] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[30] X. Song, L. Herranz, and S. Jiang, "Depth cnns for RGB-D scene recognition: Learning from scratch better than transferring from rgb-cnns," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 2017, pp. 4271–4277.

[31] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg, "Babytalk: Understanding and generating simple image descriptions," *TPAMI*, 2013.

[32] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions." *TACL*, vol. 2, no. 10, pp. 351–362, 2014.

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[34] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," 2015.

[35] A. K. J. Johnson and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] L. Yang, K. Tang, J. Yang, and L.-J. Li), "Dense captioning with joint inference and visual context)," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*, Jul 2017.

[37] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[38] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '01. New York, NY, USA: ACM, 2001, pp. 269–274. [Online]. Available: http://doi.acm.org/10.1145/502512.502550

[39] K. Cho, B. van Merri?nboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 06 2014.

[40] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Cham: Springer International Publishing, 2014, pp. 818–833.

[41] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 746–760. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33715-4_54

[42] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*, 2014.

[43] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes

Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[44] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[45] X. Song, S. Jiang, and L. Herranz, "Combining models from multiple sources for RGB-D scene recognition," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2017, pp. 4523–4529. [Online]. Available: https://doi.org/10.24963/ijcai.2017/631

[46] S. Gupta, P. Arbelaez, R. Girshick, and J. Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *Int J Comput Vis*, vol. 112, pp. 133–149, 2014.