



Place perception from the fusion of different image representation

Pei Li^a, Xinde Li^{a,b,*}, Xianghui Li^a, Hong Pan^a, M.O. Khyam^c, Md. Noor-A-Rahim^d, Shuzhi Sam Ge^e

^a School of Automation, Key Laboratory of Measurement and Control of CSE Ministry of Education, Southeast University, Nanjing, China

^b School of Cyber Science and Engineering, Southeast University, Nanjing, China

^c School of Engineering and Technology, Central Queensland University, Melbourne, VIC, Australia

^d School of Computer Science and IT, University College Cork, Cork, Ireland

^e Department of Electrical and Computer Engineering, Interactive Digital Media Institute, National University of Singapore, Singapore

ARTICLE INFO

Article history:

Received 8 January 2020

Revised 7 September 2020

Accepted 23 September 2020

Available online 24 September 2020

Keywords:

Indoor place perception

CNN

LSTM

Convolutional auto-encoder

Natural language

ABSTRACT

Inspired by the human way of place understanding, we present a novel indoor place perception network to overcome: 1). the simplicity of existing methods that only use the image features of object regions to recognize the indoor place, 2). insufficient consideration of the semantic information about object attributes and states. By utilizing multi-modal information containing the image and natural language, the proposed method can comprehensively express the attributes, state, and relationships of objects which are beneficial for indoor place understanding and recognition. Specifically, we first present a natural language generation framework based on a Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM) to imitate the process of place understanding. Next, a Convolutional Auto-Encoder (CAE) and a mixed CNN-LSTM are proposed to extract image features and semantic features, respectively. Then, two different fusion strategies, namely feature-level fusion and object-level fusion, are designed to integrate different types of features and features from different objects. The category of the indoor place is finally recognized based on fused information. Comprehensive experiments are conducted on public datasets, and the results verify the effectiveness of the proposed place perception method based on linguistic cues.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Place perception is one of the essential issues in the artificial intelligence field, which is mainly applied to image retrieval and understanding, or autonomous robot and drone. In the last few years, the related researches [1–3] mainly focused on the concept of scene recognition or place classification using approaches from the perspective of image recognition and understanding. However, most of these methods have some common drawbacks: (1) they address the place recognition as a simple image recognition problem without considering particular characteristics of place perception, (2) these works usually ignore the high-level semantic information of objects and humans in a place image, which caused an inadequate feature representation for place perception.

There are two fundamental differences between place perception and scene recognition or place classification. First of all, the

concept of the place refers to a space area in which people engage in specific activities, and this space area is abstracted in people's mind according to certain clues and can be represented by symbol labels [1]. Whereas for the concept of the scene, in some research on high-level scene perception [4], it is typically defined (often implicitly) as a semantically coherent view of a real-world environment comprising background elements and multiple foreground objects arranged in a spatially specific manner. In comparison, the conceptual scope of the scene is broader than the place. From a conceptual point of view, some places (e.g. the indoor place like a bathroom, the outdoor place like a garden or the public place like a cafeteria) can be regarded as scenes. However, not all scene (such as village, house, and river) can be classified into places. This difference is mainly because a place should be closely related to demands and actions or states of humans, but a scene is just an aggregation of some objective entities without specific functions. In another viewpoint, exploiting object entities as a clue to recognize the category of a place implicitly considers the concept of the place, that is, a place area should contain specific objects with human-related functionality, which can be regarded as

* Corresponding author at: School of Automation, Key Laboratory of Measurement and Control of CSE Ministry of Education, Southeast University, Nanjing, China.

E-mail address: xindeli@seu.edu.cn (X. Li).

prior knowledge. Several related works [5–9] studied from the aspect of the category distribution of objects contained in a place image. However, merely considering the object category ignores much valuable information, especially in a complicated place containing many different types of objects. For example, “a man lies in bed and works on a laptop” can roughly infer that this place might be a bedroom rather than an office, or “a stainless steel sink” might appear in the kitchen rather than in the bathroom. More broadly, the relationship between people and objects, states or actions of humans as well as the object attributes play essential roles in the prediction of the place category. However, these clues are kinds of high-level semantic information and hard to be formulated in an explicit form so that they cause difficulties for place perception.

The second difference is between the process of perception and recognition. We argue that the procedure of the place perception in the human brain follows typically three steps: (1) try to analyze the visual information and detect the critical objects in an input image, (2) use abstract symbols to describe the seen objects and organize symbols into semantic knowledge, (3) try to combine this knowledge to infer the possible place and generate some ideas. Thus, perception is not only to identify the category of a place but also to fully understand the containing information, e.g. symbol attribute, spatial attribute, and semantic attribute. More specifically, the perception process can also answer some semantic questions such as what kind of objects are present, what is happening in this place, or even what are the critical elements to the inference of the place category. These are fundamental requirements for service robots and human-computer interaction applications. However, current work merely addressed the label attribute problem of place perception without considering more semantic information required in the understanding process. Therefore, most methods regard it as a pattern recognition problem and use the classification model to solve it according to the type of information source and its features.

Considering those above two primary differences between place perception and scene classification, we intuitively believe that natural language information might become a valid clue for modeling feature representation in place perception. As a kind of symbolic representation form, natural language information is more abstract and suitable for human understanding, which can also cover more representative visual clues using non-redundant contents. Hence, linguistic information can adequately express the attributes, states, and relationships of objects in a place image. Additionally, although natural language can represent high-level semantic information, it also makes the information more ambiguous. Therefore, place perception can not solely rely on natural language and still needs visual information to assist in the understanding process.

The proposed method intends to mimic the human way of place perception by employing the technology natural language generation for obtaining more abundant and complementary semantic cues, which is to boost the perception performance of the current method. The semantic cue has been widely considered in the recent place recognition technology. For example, Cheng et al. [8] proposed a feature representation method from the perspective of object detection, called semantic descriptor with objectness (SDO), to get the distribution of shallow semantic descriptions between object and scene. Furthermore, Lopez-Cifuentes et al. [10] obtained the scene semantic representation by leveraging on semantic segmentation and then combined them with image features through a multi-modal CNN attention module. To eliminate the similarity among different scene categories, related studies change the core idea from discriminating single image information to fusing the semantic information of objects. However, these similar methods still rely on simple semantic elements and do not adequately consider the semantic attributes of objects.

Therefore, to the best of our knowledge, this is the first work that using the image caption technology to convert the visual modality into its relevant linguistic modality and combines both cues for place perception.

The contribution of our work is fourfold:

- (1) We propose a multi-task deep neural network to realize the indoor place understanding and recognition together, which imitates and learns the process of place perception in a human-style.
- (2) From the perspective of multi-modal information transformation and complementation, we propose an image captioning model to automatically generate natural language descriptions from place images, which is an additional information source to assist the decision-making in place recognition.
- (3) We propose a multi-modal feature extraction and fusion architecture based on a mixed-CNN-LSTM network that gathers both visual and linguistic features corresponding to instance-level and concept-level information, respectively.
- (4) We validate the effectiveness of the proposed strategy of using natural language descriptions to place perception through experiments on public image datasets, including Visual Genome, SUN397, MIT-67, and Places2.

This paper is organized as follows: [Section 2](#) provides an overview of related work in indoor place perception, with a focus on place perception and image caption based on deep learning methods. In [Section 3](#), we elaborate on the proposed approach to place perception. [Section 4](#) describes the implementation details of our algorithm and reports the experimental results. Finally, we conclude our work in [Section 5](#).

2. Related work

In this section, we briefly review two related topics: (1) place perception, and (2) image captioning. Specifically, we pay much more attention to recent works based on deep learning.

2.1. Place perception

Place perception in this paper contains two parts, namely, place understanding and place recognition. As explained in [Section 1](#), there is a difference between two concepts, and related studies have only explored the second part of them. In some current researches, place perception is often addressed as a scene classification/recognition problem and dealt with image classification methods.

According to the characteristics of the modality, there are usually two solutions to identify places, including image modality and multi-modal information. For image modality only, earlier studies are conducted in the direction of image classification, retrieval [11,12] and clustering [13,14]. These methods use artificially designed image features combined with a classifier for place image recognition. For example, Juneja et al. [15] used the Histogram of Oriented Gradients (HOG) feature encoded by the Fisher vector to describe the image regions. Similarly, Aziz et al. [16] constructed feature vectors by integrating Local Quinary Patterns (LQP), Bag of Visual Words (BoW) and HOG. In [7], the object-based probabilistic distribution was established by extracting Speeded Up Robust Features (SURF). They all applied Support Vector Machines (SVM) to realize classification. However, the recognition performance of these methods is unsatisfactory because the extracted features lack discrimination when facing complex place images.

On the other hand, due to the development of the deep neural network, various works try to exploit complex networks to learn

the image features from a large amount of data. Two types of features, including holistic features and local features of the image, are learned in these methods. In holistic feature representation, it regards an scene image as an integrated entity and extracts its global features. A comprehensive comparison of scene classification in [3] tested different network architectures on a dataset with more than 10 million images. These approaches are easy to be designed as end-to-end models, and their improvement are also benefited from the network structure itself. However, in global features based methods, the recognition accuracy of indoor scenes usually is lower than that of outdoor scenes. This poor performance is because indoor places are more complicated than outdoor places in most cases, both from the aspects of spatial structure and containing objects and stuff. Therefore, there are two directions for improvement. The first is to design additional modules to increase the complexity of image feature representation, thereby optimizing the discriminant performance. For example, Xie et al. [17] added a Fisher vector function into VGG-19. Pan et al. [18] employed two ResNets to extract foreground and background features of images. In [19], three different kinds of networks were applied for extracting features of object regions and whole images. The second is to focus on the local feature representation of elements constituting the scene image. In local feature representation, some studies represent places by using object distribution or adding prior knowledge into their models. They either designed feature descriptors manually [2,6] or automatically extracted features using deep neural network (DNN) [9,20]. It is worth noting that the object information plays a significant role in place recognition, whether from the aspect of network structure or feature selection. Thus, many methods base multi-modal models on the semantic information of objects.

For place recognition by multi-modal approaches, some researches try to enhance recognition performance by introducing additional information source like depth cue [21,22]. When there is only image modality, the main idea is utilizing linguistic cues generated from local regions of the image. Since the semantic information cannot be obtained directly, it is necessary to employ other networks for extracting semantic features. For example, In [23], the YOLOv2 method was used to generate Spatial-layout-maintained Object Semantics Features (SOSF) by Spatial Fisher Vectors (SFV). Similarly, Wang et al. [24] used VGG-19 networks to calculate Vectors of Semantically Aggregated Descriptors (VSAD). Once the semantic features are obtained, the model can fuse the image features with them to better represent the inter-class diversity and intra-class similarity of the place.

Although the method of scene recognition based on the object clue has an noticeable effect on improving recognition accuracy, it still has the problem of indistinguishable for the human-centred scene image. In order to compensate for the missing information, we attempt to tackle this problem by using natural language to describe human behaviour and state.

As for the place understanding, some researchers convert it to image understanding problem and attempt to solve it using semantic segmentation [5,25] and object detection [26,27] techniques. For example, Li et al. [28] proposed a hierarchical generative model to recognize and segment each object component in a scene. In [29], a convolutional encoder-decoder with a Bayesian framework is utilized for image segmentation. Choi et al. [30] proposed a method called 3D geometric phrases to combine object detection, layout estimation and scene classification together. Additionally, some works such as [31,32] achieve pixel-level scenes understanding by using depth information to segment images semantically. Since both semantic segmentation and object detection can only obtain the object and stuff categories, we attempt to use image caption technology to extract more information for robust place understanding.

2.2. Image captioning

The concept of image caption originates from literatures [33,34]. Its basic purpose is using natural language to describe the possible things in the image, and further achieve a more understandable way to express the image content. In recent research, some new network structures are proposed to improve the performance of the generation model using natural language fluency [35,36], richness [37,38] and accuracy [39–41]. In general, these methods use DNN to extract the global or local features of the image and then apply the sequence generation model such as RNN and LSTM to generate the corresponding natural language descriptions. To some extent, it is a kind of mapping function that converts information from the image domain to the natural language domain. Therefore, inspired by the dense caption method [37], we propose to introduce linguistic features obtained by image captioning to assist the place understanding.

3. The proposed place perception network

3.1. Overview of our model

To fully exploit the benefits of both visual and linguistic information, we propose a novel deep neural network that fuses different image representations for indoor place perception. As shown in Fig. 1, the pipeline of our proposed method is divided into four modules: region detection, image caption, feature extraction, and place recognition. Specifically, we first detect objects in an input image and extract its visual features, and then we convert the visual information into natural language. This procedure is known as dense captioning [37]. Finally, we extract both features from the image and its generated natural language and fuse the complementary cues to predict the place category. Besides, we design each module using a sub-DNN. By a combination of these sub-DNNs, we derive an end-to-end DNN for indoor place perception. Here we assume that the perceptual objects are all based on images as the information source.

As mentioned previously, to realize the process of place understanding and recognition, we utilize the natural language to enrich the representation of objects and stuff in a place image. Additionally, both image features and linguistic features are equally considered for place perception in our method. Therefore, as illustrated in Fig. 1, the proposed method is designed as a multi-task model. Since each part of the model has a distinct function, they need to be trained separately for making each module converge quickly and perform well. On the other hand, although the entire network is not trained in an end-to-end manner, it can be connected and fine-tuned together.

Specifically, as suggested in [37,38], image captioning model can be designed as an end-to-end integrated neural network, so in the actual constructing process, the proposed model is merged into three parts. The first part contains an object detection module based on ResNet-50 [42] with region proposal network (RPN) [43] and natural language generator based on LSTM. The object detector and natural language generator are trained together. The second part is a convolutional auto-encoder extracting image features of object regions and a mixed-CNN-LSTM network [44] capturing the linguistic features. The last part fuses visual and linguistic features and then input them into a softmax classifier to predict the possible category of a place image.

3.2. Dense captioning

3.2.1. Region detection network

The basic idea of dense captioning is to localize the Regions of Interest (ROIs) in an image and then express the ROIs in natural

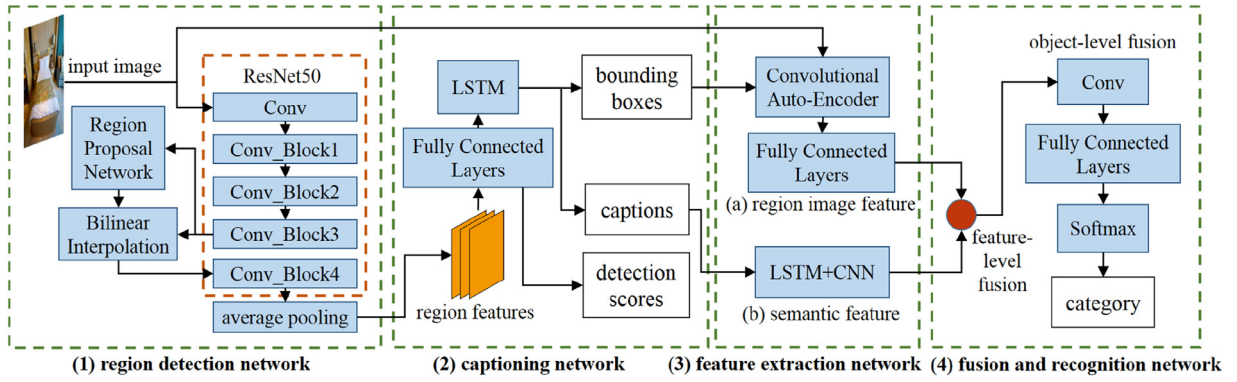


Fig. 1. The pipeline of the proposed deep neural network for place perception. Our model architecture consists of four modules: (1) a region detection network; (2) a captioning network; (3) a feature extraction network and (4) a fusion and recognition network.

language form. This process is similar to object detection, but it replaces the fixed number of object categories with plenty of visual concepts described by natural language phrases [38]. In order to improve the detection accuracy, we apply ResNet-50 [42] as the backbone in our object detector and replace the last two layers, *i.e.* fully connected layer and softmax layer with an average pooling layer. Additionally, to prevent feature map size from being too small to obtain the tiny object regions by convolution anchors, as shown in Fig. 1(1), the RPN [43] is connected with the third bottleneck layer *Conv_Block3* in ResNet-50 and a bilinear interpolation layer [45], which is then fed into the fourth bottleneck layer *Conv_Block4* to get the same-sized image features of region proposals.

Given an input image with a shape of $3 \times W \times H$, the third bottleneck layer *Conv_Block3* generates convolutional features of shape $C \times W' \times H'$, where $W' = \lfloor \frac{W}{16} \rfloor$, $H' = \lfloor \frac{H}{16} \rfloor$ and $C = 1024$ according to the structure of ResNet-50. Then RPN uses anchors centred at every grid of $W' \times H'$ feature maps with k ($k = 12$ in our method) different aspect ratios to predict region proposals by regressing offsets. Finally, since there are $W' \times H' \times k$ region proposals during training, we need to sample them to obtain a mini-batch containing B ($B = 256$) boxes with at most half of positive regions. Further, during the inference stage, we use greedy non-maximum suppression (NMS) based on the predicted detection scores to select the B' ($B' = 300$) most confident proposals. The sampling principle follows the approach in [37].

These region features sampled by anchor boxes are fed into bilinear interpolation operator [45] to obtain the same size of features. Specifically, given an input feature map $U_{c,i',j'}$ of shape $C \times W' \times H'$ and a $w_a \times h_a$ anchor box centred at (x_a, y_a) , the bilinear interpolation utilizes a kernel of $k(d) = \max(0, 1 - |d|)$ to calculate the interpolated elements in output feature map $V_{c,i,j}$ with a shape of $C \times 14 \times 14$ as follows:

$$V_{c,i,j} = \sum_{i'=1}^{W'} \sum_{j'=1}^{H'} U_{c,i',j'} k(i' - x_{i,j}) k(j' - y_{i,j}), \quad (1)$$

where $(x_{i,j}, y_{i,j})$ represents the coordinate of point (i, j) at output feature map, V corresponds to the horizontal and vertical coordinates in input feature map and U is expressed as follows:

$$\begin{cases} x_{i,j} = \frac{w_a}{14}(i + 0.5) - 0.5 + x_a - \frac{w_a}{2}, \\ y_{i,j} = \frac{h_a}{14}(j + 0.5) - 0.5 + y_a - \frac{h_a}{2}, \end{cases} \quad (2)$$

After bilinear interpolation, every proposal region outputs feature maps with the same size. These feature maps convolve with the last bottleneck layer *Conv_Block4* and perform average pooling, forming the final output with a shape of $B \times 2048$.

Meanwhile, the box regression is adopted according to a parameterization method in [43]. Given an anchor box with a size of $w_a \times h_a$, centred at (x_a, y_a) , RPN model predicts scalars (t_x, t_y, t_w, t_h) giving normalized offsets and log-space scaling transforms, so that the output region has a center at (x, y) and a shape of $w \times h$ given by Eq. (3).

$$\begin{aligned} x &= x_a + t_x w_a, & y &= y_a + t_y h_a, \\ w &= w_a \exp(t_w), & h &= h_a \exp(t_h), \end{aligned} \quad (3)$$

3.2.2. Captioning network

After the detection network calculates the region features together with their detection scores and bounding box offsets, features from each region are passed through a two-branch fully connected layers followed by rectified linear units (ReLU) non-linearity. As suggested in [38], if the bounding box offsets generated from RPN are simultaneously fed into the LSTM during caption generation, they will be predicted more precisely. According to this observation, we also adopt this joint inference module in our framework.

As for the natural language generation network, given a training sequence of tokens $[s_1, \dots, s_T]$, we firstly utilize Word2Vec method, also called word embedding [46,47], based on the entire corpus of image annotations in Visual Genome dataset, which is to convert a word into a real number vector of dimension D_{emb} . The number of token vocabulary is $|V^*| + 3$, where $|V^*|$ is the word vocabulary size and the other three tokens represent unknown or special word, start-of-sentence and end-of-sentence, *i.e.* $\langle UNK \rangle$, $\langle SOS \rangle$ and $\langle EOS \rangle$. After processed by Word2Vec, the training sequence can be converted to a tensor $[x_1, \dots, x_T]$. We add two more vectors in the training sequence, the first one x_{-1} is the region feature vector from region detection network, and the other one is $x_0 = \langle SOS \rangle$. Then, the LSTM computes a sequence of hidden states h_t and output vectors $y_t = f(h_{t-1}, x_t)$ according to the formula in [48].

At each test time, the output vector y_t corresponds to the most likely next token s_{t+1} at time $t = 0, \dots, T - 1$. Note that the output y_{-1} is ignored and the last token must be $y_T = \langle EOS \rangle$. The output of the LSTM is then evaluated by a fully connected layer with a softmax function to obtain the probability vector p_t , corresponding to each word y_t . The dimensionality of the p_t is $|V^*| + 3$, which reflects the possible word position in the vocabulary.

3.2.3. Loss function

Since we integrate the detection network and captioning network into an entire model, the loss function in each network are also combined for joint optimization of our framework. There are five loss function terms: (1) detection scores loss L_{det} : a two-class cross-entropy loss function for foreground and background regions.

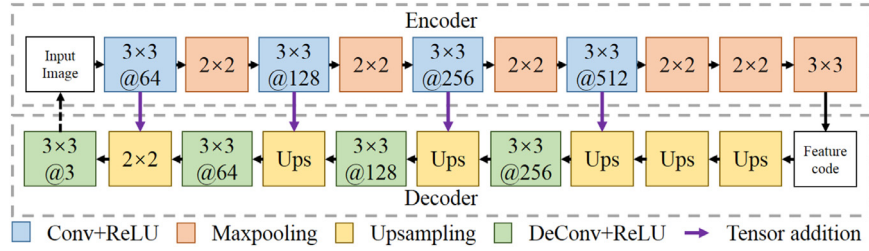


Fig. 2. The details of the convolutional auto-encoder's framework.

It needs to compute twice in RPN and fully connected layer in caption network, respectively; (2) bounding box offsets loss L_{bbox} : a smoothed-L1 loss function defined in Eq. (6). It needs to compute twice in RPN and LSTM of the captioning network, respectively; (3) captioning loss L_{cap} : the average cross-entropy given in Eq. (7), where x_i^{gt} is a logic vector reflecting the word position in the vocabulary for each word in the ground truth phrases. Specifically, gt refers to the ground-truth label, N_{reg} is the number of region proposals, and N_{cap} is the total number of output captioning words. The entire loss function L_C is a combination of these sub-loss terms, where $\alpha = 1.0$, $\beta = 1.0$ and $\gamma = 1.0$ are reasonable weights to balance each losses practically.

$$L_{det} = -\frac{1}{N_{reg}} \sum_i [p_i^{gt} \log p_i + (1 - p_i^{gt}) \log (1 - p_i)], \quad (4)$$

$$L_{bbox} = \frac{1}{N_{reg}} \sum_i p_i^{gt} \text{soomth}_{L_1}(t_i - t_i^{gt}), \quad (5)$$

$$\text{soomth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \quad (6)$$

$$L_{cap} = \frac{1}{N_{cap}} \sum_i \sum_{t=1}^T x_t^{gt} \log p_t, \quad (7)$$

$$L_C = \alpha L_{cap} + \beta L_{det} + \gamma L_{bbox}. \quad (8)$$

3.3. Feature extraction network

Since we have obtained linguistic descriptions through the captioning network, the place can be represented by extracting the semantic features, as shown in Fig. 1(3-b). Although the semantic feature is usually a high-level abstract feature with significant distinctions, it lacks low-level image details such as color, texture, and edge. In our framework, as shown in Fig. 1(3-a), we also design another network to extract low-level region image features.

3.3.1. Region image feature extraction network

Noting that the mapping between an object and the place to which it belongs is not unique, so the object region cannot associate to an accurate label of place for training a neural network. Thus, as demonstrated in Fig. 2, a convolutional auto-encoder is adopted to capture the image feature.

However, the deconvolutional operator [49,50] in the decoder has a significant problem of edge effect caused by padding with zeros. Besides, as we use bilinear interpolation, expressed in Eq. (1), to restore the size of the image feature, this upsampling operation can lead to the local averaging effect of pixels. In order to compensate for information discarded by the encoder, we modify the auto-encoder structure with an across-layers cascading method. As shown in Fig. 2, the bilinear upsampling layers in the decoder are added with the same-sized convolutional layers in the encoder.

At training time, we use mean squared error, defined in Eq. (9), as the loss function, where $x(i, j)$ and $\hat{x}(i, j)$ represent the ground truth image and restored image, respectively. Additionally, the region image is resized into 224×224 to obtain the same size as an image feature vector.

$$L_{img} = \frac{1}{N} \sum_{(i,j)} [\hat{x}(i, j) - x(i, j)]^2, \quad (9)$$

To ensure the symmetry of each feature layers between encoder and decoder, The mean square error of the feature layer, defined in Eq. (10), is added to the loss function as a regularization term:

$$L_{imgf} = \sum_{m=1}^{n-1} \frac{1}{N_m} \sum_{(i_m, j_m)} [D_m(i_m, j_m) - P_{n-m}(i_m, j_m)]^2, \quad (10)$$

where $D_m(i_m, j_m)$ is the feature map of the m th transposed convolutional layer in decoder, $P_{n-m}(i_m, j_m)$ is the feature map of the $(n - m)$ th max-pooling layer in encoder, and n represents the number of convolutional operators. The final loss function of CAE is a weighted sum of L_{img} and L_{imgf} as shown in Eq. (11), where $\alpha = 0.7$ and $\beta = 0.1$. Since the purpose of CAE is to extract image features, the training process only needs to restore the image roughly.

$$L_{CAE} = \alpha L_{img} + \beta L_{imgf}, \quad (11)$$

As shown in Fig. 3, after 20 epochs of training, the region image can be basically recovered, and the training session can be stopped. After this process, the weights of CAE are stored as a pre-trained model which is later fine-tuned with fusion and recognition networks. At test time, the input of the encoder is the region image generated by the detection network and resized into 224×224 . The output of the last average pooling layer in the encoder is then flattened to a 4608-dimensional vector. Finally, the region image feature is input to three fully connected layers which reduce its dimensionality to 128 to match with the semantic feature.

3.3.2. Semantic feature extraction network

As shown in Fig. 4, we propose a mixed CNN-LSTM model [44] to extract the features of descriptive phrases for expressing the implicit contents. Note that the ground truth descriptive phrases include a series of tokens, e.g. special symbols, numeric characters, punctuation, or meaningless auxiliary words (stop words). These tokens need to be removed or modified because they do not help train neural networks. Thus, the preprocessing step includes: (1) Remove punctuation, extra spaces, and special symbols that do not affect the original semantic meaning of descriptive phrases. (2) Replace the numeric characters in phrases with the corresponding words. (3) Remove stop words without changing the original semantic meaning.

Besides, the length of descriptive phrases needs to be fixed with the same size to satisfy the demand of data dimensionality in the neural network. According to the statistics of phrase length in Visual Genome dataset [51], we select a reasonable length parameter as $L_{max, len} = 8$. For any normalized phrase, if its length is longer



Fig. 3. Examples of ground truth region images (shown in the first row) and the recovered images through convolutional auto-encoder (shown in the second row).

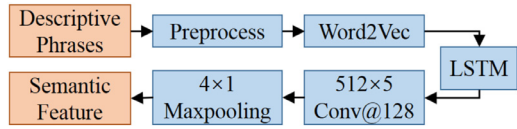
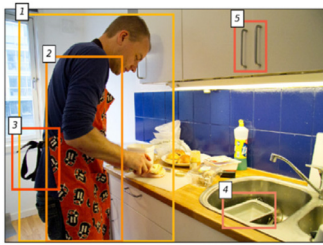


Fig. 4. The structure of the semantic feature extraction network.



- Label: kitchen
1. man wearing an apron
 2. red and black apron
 3. black ribbon tied in a bow
 4. silver pan in the sink
 5. two handles on the cabinet

Fig. 5. An example of an image and their descriptive phrases that contain humans from the SubVisGen dataset. (Parts of phrases are omitted because of space limitation.).

than L_{max_len} , we only keep its first L_{max_len} words. Otherwise, the phrase will be filled with zero until its length reaches L_{max_len} .

After preprocessing step, the normalized phrases are converted into word vector form using the Word2Vec method [46,47]. As a particular case, the filled zero is converted into a zero vector as a place-holder without any meaning. The dimensions of both word vector and zero vector are consistent with the dimension D_{emb} of word vectors in the captioning network. Both the ground truth phrases used for training and phrases generated from the captioning network for testing are required to be preprocessed.

The semantic feature extraction network consists of an LSTM, a convolutional layer with ReLU as activation function, and a max-pooling layer for dimensionality reduction. Since the phrases are pruned or filled with zero, the LSTM is designed for repairing and sharing information in a whole phrase. The size of the hidden layer is 512, which is the same length of word vector and the parameter in forget gate is set to 0.5 to keep part of previous output information. The final length of the semantic feature vector of a phrase is 128.

Since the semantic feature lacks association with an accurate label either, we combine the semantic feature extraction network with the fusion and recognition network to further train and fine-tune our model.

3.4. Fusion and recognition network

In the fusion and recognition network, we attempt to: (1) Fuse an object’s low-level image features with its high-level semantic features. (2) Fuse different objects’ features within the same image. (3) Perform place recognition using fused features.

In the feature fusion step, we propose two strategies, namely feature-level fusion and object-level fusion. Firstly, to solve the problem of feature fusion with two different patterns, we use three

distinct operators to realize feature-level fusion, respectively and test their performance under the same condition. Our compared operators include 1). Hadamard product, denoted as \odot , which means the element-wise multiplication, 2). element-wise addition denoted as \oplus , and 3). the augmented operator denoted as $[\cdot, \cdot]$, which means the concatenation of two vectors.

Secondly, for the object-level fusion, we use a convolutional layer to combine the fused features from different objects together to obtain a comprehensive feature representation. Notably, we fuse the features of the first O_{top} objects which have the highest detection scores. If there are less than O_{top} detected objects in the image, the existing objects are sampled repeatedly in the order of their detection score until the number of objects reaches O_{top} . After object-level fusion, the object features tensor with a shape of $O_{top} \times 128$ can be converted into a 128-dimensional vector.

Finally, a fully connected layer is utilized to reduce the feature dimensionality to fit with categories of indoor places we focus on. The softmax function is used as a classifier and the cross-entropy function defined in Eq. (12) is the loss function.

$$L_{rec} = -\frac{1}{N_I} \sum_i I_i^{label} \log p(I_i^{logits}), \quad (12)$$

where $p(I_i^{logits})$ represents the softmax value of the i th image corresponding to five types of indoor places and I_i^{label} is its ground truth label expressed in one-hot encoding form.

4. Experiment

We evaluate our place perception model on four public image datasets including MIT-67 [52], Places2 [3], SUN397 [53], and Visual Genome [51]. Specifically, we choose five kinds of indoor places, namely bathroom, bedroom, kitchen, living room, and office, from all datasets for evaluation and Table 1 lists the total number of samples from each category in these dataset. Since there is no clear place category label in Visual Genome dataset, we manually annotate labels for 3137 images to form the SubVisGen dataset. At the time of evaluation, we use all samples from the SubVisGen dataset to train the model and verify its performance on other datasets. Fig. 5 provides an example of a place image with its descriptive phrases.

The entire model is implemented using Tensorflow 1.12, and all experiments are conducted on Dell PowerEdge T630 Intel Xeon CPU Es-2650 v4@ 2.20 GHz x 48 processors with 64GB RAM and GeForce GTX Titan X GPU, running on Ubuntu 16.04 system.

Table 1
The public datasets used in our model.

| Dataset | Bath. | Bed. | Kit. | Liv. | Off. | Total |
|-----------|-------|------|------|------|------|-------|
| MIT-67 | 197 | 647 | 729 | 703 | 105 | 2381 |
| Places2 | 100 | 100 | 100 | 100 | 100 | 500 |
| SUN397 | 951 | 2083 | 1746 | 2361 | 136 | 7277 |
| SubVisGen | 713 | 639 | 624 | 617 | 544 | 3137 |

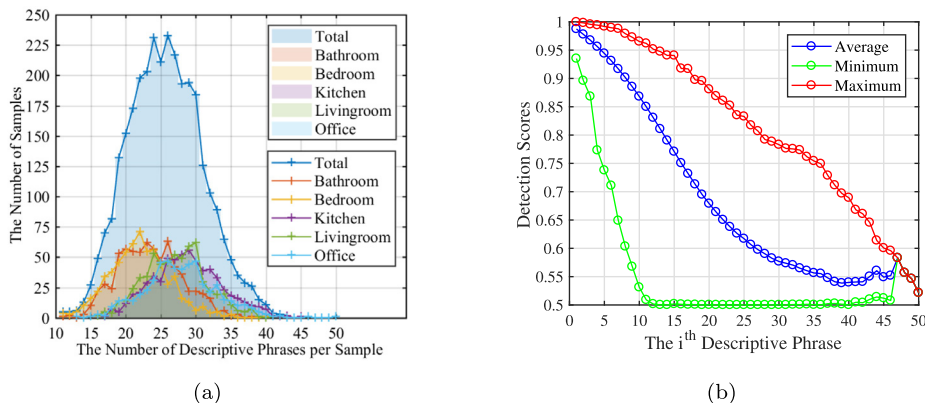


Fig. 6. Statistics of descriptive phrases generated from the captioning network on the entire SubVisGen dataset. (a) The histogram of the descriptive phrases' number for five types of places. (b) The average, minimum and maximum values of the detection scores for the i th descriptive phrase in all samples. The horizontal axis represents the ordinal value of descriptive phrases. (The output phrases are sorted according to the detection scores.)

4.1. Parameter settings and training details

For the region detection network described in Section 3.2, the input image is resized to have a longer side of 720 pixels. For the captioning network, the method of Word2Vec [46,47] is trained on the whole descriptive phrases from the Visual Genome dataset, which has a vocabulary of frequent 10K words and each word vector has the dimension of 512. The size of the sliding window is set to 5. Besides, we use the stochastic gradient descent optimizer with a mini-batch size of 1. The learning rate is set to 0.001 and decreases 50% every 100K iterations, and momentum is set to 0.98. The whole region detection and the captioning network stops training after 500K iterations. (almost 8 epochs).

In the fusion and recognition network, there are two hyper-parameters, namely fusion operator and the number of objects, required to be tuned according to the demand of feature-level fusion and object-level fusion. We choose the appropriate hyper-parameter through experimental tests. To determine a reasonable object-level fusion size, we firstly count the number of phrases generated by the captioning network. Usually, most of the images can be described by about 25 phrases. On the other hand, since we select the phrases according to the detection score, it is also an important reference index. If the detection score is too low, it may be because the detected object is a background, or the confidence of recognition accuracy is low. Therefore, we also consider the distribution of the detection score. Fig. 6(b) shows that the average score of the 25th generated phrase is as low as 0.6, so the remaining descriptive phrases may not be available. Therefore, we test several different sizes of object-level fusion on the condition, given the fusion operator as a Hadamard product, and the training and test datasets remain consistent during the experiment. As shown in Table 2, a reasonable object-level fusion size should not be too large or too small. Too small fusion size may miss object information and cannot describe the entire image effectively. On the other hand, too large fusion size may bring too much irrelevant information that decreases the recognition accuracy. Therefore, the object-level fusion size is set to 16 as an appropriate choice.

For fusion operators, we test three fusion methods and report experimental results in Table 3. It shows that the Hadamard prod-

Table 2
The classification accuracy on the test set when using different sizes of object-level fusion on the SubVisGen dataset (the operator of feature-level fusion is fixed as \odot).

| Size | 4 | 8 | 12 | 16 | 20 | 24 | 28 |
|-------------|-------|-------|-------|--------------|-------|-------|-------|
| Accuracy(%) | 75.93 | 87.18 | 91.17 | 92.34 | 89.27 | 88.21 | 86.83 |

Table 3
Classification accuracy of different fusion operators on the test set of SubVisGen dataset (the size of object-level fusion is fixed to 16).

| Fusion operator | \odot | \oplus | $[\cdot, \cdot]$ |
|-----------------|---------|----------|------------------|
| Accuracy(%) | 92.34 | 85.01 | 78.79 |

uct outperforms other fusion operators. However, such results cannot be explained explicitly by mathematical principles. We can only intuitively argue that Hadamard product introduces spatial projection, which can reduce feature dimensionality and benefit scene classification.

We use the stochastic gradient descent optimizer with a learning rate of 0.001 to train the fusion and recognition network and fine-tune the feature extraction network in 30 epochs. In the last 5 epochs, the learning rate decreases every epoch with a decay rate of 0.9. Additionally, the L2-regularization is adopted in the convolutional layers and fully connected layers of recognition network to improve the generalization ability of our model.

4.2. Experimental results and analysis

The proposed method is firstly evaluated under a 5-fold cross-validation setting. In order to perform a fair comparison, the SubVisGen dataset is evenly divided into five parts. Each fold evaluation uses one-fifth of data as a test set, and the remaining data are used as a training set. Table 4 shows the evaluation results under the same conditions in each fold. All metrics are maintained at a high level, and there is no significant fluctuation among different each fold splittings, which indicates that the proposed method is valid and has satisfying generalization capability for datasets with uneven distribution.

Furthermore, we also compare the performance of different public datasets. Table 5 provides the experimental results of the proposed model on each dataset, and Table 6 explores the corresponding statistical results of each place category. The validation model is obtained by training on all data from SubVisGen dataset.

Table 4
Evaluation results of the proposed method on the SubVisGen dataset. The last column shows the average value and standard deviation under 5-fold cross-validation setting.

| | fold-1 | fold-2 | fold-3 | fold-4 | fold-5 | average(std) |
|-------------|--------|--------|--------|--------|--------|--------------------|
| F1(%) | 92.14 | 89.93 | 91.56 | 90.52 | 93.50 | 91.58(\pm 1.28) |
| Accuracy(%) | 92.34 | 90.27 | 91.39 | 90.59 | 93.78 | 91.67(\pm 1.27) |

Table 5
Comparison of F1 and accuracy on each test dataset.

| Dataset | MIT-67 | Places2 | SUN397 | SubVisGen |
|-------------|--------|---------|--------|-----------|
| F1(%) | 71.40 | 82.58 | 81.76 | 92.25 |
| Accuracy(%) | 74.59 | 82.40 | 85.17 | 92.35 |

Table 6
Comparison of F1 on each categories.

| F1(%) | Bath. | Bed. | Kit. | Liv. | Off. |
|-----------|-------|-------|-------|-------|-------|
| MIT-67 | 79.82 | 72.71 | 82.50 | 70.82 | 51.14 |
| Places2 | 91.79 | 86.02 | 85.71 | 72.27 | 77.11 |
| SUN397 | 90.60 | 84.34 | 88.06 | 83.04 | 62.79 |
| SubVisGen | 97.32 | 92.52 | 93.95 | 85.95 | 91.50 |

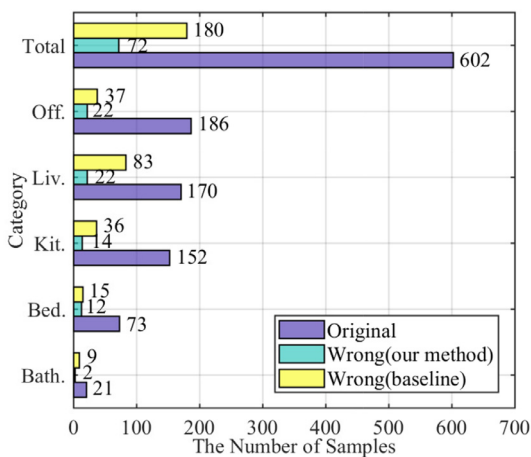


Fig. 7. Recognition results of samples with existing humans in images on SubVisGen dataset.

From the above experimental results, we observe that: (1) The proposed method performs worse on MIT-67 dataset, compared with other datasets. (2) Considering the test results among different places, the recognition results of the living room and office are slightly worse than those of the other categories.

For the first phenomenon, most images in the MIT-67 dataset have a lower resolution than those in other datasets, so it will cause difficulty for the model to detect more objects. The result also shows that the proposed model has specific requirements for the resolution of the input image. For the other phenomenon, it is because living rooms and offices contain slightly more objects than in other places, which is also consistent with our common sense. This prior knowledge can be indirectly inferred from Fig. 6(a). Therefore, this leads to a common problem in most place recognition methods that use objects as a clue, *i.e.*, when there is too much information such as categories, attributes, and relationships about the objects, it may be challenging to identify the category of place correctly. Since our method selects several objects in the order of detection score, it can not guarantee that all of the selected objects are helpful for place recognition.

4.3. Ablation studies

This section aims to analyze the influence of significant components that constitute the proposed method. First, the influence of the region detection module is discussed. Second, we evaluate the effectiveness of two feature extraction network. Meanwhile, comparing with the ResNet-50 baseline model, we consider the impact of the human cues on the performance of our model. Besides, the ablation studies of fusion strategies, namely feature fusion and object fusion are analyzed in Section 4.2, and all the ablation studies

Table 7
Ablation results for different backbones of the region detection module.

| Backbone | Number of Parameters | mAP | Accuracy(%) |
|------------|----------------------|------|-------------|
| VGG-16 | ~ 4.7M | 6.08 | 82.26 |
| ResNet-50 | ~ 20.7M | 8.29 | 92.34 |
| ResNet-101 | ~ 40.0M | 8.33 | 92.47 |

Table 8
Ablation results for different feature extraction networks.

| Model | Image feature | Semantic feature | ResNet-50 (baseline) |
|-------------|---------------|------------------|----------------------|
| Accuracy(%) | 63.87 | 86.26 | 76.31 |

are carried out using the SubVisGen dataset with the Top@1 accuracy metric.

4.3.1. Influence of the region detection network

Since the region detection network is the backbone of entire model, the performance of the detection module directly affects the rest parts. Table 7 gives the results of three different backbones of the detection network. We refer to the work [37,42] to change the ResNet-50 into the VGG-16 and ResNet-101 as the feature extraction network. Because the region detection and captioning network are designed as a whole part, we evaluate them jointly to compare the performance, which has the same metric and test settings as [37].

Results indicate that the place recognition and understanding capabilities of deeper networks are a little better than the shallower networks. As Table 7 presents, compared with the VGG-16, the ResNet-50 generates more accurate natural language descriptions and detects the objects more in line with the ground-truth, which is helpful to perceive the place image precisely. On the other hand, compared with the ResNet-50, the ResNet-101 has less potential in improving performance (mAP and accuracy metrics increase by 0.04 and 0.13%, respectively). Because when the depth size is sufficiently deep, the enhancement of depth has no significant effect on feature representation [42]. In the light of these experiments, we utilize ResNet-50 architecture as a trade-off solution between performance and complexity.

4.3.2. Influence of the feature extraction network

We have performed three ablation studies on the influence of the feature extraction module. As shown in Table 8, the first two columns present the results of using the image or semantic features separately. These two experiments are performed with the same model parameters and only reserving the object-level fusion layer. Additionally, the last column presents the results of removing both the region image and semantic feature extraction network. We use the ResNet-50 as a baseline model for place image classification, which is derived from reference [42] and fine-tuned on the SubVisGen dataset.

Results indicate that the semantic feature plays a more critical role in inferring the place's category, which is consistent with the theoretical analysis in Section 3.3. On the other hand, compared with the baseline model, the model using the image feature of objects region has a significant decrease in the recognition accuracy. This phenomenon probably due to the lack of clear correspondence between the place category and the object information, and the local features are also more complex than the global, which leads to the difficulty in identification.

Besides, considering the results of Tables 7 and 8 together, we notice that the LSTM network in the captioning module has the most significant impact on the recognition process. If the LSTM can


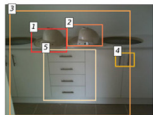


| Dataset | MIT-67 | Places2 | SUN397 | SubVisGen |
|---------------------|---|--|---|---|
| Image | (a)  | (b)  | (c)  | (d)  |
| Descriptive Phrases | 1.the tv is white 2.a white paper 3.a book on the table 4.the tv is open 5.a blue box | 1.a white pot 2.a white lid 3.the wall is white 4.a handle on the wall 5.the drawers are white | 1.a picture on the wall 2.a brown wooden door 3.a wooden table 4.the clock is black 5.a black pot | 1.a man is holding a pizza 2.a man is holding a pizza 3.a bottle of water 4.a woman is standing 5.a silver faucet |
| True | office | kitchen | kitchen | bathroom |
| Predicted | living room | bathroom | living room | kitchen |

Fig. 8. Examples of incorrect place perception. The bounding boxes and captions are sorted in the descending order of the detection score.

receive more precise image features of objects as the initial state from the detection network, the results of place recognition and understand will be better. In another aspect, compared with the single module, the entire model results in a 28.47% and a 6.08% increase in term of accuracy with respect to the image feature and semantic feature. To some extent, the image features of objects are very diverse but may be described by the same word. Therefore, we think that the intra-class similarity of place images is significantly improved after represented by natural language, and the role of image features is to enhance the difference in inter-class.

Apart from the attributes and states of objects, the humans in the images are also our primary focus. To this end, we count the original images with existing human from the SubVisGen dataset, as well as the images with wrong recognition results in our model and baseline model. Fig. 7 shows that our method has a good effect on place perception based on the human clue. Specifically, it obtains only 11.96% error rate, which is lower than 29.90% error rate in the baseline model.

4.4. Comparison to previous works

Along this section, the proposed method is compared with 10 recent approaches, ranging from common CNN architectures to using semantic information to drive scene recognition. Comparison is performed on MIT-67 dataset. Since this dataset does not contain any natural language descriptions of images, we use captioning module pre-trained on Visual Genome dataset to generate them and further compose a new MIT-67 dataset, which is used to train our entire model. Unless explicitly mentioned, the results of all approaches are extracted from their respective papers.

Table 9 depicts the performance of each method. As these datasets are balanced, we only compare the Top@1 accuracy metric. All listed algorithms are based on CNNs (see the Backbone column for details) and parts of them consider the semantic infor-

mation (marked in the Semantic column). It is worth noting that the training datasets in Tables 5 and 9 are different. In the experiment of Table 5, the model is trained on SubVisGen dataset and does not carry out transfer learning on MIT-67 dataset, so it obtains unsatisfactory recognition results. On the other hand, the linguistic information of MIT-67 dataset in Table 9 is not from the ground-truth annotations, but from the pre-trained model on Visual Genome dataset, which is a transfer learning process. These two results indicate that the proposed model is sensitive to semantic data, which is consistent with the conclusion in Section 4.3.2.

As the discussion in Section 2, the listed methods are divided into two main types. The first is from the perspective of image features, and these methods employ strategies of improving network structures or feature representations. The second is to consider the different information modalities, and the semantic information of objects are mostly taken into account. To some extent, both of two core ideas are beneficial to improving recognition performance. On the other hand, we notice from methods [10] and [19] that considering object information can significantly increase the recognition accuracy, whether from the perspective of image features or semantic features.

Compared with the methods using semantic information, our model considers more complete linguistic cues rather than the distribution of semantic labels for objects or regions, which makes the semantic features more abundant and accurate. Therefore, our approach provides 4.50% and 4.16% performance increments over methods [8] and [10], and leads other methods by a relative margin in terms of recognition accuracy.

4.5. Limitations of the proposed method

In addition to the above discussion, we show some failed examples of our method in Fig. 8, which mostly occur when the generated natural language description conflicts with the image. Sev-

Table 9
State-of-the-art results on MIT-67 dataset.

| Method | Backbone | Semantic | Accuracy(%) |
|-----------------------------|-----------------------|----------|-------------|
| Zhou, et al.[3] | VGG-16 | | 79.76 |
| Xie, et al.[17] | VGG-19 | | 82.24 |
| Wang, et al.[24] | 2 × VGG-19 | ✓ | 86.20 |
| Wang, et al.[20] | 2 × BN-Inception | | 86.70 |
| Cheng, et al.[8] | 2 × VGG-19 | ✓ | 86.76 |
| Pan, et al.[18] | ResNet-101+ResNet-152 | | 87.60 |
| Laranjeira, et al.[9] | ResNet-50+BiLSTM | | 88.25 |
| Sun, et al.[23] | 2 × VGG-16+YOLOv2 | ✓ | 89.51 |
| López-Cifuentes, et al.[10] | ResNet-18+ResNet-50 | ✓ | 87.10 |
| Seong, et al.[19] | 2 × SE-ResNeXt-101 | | 90.30 |
| Ours | ResNet-50+LSTM+CAE | ✓ | 91.26 |

eral reasons are causing the misidentification of our method: (1) Misclassification of the object category. As listed in Fig. 8(a) and (d), two important objects are classified into wrong categories, e.g. the computer monitor is recognized as “the tv is white” and the toothbrush is recognized as “a man is holding a pizza”. (2) Object information does not play a decisive role in perception. As listed in Fig. 8(b) and (c), since the recognized objects with high detection scores are not decisive for place understanding, it is difficult to determine the category of place accurately. (3) The complexity and ambiguity of the original image. As shown in Fig. 8(a) and (b), due to the shooting range and angle of the camera, the objects contained in the image are inherently ambiguous, which cannot be solved by the model itself.

In general, the main factor that limits the performance of the proposed model is whether the descriptive phrases generated from the captioning network are accurate, especially the category of the described objects. Although conflicts can lead to semantic information not complementing with image modality, and further cause recognition errors, the ablation experiments suggest that two kinds of fusion strategies can ensure that linguistic cue has a certain credibility. Therefore, the performance of our model would improve if it can successfully detect and describe the objects that play a decisive role in the perception.

5. Conclusion

In this paper, we attempt to leverage both semantic features of natural language and low-level image features to imitate the human’s way of place understanding, so a novel deep neural network structure is designed to realize indoor place perception. This method mainly employs the image captioning technology to generate linguistic clues and then fuses with the image features. Besides, two fusion strategies, namely object-level fusion and feature-level fusion, are proposed to realize the fusion of multi-modal information.

Experimental results indicate that our method is valid for the place perception. Besides, from the ablation study analysis, semantic features significantly impact place recognition, and adding natural language information helps improve recognition accuracy. Therefore, we think that considering natural language information is an effective way to solve the problem of place perception.

Although the proposed method has acceptable performance, some limitations still exist: (1) The captioning network’s performance needs to be improved because the accuracy and richness of natural language information can influence the recognition progress. (2) The model does not consider the decision problem when information conflict occurs, leading to wrong recognition results. For future work, we will focus on boosting the model performance from the aspect of the above two shortcomings. Besides, the generalization performance of the proposed model needs further improvement.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the [National Natural Science Foundation of China](#) (Grant No.91748106 and 61671151), in part by Key Laboratory of Integrated Automation of Process Industry (PAL-N201704), in part by the Advanced Research Project of the 13th Five-Year Plan under Grant 31511040301, and in part by

Guangdong Innovative and Entrepreneurial Research Team Program (No. 2019ZT08Z780).

References

- [1] S. Lowry, N. Sünderhauf, P. Newman, J.J. Leonard, D. Cox, P. Corke, M.J. Milford, Visual place recognition: a survey, *IEEE Trans. Rob.* 32 (1) (2015) 1–19.
- [2] S.H. Khan, M. Hayat, M. Bennamoun, R. Togneri, F.A. Sohel, A discriminative representation of convolutional features for indoor scene recognition, *IEEE Trans. Image Process.* 25 (7) (2016) 3372–3383.
- [3] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (6) (2017) 1452–1464.
- [4] J.M. Henderson, A. Hollingworth, High-level scene perception, *Annu. Rev. Psychol.* 50 (1) (1999) 243–271.
- [5] L.-J. Li, H. Su, L. Fei-Fei, E.P. Xing, Object bank: a high-level image representation for scene classification & semantic feature sparsification, in: *Advances in Neural Information Processing Systems*, 2010, pp. 1378–1386.
- [6] L. Zhang, X. Zhen, L. Shao, Learning object-to-class kernels for scene classification, *IEEE Trans. Image Process.* 23 (8) (2014) 3241–3253.
- [7] A.C. Hernández, C. Gómez, R. Barber, O.M. Mozos, Object-based probabilistic place recognition for indoor human environments, in: *2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, IEEE, 2018, pp. 177–182.
- [8] X. Cheng, J. Lu, J. Feng, B. Yuan, J. Zhou, Scene recognition with objectness, *Pattern Recognit.* 74 (2018) 474–487.
- [9] C. Laranjeira, A. Lacerda, 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. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, Á. García-Martín, Semantic-aware scene recognition, *Pattern Recognit.* 102 (2020) 107256.
- [11] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, D. Tao, Unsupervised semantic-preserving adversarial hashing for image search, *IEEE Trans. Image Process.* 28 (8) (2019) 4032–4044.
- [12] C. Deng, E. Yang, T. Liu, D. Tao, Two-stream deep hashing with class-specific centers for supervised image search, *IEEE Trans. Neural Netw. Learn. Syst.* (2019) 1–13.
- [13] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, J.T. Zhou, Deep clustering with sample-assignment invariance prior, *IEEE Trans. Neural Netw. Learn. Syst.* (2019) 1–12.
- [14] X. Peng, Z. Huang, J. Lv, H. Zhu, J.T. Zhou, COMIC: Multi-view clustering without parameter selection, in: *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, PMLR, 2019, pp. 5092–5101.
- [15] M. Juneja, A. Vedaldi, C. Jawahar, A. Zisserman, Blocks that shout: distinctive parts for scene classification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 923–930.
- [16] S. Aziz, Z. Kareem, M.U. Khan, M.A. Imtiaz, Embedded system design for visual scene classification, in: *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2018, pp. 739–743.
- [17] G. Xie, X. Zhang, S. Yan, C. Liu, Hybrid CNN and dictionary-based models for scene recognition and domain adaptation, *IEEE Trans. Circuits Syst. Video Technol.* 27 (6) (2017) 1263–1274.
- [18] Y. Pan, Y. Xia, D. Shen, Foreground fisher vector: encoding class-relevant foreground to improve image classification, *IEEE Trans. Image Process.* 28 (10) (2019) 4716–4729.
- [19] H. Seong, J. Hyun, E. Kim, FOSNet: an end-to-end trainable deep neural network for scene recognition, *IEEE Access* 8 (2020) 82066–82077.
- [20] L. Wang, S. Guo, W. Huang, Y. Xiong, Y. Qiao, Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs, *IEEE Trans. Image Process.* 26 (4) (2017) 2055–2068.
- [21] H. Zhu, J.-B. Weibel, 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.
- [22] G. Li, W. Chou, F. Yin, Multi-robot coordinated exploration of indoor environments using semantic information, *Sci China Inf Sci* 61 (7) (2018) 79201.
- [23] N. Sun, W. Li, J. Liu, G. Han, C. Wu, Fusing object semantics and deep appearance features for scene recognition, *IEEE Trans. Circuits Syst. Video Technol.* 29 (6) (2019) 1715–1728.
- [24] Z. Wang, L. Wang, Y. Wang, B. Zhang, Y. Qiao, Weakly supervised patchnets: describing and aggregating local patches for scene recognition, *IEEE Trans. Image Process.* 26 (4) (2017) 2028–2041.
- [25] W. Liu, Y. Li, Q. Wu, An attribute-based high-level image representation for scene classification, *IEEE Access* 7 (2018) 4629–4640.
- [26] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: quantifying interpretability of deep visual representations, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6541–6549.
- [27] X. Song, S. Jiang, L. Herranz, Multi-scale multi-feature context modeling for scene recognition in the semantic manifold, *IEEE Trans. Image Process.* 26 (6) (2017) 2721–2735.
- [28] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: classification, annotation and segmentation in an automatic framework, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 2036–2043.

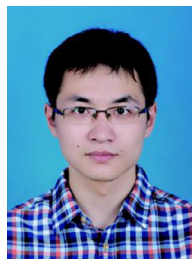
- [29] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, arXiv preprint arXiv:1511.02680(2015).
- [30] W. Choi, Y.-W. Chao, C. Pantofaru, S. Savarese, Indoor scene understanding with geometric and semantic contexts, *Int. J. Comput. Vis.* 112 (2) (2015) 204–220.
- [31] Y. Chen, D. Pan, Y. Pan, S. Liu, A. Gu, M. Wang, Indoor scene understanding via monocular RGB-D images, *Inf. Sci.* 320 (2015) 361–371.
- [32] A. Handa, V. Pătrăucean, S. Stent, R. Cipolla, SceneNet: an annotated model generator for indoor scene understanding, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 5737–5743.
- [33] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
- [34] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
- [35] Z. Ren, X. Wang, N. Zhang, X. Lv, L.-J. Li, Deep reinforcement learning-based image captioning with embedding reward, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 290–298.
- [36] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.
- [37] J. Johnson, A. Karpathy, L. Fei-Fei, DenseCap: fully convolutional localization networks for dense captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565–4574.
- [38] L. Yang, K. Tang, J. Yang, L.-J. Li, Dense captioning with joint inference and visual context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2193–2202.
- [39] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
- [40] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, Context and attribute grounded dense captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6241–6250.
- [41] A. Deshpande, J. Aneja, L. Wang, A.G. Schwing, D. Forsyth, Fast, diverse and accurate image captioning guided by part-of-speech, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10695–10704.
- [42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [43] S. Ren, K. He, R. Girshick, 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.
- [44] C. Zhou, C. Sun, Z. Liu, F. Lau, A C-LSTM neural network for text classification, arXiv preprint arXiv:1511.08630(2015).
- [45] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.
- [46] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
- [47] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of Workshop at ICLR, 2013, 2013.
- [48] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
- [49] V. Dumoulin, F. Visin, A guide to convolution arithmetic for deep learning, arXiv preprint arXiv:1603.07285(2016).
- [50] A. Odena, V. Dumoulin, C. Olah, Deconvolution and checkerboard artifacts, *Distill* 1 (10) (2016) e3.
- [51] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vis.* 123 (1) (2017) 32–73.
- [52] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 413–420.
- [53] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, A. Torralba, Sun database: large-scale scene recognition from abbey to zoo, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3485–3492.



Pei Li is currently studying for the Ph.D. and Master's degree in Control Science and Engineering at the School of Automation, Southeast University, Nanjing, China. He received his Bachelor's degree with a major in Automation from the Tongji University, Shanghai, China. His research interests include place perception, deep learning, natural language processing.



Xinde Li received his Ph.D. from the Department of Control, Huazhong University of Science and Technology in June 2007. In December of the same year, he worked in the School of Automation, Southeast University. From January 2012 to January 2013, he visited Georgia Polytechnic University as a national visiting scholar for one year. From January 2016 to the end of August 2016, he worked as a research fellow in the Department of ECE, National University of Singapore. His main research interests include intelligent robots, machine vision perception, machine learning, human-computer interaction, intelligent information fusion and artificial intelligence.



Xianghui Li received the B.Sc. degrees in automation from Southeast University, Nanjing, Jiangsu Province, China in 2014. He is currently pursuing Ph.D. degree in pattern recognition at Southeast University. His research interests include deep learning and place identification.



Hong Pan is associate researcher at the School of Automation, Southeast University. In 2004, he graduated from Southeast University with his Ph.D. in pattern recognition and intelligent systems. His research interests include machine learning, deep learning, computer vision, medical image processing and analysis, multimedia signal processing (image/video codec, retrieval and analysis).



M. O. Khyam received the B.Sc. degree in electronics and telecommunication engineering from the Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, in 2010, and the Ph.D. degree from the University of New South Wales, Australia, in 2015. His research interests include signal processing and wireless communication.



Md. Noor-A-Rahim received the Ph.D. degree from Institute for Telecommunications Research, University of South Australia, Adelaide, SA, Australia, in 2015. He was a Postdoctoral Research Fellow with the Centre for Information Technology, Nanyang Technological University, Singapore. He is currently a Senior Postdoctoral Researcher (MSCA Fellow) with the School of Computer Science and IT, University College Cork, Cork, Ireland. His research interests include control over wireless networks, intelligent transportation systems, information theory, signal processing, and DNA-based data storage.



Shuzhi Sam Ge is with the Social Robotics Laboratory, Department of Electrical and Computer Engineering, Interactive Digital Media Institute, National University of Singapore, Singapore. S. S. Ge is also with Institute for Future (IFF), Qingdao University, China.