Pattern Recognition 90 (2019) 436–449

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

MAPNet: Multi-modal attentive pooling network for RGB-D indoor scene classification

Yabei Li[a,b], Zhang Zhang[a,b,*], Yanhua Cheng[c], Liang Wang[a,b,d], Tieniu Tan[a,b,d]

[a] CRIPAC & NLPR, CASIA, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Tencent WeChat AI, Beijing, China
[d] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

ARTICLE INFO

Article history:
Received 20 August 2018
Revised 8 January 2019
Accepted 7 February 2019
Available online 8 February 2019

Keywords:
Indoor scene classification
Multi-modal fusion
RGB-D
Attentive pooling

ABSTRACT

RGB-D indoor scene classification is an essential and challenging task. Although convolutional neural network (CNN) achieves excellent results on RGB-D object recognition, it has several limitations when extended towards RGB-D indoor scene classification. 1) The semantic cues such as objects of the indoor scene have high spatial variabilities. The spatially rigid global representation from CNN is suboptimal. 2) The cluttered indoor scene has lots of redundant and noisy semantic cues; thus discerning discriminative information among them should not be ignored. 3) Directly concatenating or summing global RGB and Depth information as presented in popular methods cannot fully exploit the complementarity between two modalities for complicated indoor scenarios. To address the above problems, we propose a novel unified framework named Multi-modal Attentive Pooling Network (MAPNet) in this paper. Two orderless attentive pooling blocks are constructed in MAPNet to aggregate semantic cues within and between modalities meanwhile maintain the spatial invariance. The Intra-modality Attentive Pooling (IAP) block aims to mine and pool discriminative semantic cues in each modality. The Cross-modality Attentive Pooling (CAP) block is extended to learn different contributions across two modalities, which further guides the pooling of the selected discriminative semantic cues of each modality. We further show that the proposed model is interpretable, which helps to understand mechanisms of both scene classification and multi-modal fusion in MAPNet. Extensive experiments and analysis on SUN RGB-D Dataset and NYU Depth Dataset V2 show the superiority of MAPNet over current state-of-the-art methods.

1. Introduction

Scene classification is one of the fundamental problems in computer vision. The goal of scene classification is to annotate the images with scene classes, such as mountain, football field and classroom. Although many studies focus on the outdoor scene [1–3], recently more and more researches are conducted on the indoor scene [4,5] due to its wide applications in robotics, home intelligence and surveillance. Compared with the outdoor scene, the indoor scene is much more complicated for larger variations in light, shape, layout and severer occlusions. Some of these challenges are intrinsic due to the loss of 3D information in image capturing. Fortunately, with the release of the affordable depth sensors such as Microsoft Kinect, it is promising to overcome some of these problems. In this paper, we focus on the RGB-D indoor scene classification.

Indoor scene images are very different from the object-centric images. An indoor scene is the abstract of various semantic cues which include multiple objects of which classes are open-set and contextual relationships between them. Classifying the scene needs aggregation of all these semantic cues. For example, recognizing the *chair* or the *computer* alone can not classify the scene as the *home office*. The *chair* and the *computer* can also exist in other scene categories such as the *office* and the *computer room*. People thus need to recognize multiple objects in the scene, perceive the context and then aggregate such information to infer the scene as the *home office* correctly. Although Convolutional Neural Network (CNN) achieves superior performance on the classification of object-centric images, it gets less ideal result when directly applied to complex scene images. The reason lies in the fact that the semantic cues could be highly spatially variant in the scene. As shown in Fig. 1(a), the *bed* and the *closet* could be at any positions in the *bedroom*. The CNN features extracted from global scene

* Corresponding author.
E-mail addresses: yabei.li@cripac.ia.ac.cn (Y. Li), zzhang@nlpr.ia.ac.cn (Z. Zhang), breezecheng@tencent.com (Y. Cheng), wangliang@nlpr.ia.ac.cn (L. Wang), tnt@nlpr.ia.ac.cn (T. Tan).

https://doi.org/10.1016/j.patcog.2019.02.005
0031-3203/

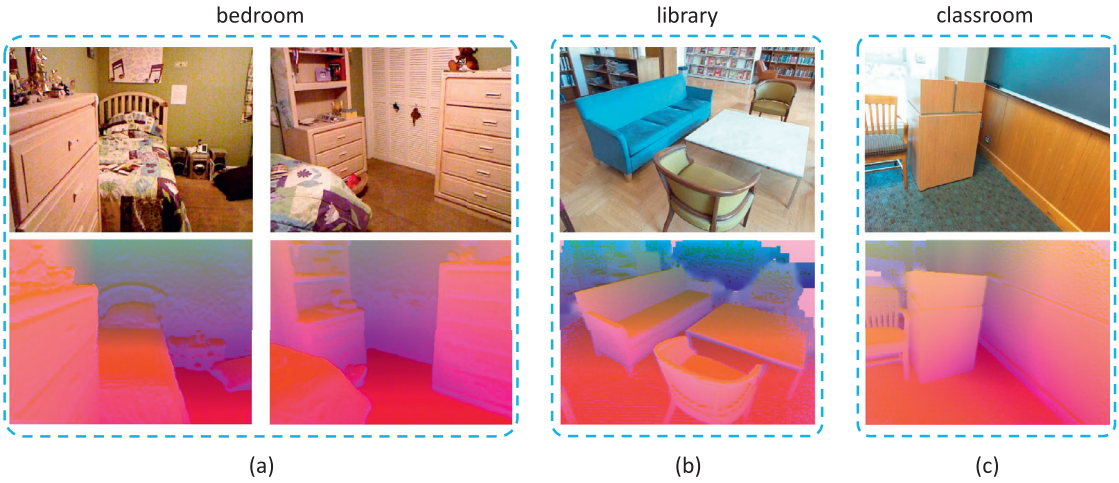


Fig. 1. Illustration of the difficulties for the RGB-D indoor scene classification. The RGB and Depth (We use HHA (Horizontal Disparity, Height above the ground, Angle of surface normal) [15] to encode the depth images in this paper.) pairs in SUN RGB-D Dataset are shown. (a) The semantic cues are highly spatially variable in indoor scene, for example the *bed* and the *closet* can be in any positions in *bedroom*. (b) The indoor scene is cluttered so that some semantic cues are not useful for classification. For example, although the *sofa* is salient in the image, it is not a helpful cue to recognize *library*. (c) The RGB and Depth sensor emphasize different semantic cues. The RGB sensor can well delineate the *blackboard* while the Depth sensor can better capture the shape of the *table*.

images are too spatially rigid to capture the invariance of semantic cues [6].

To address the high spatial variation problem in indoor scene, a lot of previous researches propose to utilize orderless pooling to maintain spatial invariance in aggregation for scene classification. The orderless pooling means that the pooling results are independent to the arrangement of the sequence to be pooled. These methods usually first extract local descriptors in different locations and scales and then use encoding methods, such as Fisher Vector (FV) [7] and Vector of Locally Aggregated Descriptors (VLAD) [8], to aggregate these local descriptors [9–12]. The local descriptors includes Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and features densely extracted from CNN or semantic cues such as objects and semantic attributes. These encoding-based methods, though utilized widely, have two common problems for indoor scene classification. Firstly, the codebook is usually formed by clustering all the local descriptors. The codebook is able to discard some noise. However, for indoor scenes that are always cluttered, local descriptors that commonly appeared are not necessarily helpful for scene classification. For example, as illustrated in Fig. 1(b), although the *sofa* always appears in indoor scenes and it's salient in the image, it's not a useful cue to recognize the scene as *library*. It may even cause the scene misclassified as the *living room*. Secondly, the encoding methods such as Fisher Vector are hard to be integrated into CNN [13,14]. The existing methods usually process feature extraction and feature encoding independently, which also leads to sub-optimal feature representation.

Another issue is how to fuse RGB and Depth information effectively. Traditional methods usually use handcrafted methods to extract the RGB and Depth cues and combine them using summation or concatenation [16–18]. In deep learning based methods, most previous work directly concatenates or sums information of two modalities at different levels [4,15,19,20]. These popular methods are too simple to fully exploit the relationships across modalities and they suffer from two drawbacks when applied to indoor scene classification. Firstly, at global level, the contributions of the RGB and Depth information are presumed to be fixed for all image pairs. However, for different scenarios, RGB and Depth information should have different roles. For example, for images captured in dim scenarios, Depth channels should be paid more attention. While for noisily captured depth images, RGB channels tend to be more reliable. Secondly, at local level, the contributions of the RGB and Depth modalities are set to be the same for all local semantic cues in a single RGB-D image pair. However, the RGB sensor captures more color information and the depth sensor focuses more on geometric information. For example in Fig. 1(c), in the *classroom*, the RGB image can delineate the color and texture features better for the *blackboard* which is ignored by the depth sensor while the depth sensor captures more invariant shape information of the *table* and the *chair* than the RGB sensor.

To address the above issues, in this paper, we propose a unified framework named Multi-modal Attentive Pooling Network (MAPNet) tailored specifically to the task of RGB-D indoor scene classification. The MAPNet firstly obtains a set of local semantic cues by extracting CNN features on region proposals. To compute efficiently, we use ROI pooling [21] on the top of last convolutional layer to obtain features of the corresponding region proposals. The local features which represent local semantic cues are aggregated in two stages. In the first stage, the local semantic cues in each modality are pooled by Intra-modality Attentive Pooling (IAP) block. To select out discriminative semantic cues, the attention mechanism is incorporated into orderless pooling. Two attentive pooling strategies, the class-agnostic attentive pooling and class-aware attentive pooling, are further explored and compared in IAP block. In the second stage, the local semantic cues across two modalities are pooled by Cross-modality Attentive Pooling (CAP) block, which extends the attention mechanism to learn different contributions across modalities for different types of semantic cues. The obtained contributions across two modalities further guide the pooling of the selected discriminative semantic cues from two modalities. Although the MAPNet is learned in two stages, the parameters are optimized in an end-to-end manner. Simple yet effective, the proposed framework is easy to re-implement.

Our contributions can be summarized in fourfold. The first contribution is towards semantic cues aggregation. Instead of directly using fully connected layers or encoding methods such as FV and VLAD in aggregation, we propose attentive pooling structures which can maintain spatial invariance meanwhile discern discriminative semantic cues. The second contribution addresses

the RGB and Depth data fusion. Unlike other popular methods that directly presume equal contributions for RGB and Depth modalities, the proposed fusion model allows for adjustments of contributions across RGB and Depth for different discriminative semantic cues. Thirdly, we show that the proposed MAPNet is interpretable. Through visualization, we analyze the mechanisms of both scene classification and RGB-D fusion. Finally, we verify our methods on two popular RGB-D datasets: SUN RGB-D Dataset and NYU Depth V2 Dataset. The proposed MAPNet achieves the state-of-the-art results on both datasets.

2. Related work

2.1. Scene classification

Most existing methods for scene classification consist of three steps: 1) learn feature representations in different locations or scales of scene images, 2) aggregate or pool these features to obtain the scene representation, 3) based on the representation, learn some widely used classifiers such as Support Vector Machine (SVM) and Neural Network. For the first step, traditional methods extract hand-crafted features such as SIFT [22], GIST [23] and HOG [24]. In deep learning based methods, some studies [25] directly extract features from the penultimate fully connected layer of the CNN model that is pre-trained on the ImageNet or Places Dataset [26]. Some other studies [1,27,28] propose to use the posterior probabilities of objects in the scene to represent the scene image. For the aggregation step, some existing researches use information concatenation [1,25,29]. The concatenation usually leads to high-dimensional representation and thus will increase the computing consumption. Besides, these methods concatenate all local semantic cues without information selection, which may lead to models less robust towards redundant and noisy information in indoor scenes [30,31]. The mainstream aggregation methods for scene classification, instead, use encoding-based pooling methods such as FV and VLAD [10,11,32,33]. The encoding methods, though effective and popular, for they can overcome the spatial variabilities in scene classification, tend to ignore information selection in aggregation and are hard to be integrated into an end-to-end deep learning framework. Recently, some researches [11,13,14] make efforts to implement the encoding process into an end-to-end network. However, they either use the approximation in encoding or need complex implementation. In this paper, instead of applying simple concatenation or using encoding strategies, we explore and propose simple but effective attentive orderless pooling structures for aggregation in scene classification. Incorporated with the attentional mechanisms, it can attend on discriminative semantic cues in the pooling and it's optimized in an end-to-end manner.

2.2. RGB and depth data fusion

With the release of affordable depth sensors recently, how to utilize the complementary information in RGB and Depth modalities attracts a lot of researchers. Most previous methods on RGB and Depth data fusion can be summarized into three categories: image-level fusion [4], feature-level fusion [19], score-level fusion [34]. These methods all regard the RGB and Depth information with the same and equal contributions for all images. Recently, some researchers propose to further emphasize the correlations between the RGB and Depth [35,36]. However, these methods all focus on the relations in aggregated feature level. Nevertheless, enforcing all the local semantic cues in two modalities to follow the same relations in fusion is not suitable for indoor scene images. Furthermore, all the above mentioned methods are lack of interpretability. In this paper, the proposed framework models the relations between RGB and Depth modalities at local level, adapting the attention weights across the two modalities for different local semantic cues. The proposed method is able to explicitly visualize the contributions from RGB and Depth modalities, providing better understanding of the fusion process.

2.3. Attention mechanisms

One of the elaborate designations of the human visual system is the attention mechanism. Instead of absorbing all the information in a fixed procedure, human attention mechanisms choose most salient features adaptively for certain requirements. Recently, a lot of studies show that by incorporating attention mechanisms into computer vision, the context information can be better utilized [37,38]. The common idea is to read external memories/contexts through an attention scheme. These architectures usually employ the Long Short Term Memory network (LSTM) [39] to model the sequential input and output. In our case on RGB-D scene classification, the attention mechanism is used for pooling the semantic cues and fusing multi-modal information. In practice, the local semantic cues can be viewed as memories, and the context can be aggregated from this information. Recently, some papers also propose to incorporate attention mechanisms into the pooling operation. Yang et al. [40] propose to aggregate features of frames in videos to emphasize more discriminative frames for video face recognition. Girdhar and Ramanan [41] propose to use the combination of bottom-up saliency and top-down attention to approximate the second order pooling for action recognition. The attention mechanism architecture in this paper is different from them. Firstly, our attention model aims to aggregate the local semantic cues in each modality and fuse semantic cues across the multiple modalities for RGB-D indoor scene classification. Secondly, the proposed attentive pooling model is class-aware. The attention weight of certain semantic cues varies according to different potential scene categories. Moreover, the proposed model does not need extra information, unlike Girdhar and Ramanan [41], in which the bottom-up saliency needs to be learned from human pose key points in the implementation of attentive pooling.

3. Method

3.1. Overview of framework

The framework of the proposed MAPNet for RGB-D indoor scene classification is illustrated in Fig. 2. The local semantic cues are firstly extracted on RGB and Depth pairs. The MAPNet then aggregates the RGB and Depth local semantic cues in two learning stages. The first learning stage focuses on discriminative semantic cues aggregation for each modality of the scene image. The second learning stage aims to aggregate selected discriminative semantic cues across RGB and Depth modalities.

To extract local semantic cues in each modality, we propose to utilize region proposals of the scene image. The public available region proposal extractor [42] is employed, which needs no extra bounding box annotations and object labels. The region proposals can be produced on single modal image or RGB and Depth image pair. In practice, since the extractor [42] is originally designed for color images, we generate region proposals based on the RGB image. Notice that other region proposal extractors and strategies can also be integrated into MAPNet off the shelf.

Suppose a pair of RGB and Depth image (x^r, x^d), the region proposals extracted on the x^r are $\{p_1, \ldots, p_i, \ldots, p_L\}$, where p_i represents position of the ith region proposal in x^r and L is the number of region proposals. The regions the ith proposal crops out on the image x^r and x^d are denoted as $x^r_{p_i}$ and $x^d_{p_i}$. The local semantic cues in $\{x^r_{p_i}\}$ and $\{x^d_{p_i}\}$ include part of object, single object and multiple co-occurrent objects. After we get regions of interest, two Convolutional Neural Networks (CNNs) are employed to

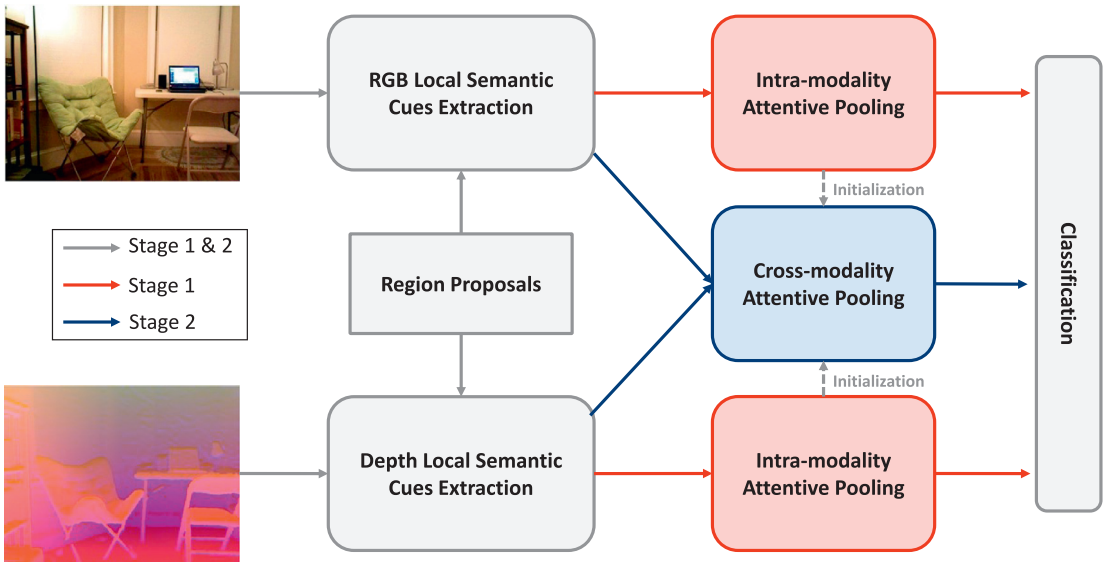


Fig. 2. Illustration of the whole framework of MAPNet. The MAPNet is learned in two stages. The first learning stage aggregates discriminative semantic cues in each modality by using Intra-modality Attentive Pooling block. The second learning stage aggregates selected discriminative semantic cues of each modality across modalities by using Cross-modality Attentive Pooling block.

extract features for $\{x^r_{p_i}\}$ and $\{x^d_{p_i}\}$, respectively. We use AlexNet [43] as our backbone CNN structure. The AlexNet consists of 5 convolutional layers denoted as $conv_1 - conv_5$ and 3 fully connected layers denoted as $fc_6 - fc_8$ in our paper. For each modality, we put all the cropped regions $\{x_{p_i}\}$ of an image through the CNN and use the outputs from the fc_6 layer as the local semantic cues. In order to obtain representations of the proposed regions efficiently, we use Region of Interest (RoI) pooling [21] on the global representations $conv_5(x^r)$ and $conv_5(x^d)$ from the last convolutional layers of the networks. Instead of calculating the representations for each proposal, the ROI pooling crops the representation of each region proposal from global representation from $conv_5(x)$. Each cropped local representation is max-pooled to small feature map with fixed size (such as 7×7) in ROI pooling. The obtained local representations are further transformed by the fully connected layer fc_6, the outputs of which are denoted as $\{\mathbf{f}^r\}$ and $\{\mathbf{f}^d\}$ for RGB and Depth respectively. The features of the proposed regions $\{\mathbf{f}^r\}$ and $\{\mathbf{f}^d\}$ can represent various local semantic cues in RGB and Depth scene image.

In the first learning stage, the local representations $\{\mathbf{f}^r\}$ and $\{\mathbf{f}^d\}$ are aggregated to obtain the scene classification decision in each modality respectively by the Intra-modality Attentive Pooling (IAP) block. The IAP blocks utilized in two modalities are of the same structure but do not share weights. For brevity, we denote either x^r or x^d as x in the first learning stage. The details of the IAP block will be described in Section 3.2.

The network constructed in the second learning stage consists of two streams, operating on RGB and Depth image respectively. Each stream is initialized with the weights optimized in the first stage learning. The Cross-modality Attentive Pooling (CAP) block is constructed to aggregate local representations $\{\mathbf{f}^r\}$ and $\{\mathbf{f}^d\}$ from both modalities based on the learned aggregation of each modality in the first learning stage. The details of the CAP block will be described in Section 3.3.

Towards network optimization, the classification loss is designed based on the outputs of the IAP and CAP in two learning stages. In the first learning stage, two networks for RGB and Depth local semantic cues aggregation are optimized independently. Although trained in two stages, the parameters in MAPNet are optimized in an end-to-end manner.

3.2. Intra-modality attentive pooling block

The inputs of the Intra-modality Attentive Pooling (IAP) block are local representations $\{\mathbf{f}_i\}$, where $i = 1, 2, \ldots, L$, and L is the number of region proposals. The IAP block aims to aggregate them and outputs the classification decision.

3.2.1. Average pooling

As shown in Fig. 1(a), the local semantic cues (*closet* and *bed*) can appear anywhere in the scene image and the locations of these semantic cues do not influence the definition of the scene as *bedroom*. Thus we assume the aggregation of local semantic cues is invariant to spatial locations. The most common operation to achieve this goal is average pooling.

The structure to achieve average pooling strategy in the IAP block is designed as illustrated in Fig. 3(a). Specifically, it consists of fully connected layers and average pooling operator. In the structure, the $\{\mathbf{f}_i\}$ are firstly transformed through fully connected layer $fc_7(\cdot)$, the outputs of which are represented as $\{fc_7(\mathbf{f}_i)\}$. The average pooling operator is then conducted to the $\{fc_7(\mathbf{f}_i)\}$ to obtain the aggregated representation. Ultimately, the final classification logits can be obtained by using another fully connected layer $fc_8(\cdot)$ to transform the aggregated representation. The average pooling strategy of the IAP block can be written as:

$$IAP_{ave\_pooling}(x) = fc_8\left(\frac{1}{L}\sum_{i=1}^{L} fc_7(\mathbf{f}_i)\right) \tag{1}$$

3.2.2. Class-agnostic attentive pooling

Although the average pooling strategy is easy to implement and has been widely used, it always focuses more on salient regions where the proposals are frequently extracted. However, the indoor scene images are captured from different viewpoints, the most salient region do not necessarily represent the most discriminative semantic cues for indoor scene classification. Moreover, some of these salient regions may mislead classification. For example, the salient region *computer* is not discriminative enough to discern the *office* and the *computer room*.

To overcome this problem, we incorporate the attention mechanism to enhance the discriminative capability of pooling. The intu-

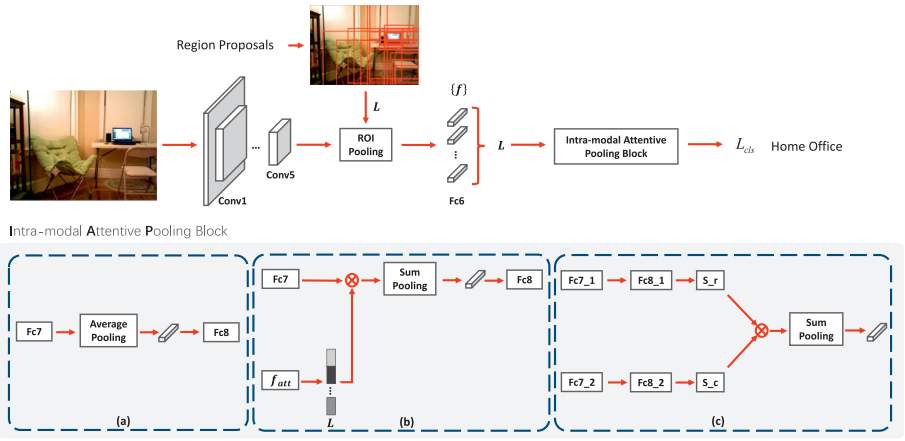


Fig. 3. The detailed MAPNet structure in the first learning stage. The input image can be RGB or HHA image. Three pooling strategies for Intra-modality Attentive Pooling (IAP) block are explored and compared: (a) average pooling. (b) class-agnostic attentive pooling. (c) class-aware attentive pooling. The "S_c" represents softmax operation across the attention scores of region proposals and the "S_r" represents softmax operation across the classification scores of each region proposal.

ition behind this is that human being recognizes a scene by selecting and attending on some key parts of the scene. The structure that achieves class-agnostic attentive pooling in the IAP block is shown in Fig. 3(b). In the block, an extra attention branch is constructed to read all local features $\{\mathbf{f}_i\}$ and output the importance of each proposed region for indoor scene classification. The attention branch can be written as:

$$e_i = f_{att}(\mathbf{f}_i) \tag{2}$$

$$a_i = \frac{\exp(e_i)}{\sum_{i=1}^{L}\exp(e_i)} \tag{3}$$

where the specific form of $f_{att}(\cdot)$ can be represented by two-stacked fully connected layers (1024-1). $\{e_i\}$ are passed to the softmax operator to generate positive weights $\{a_i\}$ with $\sum_{i=1}^{L} a_i = 1$. The a_i represents the importance of the semantic cues in the ith proposed region regardless of scene class. After obtaining the attention weights, the class-agnostic attentive pooling strategy in the IAP block can be represented as:

$$IAP_{cag\_pooling}(x) = fc_8\left(\sum_{i=1}^{L} g(\mathbf{f}_i)\circ a_i\right) \tag{4}$$

where the operation $\circ$ means scalar multiplication and $g(\cdot)$ represents the linear transformation for local features. In practice, $g(\cdot)$ is formed by a fully connected layer with an output dimension of 1024.

3.2.3. Class-aware attentive pooling

The class-agnostic attentive pooling assigns different weights to different proposed regions, enabling the discriminative semantic cues to be salient in aggregation. However, it does not consider class information when learning the attention weights. The $\{a_i\}$ are learned regardless of the scene classes. However, intuitively, the same semantic cue has different roles for different types of scenes. For example, the *book* is a vital semantic cue for *library* while it is not such an important clue for *living room*. To address such problem, as illustrated in Fig. 3(c), we design the class-aware attentive pooling strategy in IAP block inspired by Bilen and Vedaldi [44].

The key difference between the class-aware attentive pooling and class-agnostic attentive pooling lies in the attention branch. The attention branch in class-aware attentive pooling learns class-specific attention weights $\{\mathbf{a}_i\}$, where $\mathbf{a}_i$ is a vector with dimension C. The element a_{ic} of $\mathbf{a_i}$ represents the attention weight of the ith proposed region for the cth scene class. $\mathbf{a}_i$ can be computed by:

$$\mathbf{e}_i = t_{att}(\mathbf{f}_i) \tag{5}$$

$$\mathbf{a}_i = \frac{\exp(\mathbf{e}_i)}{\sum_{i=1}^{L}\exp(\mathbf{e}_i)} \tag{6}$$

where the form of $t_{att}(\cdot)$ can be represented by two-stacked fully connected layers (1024-C). The softmax operator is conducted on $\{\mathbf{e}_i\}$ to generate positive weights and guarantee $\sum_i a_{ic} = 1$.

Due to $\mathbf{a}_i \in \mathbb{R}^C$, directly combining the class specific attention weights with the local features $\mathbf{f}_i \in \mathbb{R}^D$, where D is feature dimension, is not compatible. Observing that due to the linearity of the pooling operator, the attention weights can be applied to aggregate not only the local features but also the classification logits of proposed regions. The class-aware attentive pooling structure in IAP block is thus designed and represented as:

$$\mathbf{b}_i = g(\mathbf{f}_i) \tag{7}$$

$$IAP_{caw\_pooling}(x) = \sum_{i=1}^{L}\mathbf{a}_i \odot \mathbf{b}_i \tag{8}$$

where $\mathbf{b}_i \in \mathbb{R}^C$ and $\odot$ means the element-wise multiplication. $g(\cdot)$ here is constructed by two stacked fully connected layers (1024-C) and the softmax operator. Suppose the outputs from the two stacked fully connected layers are $\{\mathbf{o_i}\}$, the softmax operator is applied to $\{\mathbf{o_i}\}$ to generate the classification logits $\{\mathbf{b_i}\}$, where for each i, $\sum_c b_{ic} = 1$. The $\mathbf{b_i}$ represents the probability of the scene belonging to a specific scene category in terms of semantic cues in the ith propose regions. The $IAP_{caw\_pooling}(x)$ aggregates the local semantic cues by utilizing class-aware attention weights to sum the classification probabilities of the discriminative semantic cues.

The class-aware attentive pooling can be also regarded as processing the local features with two branches, the attention branch and the classification branch. When the MAPNet classifies indoor scenes with class aware attentive pooling, the attention branch outputs $\{\mathbf{a_i}\}$ which represent which semantic cues should be attended more on for a specific scene class. For example, for the illustrated image in Fig. 3, the attention branch chooses where to focus on when classifying the scene as *home office*. The proposed semantic cues include *fluffy chair, computer and table, floor*, etc. The classification branch, on the other hand, outputs $\{\mathbf{b_i}\}$ that represent the probabilities of the scene image belonging to different categories according to the proposed semantic cues, e.g., the probability of the illustrated image captured in *home office* given the

semantic cue *computer*. The scene is inferred as *home office* only when the selected discriminative semantic cues in the attention branch are also decided as strong evidence for *home office* in the classification branch.

The parallel branches structure in class-aware attentive pooling is similar to that in [44,45]. However, the aim of our task is different from theirs. Bilen and Vedaldi [44] aim to solve the weakly supervised detection problem, where the multiple instance learning process is improved by using the parallel branches to select regions. The region scores are utilized to infer the object detection results. Cui et al. [45] use the attention over attention structure to improve the reading comprehension task. The motivation is to exploit mutual information between the document and query. Our work, on the other hand, aims to use parallel branches for class-aware attentive pooling, so as to further address the discriminative semantic cues aggregation problem in indoor scene classification.

3.2.4. Optimization for IAP

For average pooling and class-agnostic attentive pooling, the optimization goal is set as softmax loss (defined as the combination of the softmax operation and category cross entropy loss):

$$L_{softmax} = -\frac{1}{N}\sum_{n}\log\left(\frac{\exp(IAP(x_n)_{y_n})}{\sum_{j=1}^{C}\exp(IAP(x_n)_j)}\right) \tag{9}$$

where N is the number of the training images and y_n is the scene class label for the nth training image.

For class-aware attentive pooling, the output logits are the combination of probabilities. The logits in the output are not mutually exclusive. In practice, if the category cross entropy loss is employed to train the network, all the elements in the $IAP(x)$ will be optimized to be nearly 1. To overcome this problem in training, we instead use the binary cross entropy loss as follows:

$$L_{cls} = -\frac{1}{N}\sum_{n}\left(\sum_{j=1}^{C}\mathbb{1}(y_n \neq j)\log(1 - IAP(x_n)_j) + \sum_{j=1}^{C}\mathbb{1}(y_n = j)\log(IAP(x_n)_j)\right) \tag{10}$$

where $\mathbb{1}(\cdot)$ means the indicator function, which only equals 1 when the condition in the parenthesis is satisfied. In the later sections, Eq. (10) is used as the classification training loss.

3.3. Cross-modality attentive pooling block

For RGB-D indoor scene classification, another challenge is multi-modal fusion of RGB and Depth information. In the multi-modal scenario, the indoor scene can be regarded as the abstract of local semantic cues from both RGB and Depth modalities. Each modality emphasizes different types of semantic cues. Based on the above observations, the Cross-modality Attentive Pooling (CAP) block is constructed as illustrated in Fig. 4.

The CAP block is an extension of the IAP block for multi-modal local semantic cues aggregation. The inputs of the CAP block are RGB local semantic cues $\{\mathbf{f}^r\}$ and Depth local semantic cues $\{\mathbf{f}^d\}$ from the RGB stream and Depth stream respectively. Both RGB and Depth streams in the CAP block have the same structure as that of the IAP block in Fig. 3(c). The aggregation of local discriminative semantic cues in each modality has been learned in the IAP block in the first learning stage. To further fuse the learned discriminative semantic cues across the two modalities, we propose an additional modality attention branch (where $\{fc_7(\mathbf{f}^r)\}$ and $\{fc_7(\mathbf{f}^d)\}$ flow in illustrated in Fig. 4) in CAP block to learn the contributions of RGB and Depth modalities for each local semantic cue. All outputs from the penultimate fully connected layers in classification branches and attention branches in RGB and Depth streams are concatenated as the inputs $\{\mathbf{q}_i\}$ of the modality attention branch. The modality attention branch can be represented as:

$$\mathbf{u}_i = h(\mathbf{q}_i) \tag{11}$$

$$z_i^m = \frac{\exp(u_{im})}{\sum_{m\in r,d}\exp(u_{im})} \tag{12}$$

where $h(\cdot)$ is a linear transformation and m denotes the modality. $\mathbf{u}_i \in \mathbb{R}^2$. The z_i^r and z_i^d represent the importance of RGB and Depth information for the ith local semantic cue respectively. $\mathbf{z_i} = [z_i^r, z_i^d]$. In practice, $h(\cdot)$ is constructed by two stacked fully connected layers, of which the output numbers are 256 and 2. The softmax operator here is utilized to guarantee that for each semantic cue, the sum of attention weights for RGB and Depth modalities equals 1.

The z_i^r and z_i^d are further utilized to guide the aggregation of semantic cues in the two modalities by fusing the attention branches and classification branches in the two streams. As illustrated in Fig. 4, the outputs from the two branches in the two streams can be denoted as $\{\mathbf{e}_i^r\}$, $\{\mathbf{o}_i^r\}$ and $\{\mathbf{e}_i^d\}$, $\{\mathbf{o}_i^d\}$ respectively. The cross-modality attentive pooling can thus be represented as:

$$a_{ic}^* = \frac{\exp(e_{ic}^r \cdot z_i^r + e_{ic}^d \cdot z_i^d)}{\sum_{i=1}^{L}(e_{ic}^r \cdot z_i^r + e_{ic}^d \cdot z_i^d)} \tag{13}$$

$$b_{ic}^* = \frac{\exp(o_{ic}^r \cdot z_i^r + o_{ic}^d \cdot z_i^d)}{\sum_{c=1}^{C}(o_{ic}^r \cdot z_i^r + o_{ic}^d \cdot z_i^d)} \tag{14}$$

$$CAP(x^r, x^d) = \sum_{i=1}^{L}\mathbf{a}_i^* \odot \mathbf{b}_i^* \tag{15}$$

where $\mathbf{a}_i^* = [a_{i1}^*, a_{i2}^*, \ldots, a_{iC}^*]$, $\mathbf{b}_i^* = [b_{i1}^*, b_{i2}^*, \ldots, b_{iC}^*]$, $CAP(x^r, x^d) \in \mathbb{R}^C$, and $\odot$ means the element-wise multiplication.

The attention weights $\mathbf{z_i}$ are predicted based on a single proposed region, which cannot guarantee the consistency with its spatially neighboring proposed regions. Intuitively, if the RGB cues contribute more to one proposed region, people can infer that the RGB cues should also be attended more on its neighboring regions. A smoothness constraint is then proposed to describe this relationship.

$$L_{smoothness} = \frac{1}{NL}\sum_{n=1}^{N}\sum_{i=1}^{L}\sum_{k\in K_i}\frac{1}{2}(a_{iy_n}^* \cdot b_{iy_n}^*)^2(\mathbf{z}_k - \mathbf{z}_i)(\mathbf{z}_k - \mathbf{z}_i)^T \tag{16}$$

where y_n is the label of the nth training image, K_i indicates the neighboring proposed regions of the ith proposed region. We define the neighboring proposed regions as the regions that have at least 0.6 Intersection over Union (IoU) with the given proposed region. The term $(a_{iy_n}^* \cdot b_{iy_n}^*)$ measures the importance of the ith proposed region for correct classification. The more important is the proposed region, the stricter constraints are for the neighboring regions of it. The reason is that the neighboring proposed regions of the high-score proposed region are also more informative for classification. In practice, constraining all the proposed regions to be consistent is time-consuming. Instead, we only constrain the proposed region with the highest score. The number of the neighboring proposed regions are constrained to be at most 10. The selection of the neighbors k is based on $(a_{ky_n}^* \cdot b_{ky_n}^*)$, i.e., their importance to the classification. Neighbors with high importance are selected preferentially. We find that this strategy is effective enough in our experiments.

The second-stage training of the proposed MAPNet is optimized by the following function:

$$L = L_{cls} + \beta L_{smoothness} \tag{17}$$

In the test phase, only the architecture constructed in the second learning stage is used for inference.

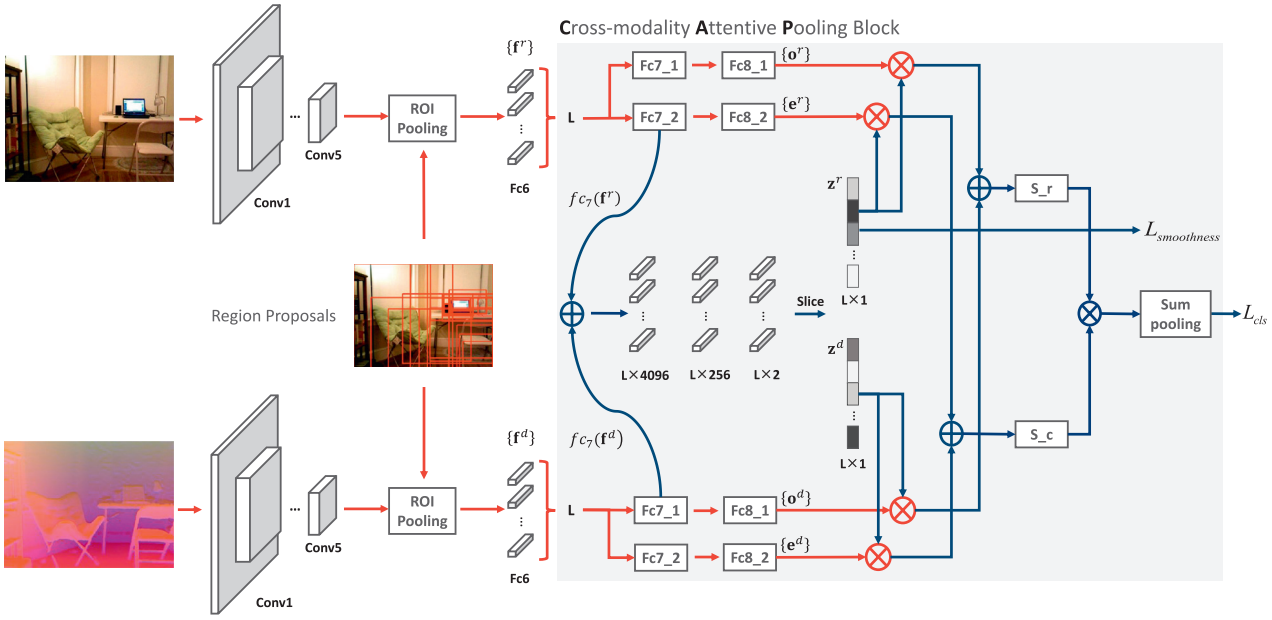


Fig. 4. The MAPNet structure in the second learning stage. In Cross-modality Attentive Pooling (CAP) block, the attention weights across RGB and Depth modalities for each proposed region are learned to guide the aggregation of discriminative semantic cues across modalities. The parameters that flow through the red arrows are initialized with the weights learned in the first stage. The parameters that flow through the blue arrows are initialized from scratch. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Experiments

The proposed method is validated on two popular indoor RGB-D scene classification datasets: SUN RGB-D Dataset [5] and NYU Depth Dataset V2 [46].

4.1. Datasets

The SUN RGB-D Dataset contains 10,355 RGB and Depth image pairs captured from different cameras including Kinect v2, RealSense, Kinect v1 and AsusXtion. We follow the experimental settings in [5]. The categories including more than 80 images are used as scene categories. 19 categories are kept for our experiments with 4845 images for training and 4659 images for test.

The NYU Depth Dataset V2 includes 1449 RGB and Depth image pairs. To compare with other methods, we follow the experimental settings in [16], in which the original 27 categories are further gathered into 10 categories (including one "other" category). The original training and test splits are kept, which results in 795 training images and 654 test images for 10 scene categories.

4.2. Implementation details

We implement the whole architecture in the popular framework Caffe [47]. To fairly compare the results with others, the AlexNet is used as the base architecture. We encode Depth image into three-channel representation HHA (Horizontal Disparity, Height above the ground, Angle of surface normal) [15] which has been widely utilized and proven effective in RGB-D indoor scene understanding [6,36,48,49]. The whole architecture is trained in two stages. Stochastic Gradient Descent (SGD) and "step" policy are used to train the framework in both stages. For each RGB-D image pair, 300 region proposals are extracted. Since we use the ROI pooling strategy, the input images can be in arbitrary sizes. In our implementation, to keep more detailed semantic cues, the input image pairs are resized based on the principle that the shorter axis of the image is resized to 480 and the aspect ratio is kept. For the SUN RGB-D Dataset, in the first stage, the weights are initialized using Places CNN and the batch size n_0, initial learning rate γ_0, stepsize n_s and max iterations N_t are respectively set as 1, 0.001, 20000, 70,000 for both RGB and Depth modality. In the second stage, n_0, γ_0, n_s and N_t are set as 10, 0.0001, 4000, 10000. For the NYU Depth Dataset V2, since the number of samples is very small, following the training strategy in [6,50], we initialize the first stage's network for RGB and depth with Places CNN fine-tuned on the corresponding modality of the SUN RGB-D Dataset. The n_0, γ_0, n_s and N_t are set as 1, 0.0001, 4000, 10,000 in the first stage and they are modified as 10, 0.00001, 4000, 10,000 in the second stage. The weights between the classification loss and the smoothness loss β are set as 0.01 and 0.1 for the SUN RGB-D Dataset and the NYU Depth Dataset V2 respectively through cross validation. Both the SUN RGB-D Dataset and NYU Depth Dataset V2 have highly imbalanced number of images between classes. To train the network effectively for the class that have fewer images, we use frequency weighted classification loss:

$$L_{cls} = -\frac{1}{N}\sum_{n} w(y_n)\left(\sum_{j=1}^{C}\mathbb{1}(y_n \neq j)\log(1 - IAP(x_n)_j) + \sum_{j=1}^{C}\mathbb{1}(y_n = j)\log(IAP(x_n)_j)\right) \quad (18)$$

where $w(t)$ is defined as:

$$w(t) = \frac{N_{c\_\max} - N_{c\_\min}}{N_t - N_{c\_\min} + \delta}$$

in which N_t is the number of images of class t in the training set. $c\_\min$ and $c\_\max$ represents the classes with the least and the most number of training images, respectively. δ is empirically set as 0.01.

4.3. Results on SUN RGB-D dataset

4.3.1. Comparison with state-of-the-art methods

We compare our final results with the state-of-the-art results as shown in Table 1. Our model achieves superior results compared with other methods. For local features aggregation, Wang et al. [35], Liao et al. [51] and Zhu et al. [36] use the fully connected layers, which cannot maintain the spatial invariance. Wang

Table 1
Performance comparison with the state-of-the-art methods on SUN RGB-D dataset.

Methods	Extra annotation	Accuracy(%)
(Wang et al. 2015 [35])	No	26.5%
(Liao et al. 2016 [51])	Yes	41.3%
(Zhu et al. 2016 [36])	No	41.5%
(Wang et al. 2016 [6])	No	48.1%
(Song et al. 2017 [48])	No	52.4%
(Song et al. 2017 [52])	No	52.3%
Local + OOR (Song et al. 2017 [50])	Yes	50.3%
MAPNet (ours)	**No**	**54.6%**
Local + Global + OOR (Song et al. 2017 [50])	Yes	54.0%
MAPNet + Global(ours)	**No**	**56.2%**

et al. [6] share some common ideas with our work. They also use region proposals to extract local semantic cues, however, they propose to use the FV to aggregate local information. It enhances the discriminative power by regularizing the classifier weights of the Gaussian Mixture Model components. Our model, on the other hand, learns the discriminative semantic cues in aggregation using the attention branch. The results confirm that our model can learn the discriminative information better through the end-to-end optimization. Another work that shares similar idea with us is Song et al. [50]. They propose to detect objects in the scene and construct the object-to-object relations as semantic cues. Average pooling then is employed to combine these local cues. Additional object bounding box annotations, which are costly to obtain, are required in [50] to learn to detect objects in the scene. Since the object classes in the scene are open-set, annotations for the objects in practice are also limited. We will show through visualization that benefiting from the designed attention mechanism and the end-to-end learning, the proposed MAPNet can latently discover some object co-occurrence and object-to-object spatial relationships, without requiring extra object annotations. For RGB and Depth fusion, all the state-of-the-art methods concatenate the aggregated representation of RGB and Depth or build relationships on aggregated feature level. The MAPNet first proposes to model the relations between modalities for different RGB and Depth local semantic cues. With the proposed aggregation and multi-modal fusion model, our final test result on the SUN RGB-D Dataset achieves 54.6% accuracy, outperforming all the state-of-the-art methods.

To fairly compare with Song et al. [50] which combines both the aggregated local features and global features for RGB-D indoor scene classification, we also incorporate global information in our model. In practice, the global model we utilize is a two-stream CNN, each stream of which represents one modality and is also constructed based on the AlexNet model. The local model and the global model are fused in the score level and trained in an end-to-end manner. The combined model can further achieve 56.2% accuracy. The obtained further improvement suggests the global model can provide complementary global spatial layout information for the MAPNet.

The confusion matrix is illustrated in Fig. 5. It shows the ground truth classes of most misclassified images are semantically similar with the inferred classes. For example, *lecture theatre* is liable to be confused with *classroom*. Notice the accuracy of *study space* is much lower than other categories, for there are much less training samples for *study space*. Other methods [6,36] that report results on the SUN RGB-D Dataset also have lower performances on *study space*.

4.3.2. Comparison of different pooling strategies in IAP

In Section 3.2, three pooling strategies are proposed in Intra-modality Attentive Pooling block for discriminative semantic cues aggregation. Here, we quantitatively test their effectiveness. The results are shown in Table 2. The *baseline* model uses Alexnet as the backbone and is initialized with Places CNN in [26]. It can be regarded as using the fully connected layer to aggregate local information. The *average pooling* corresponds to the structure in Fig. 3(a). As shown in Table 2, it achieves lower performance than the baseline model. The reason is that there are a lot of redundant local semantic cues, some of which are disturbing for correct classification. The *class-agnostic attentive pooling* corresponds to the structure in Figure 3(b). By adding the attention branch to discover the discriminative semantic cues, the performance increases dramatically to 45.2% and 40.0% for RGB and Depth modalities respectively. Finally, the *class-aware attentive pooling* corresponding to the structure in Fig. 3(c), further improves the accuracy to 46.0% and 40.8% for RGB and Depth modalities attributed to the considerations of class information in the attention branch.

Table 2
Comparison of different pooling strategies in Intra-modality attentive pooling (IAP) block.

Accuracy(%)	RGB	Depth
Baseline	42.1%	38.9%
Average pooling	40.1%	38.5%
Class-agnostic attentive pooling	45.2%	40.0%
Class-aware attentive pooling	46.0%	40.8%

4.3.3. Visualization for the intra-modality attentive pooling

The deep learning models have been criticized for the lack of interpretability for a long time. One of the advantages of the proposed attentive pooling method is that it can clearly illustrate the decision-making process for scene classification by utilizing the attention weights to visualize the importance of the proposed regions corresponding to local semantic cues. As an illustration, the class-aware attentive pooling in IAP block is employed in visualization. Assume the evaluated image is inferred as class c, each proposal's score is calculated as $a_{ic} \cdot b_{ic}$. The a_{ic} and b_{ic} are defined by Eqs. (6) and (7). In Fig. 6, the top 3 most influential regions are shown with different colors. The red bounding box represents the most influential region; the blue shows the least influential region among the top 3. The scores of the three regions are listed below each picture.

In Fig. 6, we can see the IAP can automatically discover the most informative semantic cues in different scales for indoor scene classification: 1) For the scenes with some highly discriminative objects, the IAP will highlight these discriminative objects in the scene. For example, in *bathroom*, the *toilet* and *sink* are always assigned strong attention than other objects. It is worth noting that the IAP can also find some discriminative objects that have rarely been annotated in object detection datasets, such as the *toilet paper roll* in *bathroom*. 2) For scenes that need co-occurrence cues to be recognized, IAP assigns higher scores for highly discriminative regions within which multiple relevant objects occur together than the regions only containing less discriminative individual objects within the regions. The typical example is that for *library*, the region containing both books and chairs gets higher scores than the regions only containing books or chairs. The reason is that the books and chairs alone can also be evidence for other scenes such as *bedroom* and *office*. 3) For scenes that need object-to-object spatial relations to be distinguished from each other, the IAP tends to select regions within which discriminative spatial relations between objects exist. For example, for *dining area* and *dining room*, they all consist of tables and chairs, but the spatial relations between tables and chairs are different in the two scenes. The *dining area* usually have multiple dining tables, and they are arranged in rows, while the *dining room* only has one table and the chairs are laid around the table. Notice the IAP does not explicitly learn the

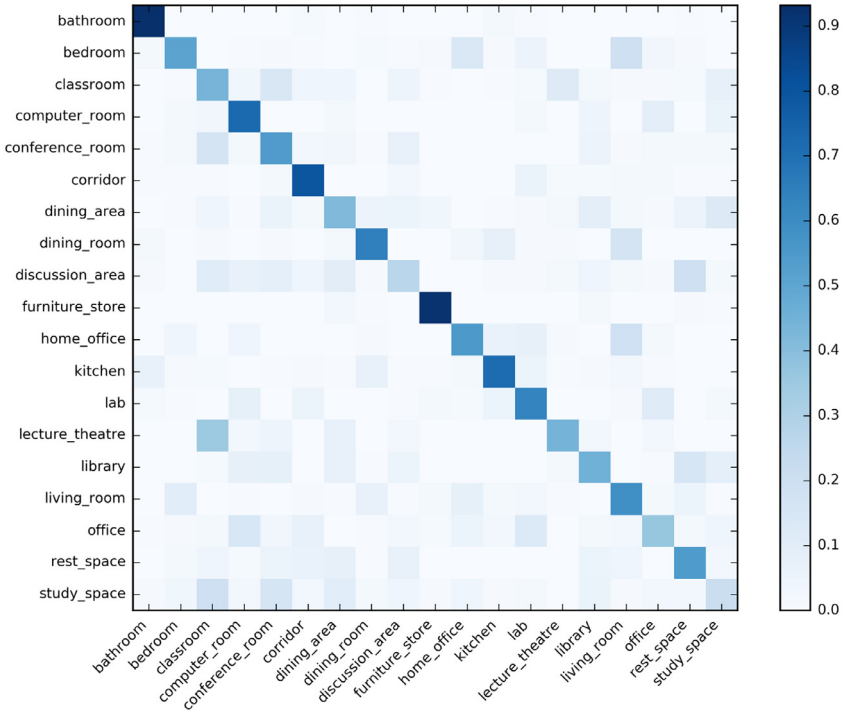


Fig. 5. The classification confusion matrix of MAPNet on the SUN RGB-D Dataset. The vertical axis shows the ground-truth classes. The classes in the horizontal axis are in the same order with that in the vertical axis. The horizontal axis shows the predicted classes.

Table 3
Ablation study for MAPNet on SUN RGB-D dataset.

Methods	Accuracy(%)
Baseline RGB+D (Global)	48.3%
IAP(class-aware) for RGB+D	53.7%
CAP for RGB+D (classification loss only)	54.2%
CAP for RGB+D (classification & smoothness loss)	54.6%
CAP for RGB+D (classification loss only) + Global	55.5%
CAP for RGB+D (classification & smoothness loss) + Global	56.2%

spatial relations of attentive objects to distinguish different scenes. However, the regions highlighted by IAP show that the spatial relations of objects might be encoded implicitly, which corresponds to how human beings distinguish these scenes through spatial relations.

4.3.4. Comparison of different fusion strategies in MAPNet

In the second learning stage, MAPNet learns to aggregate semantic cues across multiple modalities. Different strategies are compared and analyzed for multi-modal fusion in the second stage, as shown in Tabel 3. The *Baseline RGB+D* model utilizes two-stream CNN that the stream of which fuses at the score level. The weights in each stream are initialized with the fine-tuned Places CNN on RGB and Depth modalities respectively. It achieves 48.3% accuracy, which is already a high baseline compared with the state-of-the-art methods in Table 1. The *IAP for RGB+D* also utilizes two-stream CNN, each of which is constructed with the IAP block as shown in Fig. 3. The pooling strategy employed in IAP is class-aware attentive pooling. In the fusion of the two modalities, it adds $\mathbf{e^r}$, $\mathbf{e^d}$ and $\mathbf{o^r}$, $\mathbf{o^d}$ for each proposed region respectively. The combined logits are then aggregated as that in the class-aware attentive pooling strategy. It can achieve better performance compared with the Baseline model, validating the effectiveness of the proposed IAP block. However, the *IAP for RGB+D* still presumes the RGB and Depth modalities have the same contribution. The proposed CAP block further learns different contributions across modalities for different local semantic cues. The *CAP for RGB-D (classification loss only)* uses cross-modality attentive pooling block with only classification loss in Eq. (10). It increases the performance to 54.2%. The *CAP for RGB-D (classification & smoothness loss)*, i.e., the final structure in the second stage of learning, shows the effectiveness of the smoothness loss in Eq. (18). The final performance for MAPNet achieves 54.6% accuracy. Since the MAPNet classifies the indoor scenes based only on the local semantic cues, the global model can further provide complementary global spatial layout cues. The global model boosts the accuracy by about 2% to 56.2%.

4.3.5. MAPNet-V for feature-level RGB-D fusion

The proposed MAPNet in Fig. 4 fuses RGB and Depth modalities at score-level since the class-aware attentive pooling needs score-level information. We show in this section that the MAPNet can also be extended for feature-level RGB-D fusion by applying some modifications. We denoted the modified MAPNet as MAPNet-V. Following the same training strategy of MAPNet, the MAPNet-V is also trained in two stages. In the first training stage, we use the class-agnostic intra-modal attentive pooling as illustrated in Fig. 3 (b) for RGB and Depth streams respectively. In the second training stage, we design the variant CAP (CAPv) block based on the class-agnostic intra-modal attentive pooling. In accordance with the original CAP block, the attention weights z_i^r, z_i^d can be obtained by Eqs. (11) and (12). The formulation of the block then can be extended from Eq. (4) to:

$$CAPv(x) = fc_8\left(\left[\sum_{i=1}^{L} g_r(\mathbf{f}_i^r)\circ a_i^{r*}, \sum_{i=1}^{L} g_d(\mathbf{f}_i^d)\circ a_i^{d*}\right]\right) \tag{19}$$

where,

$$a_i^{m*} = \frac{\exp(a_i^m \cdot z_i^m)}{\sum_{i=1}^{L}(a_i^m \cdot z_i^m)}, m \in \{r, d\} \tag{20}$$

Fig. 6. Visualization of the Intra-modality Attentive Pooling block (best viewed in color). Top 3 most influential regions for scene classification are annotated with red, yellow and blue bounding boxes respectively. The corresponding scores are shown in the histograms below them. It shows our model can latently discover different types of discriminative semantic cues such as objects (the first row), the co-occurrence between the objects (the second row) and the spatial relations among the objects (the third row). More detailed analysis can be found in Section 4.3.3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The $g_r(\cdot)$, $\mathbf{f}_i^r$, e_i^r and $g_d(\cdot)$, $\mathbf{f}_i^d$, e_i^d have the same meanings as those in Eq. (4) for the RGB and Depth streams respectively. We only use the classification loss in the second stage training for MAPNet-V.

In Table 4, we compare the MAPNet-V with other strategies which also combine RGB and Depth at feature level. The *Baseline RGB-D* is similar with *Baseline RGB+D (Global)* except for that the RGB and Depth streams are combined by concatenating the outputs from the penultimate fully connected layers in the second training stage. It shows that the *Baseline RGB-D* gets similar accuracy with the *Baseline RGB+D (Global)* that combines RGB and Depth at score-level. The *IAP(class-agnostic) for RGB-D* uses class-agnostic attentive pooling in the first learning stage for RGB and Depth modality respectively. In the second learning stage, it di-

Table 4
Ablation study for MAPNet-V on SUN RGB-D dataset.

Methods	Accuracy(%)
Baseline RGB-D	48.3%
IAP(class-agnostic) for RGB-D	52.3%
MAPNet-V	53.1%

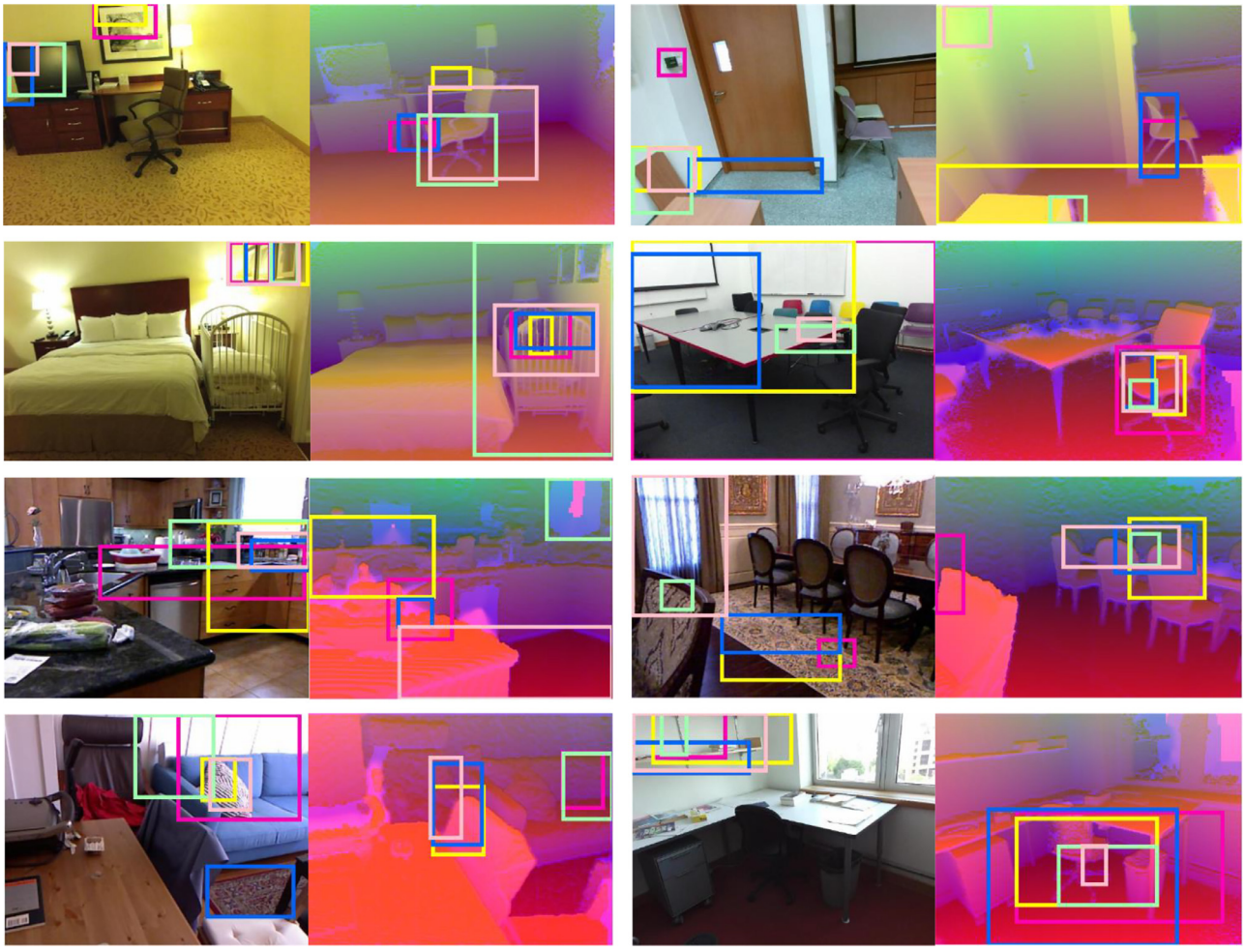


Fig. 7. Visualization of the obtained contributions across modalities in the Cross-modality Attentive Pooling block (best viewed in color). The 5 highest scored proposed regions in each modality are shown with red, yellow, blue, green and pink bounding boxes respectively. The proposed regions are assigned high scores in RGB modality where there is strong color contrast, while the proposed regions get high scores in Depth modality where the shape cues are prominent. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rectly concatenates the pooled features from penultimate fully connected layers of RGB and Depth streams. It achieves 52.3% accuracy, which further validates the effectiveness of the IAP block. By applying the MAPNet-V which uses the CAPv block instead of directly concatenating features, the classification accuracy can be further improved to 53.1%. Compared with the *Baseline RGB-D* and *IAP(class-agnostic) for RGB-D*, it validates the effectiveness of the proposed attention blocks (IAP and CAP) over direct feature concatenation in feature-level RGB-D fusion scenario. Notice that the MAPNet-V achieves about 1% less accuracy than the original MAPNet((*CAP for RGB+D (classification loss only)* in Table 3)). One possible reason is that we can only use the class-agnostic attentive pooling which lacks the consideration of the class information in attentive pooling as in MAPNet.

4.3.6. Visualization of contributions across modalities

The proposed model can help to explain and understand the multi-modal fusion process. In CAP block, the scores z_i^r and z_i^d in Eq. (12) represent the relative importance between the RGB and Depth modalities for the ith proposed region. We visualize regions that correspond to the 5 highest scores for RGB and Depth modalities respectively in Fig. 7. It shows that the model attends more on RGB regions where strong color or texture cues exist such as *picture on the wall* and *pillow on sofa*, while it attends more on Depth regions where shape cues in objects are prominent and discriminative, or the corresponding RGB regions are in dim environment such as *legs of chairs* and *corner of table*. The visualization proves that the CAP block can automatically learn the importance between the RGB and Depth for different semantic cues.

4.3.7. Visualization of occlusion scenario

In Fig. 8 we visualize the classification results from MAPNet in some occlusion scenarios. The occlusion here means some important discriminative cues indicating scene categories, e.g., the *bed* for the scene *bedroom*, are not presented or partially presented due to some occluisions or out of view. Assume the evaluated image is inferred as class c. The scores e_{ic}^r, e_{ic}^r in Eq. (13) and $a_{ic}^* \cdot b_{ic}^*$ in Eq. (15) represent importance of proposal i in RGB modality, Depth modality and combined RGB-D modalities when classifying the image as class c respectively. We show the 5 highest scores of e_{ic}^r, e_{ic}^r, $a_{ic}^* \cdot b_{ic}^*$ in Fig. 8 for each modality respectively. We can see that in the first row of Fig. 8 MAPNet can be robust in discovering occluded discriminative cues (*bed* and *row of chairs*). This is because that we use region proposals as our local cues. Even if the object is occluded, the part of the object can be extracted as local cues with region proposals. We can also find that as illustrated in the first image of Fig. 8, for the dim occluded scenario, the Depth modality can provide strong cues that help to overcome the heavy occlusion. When the most discriminative cues (*bed* and *toilet*) are occluded heavily, MAPNet also cannot assign them with

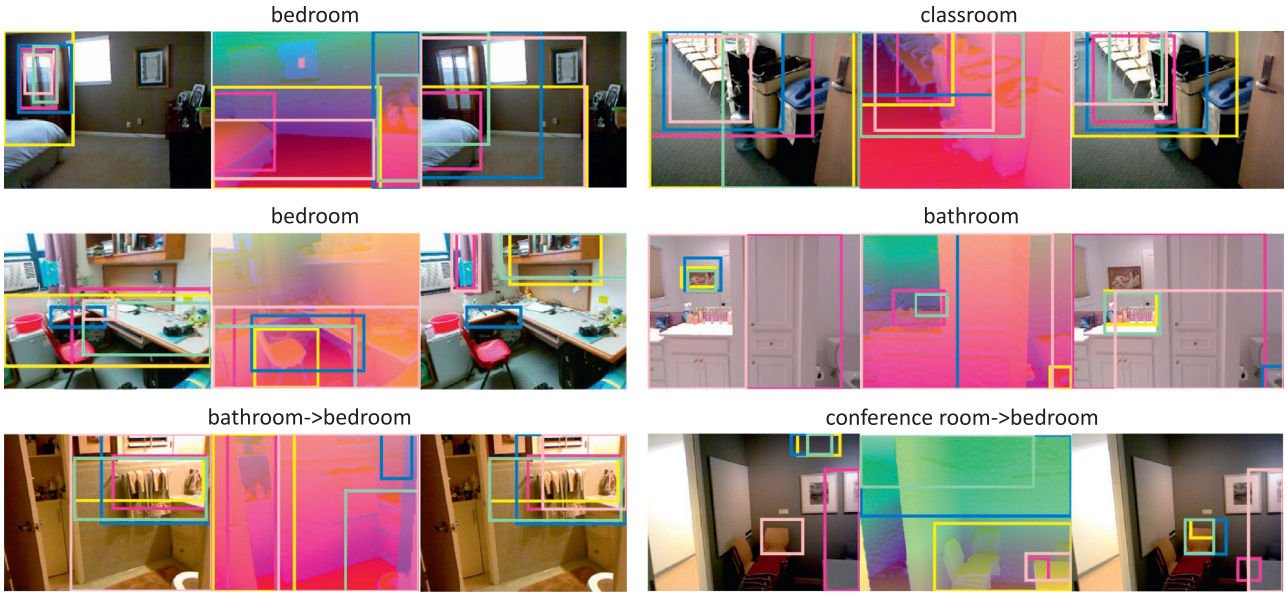


Fig. 8. Visualization of the occlusion scenarios. The occlusion here means some important discriminative cues indicating scene categories, e.g., the *bed* for the scene of *bedroom* are not presented or partially presented. For each image we show the 5 highest scored proposed regions from RGB modality, Depth modality and fused RGB-D modalities respectively. The 5 highest scored proposed regions are shown with red, yellow, blue, green and pink bounding boxes respectively. In the first two rows, we show some examples that MAPNet classifies correctly. In the last row, we show some examples that MAPNet classifies wrong.(*bathroom* and *conference room* are recognized as *bedroom*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

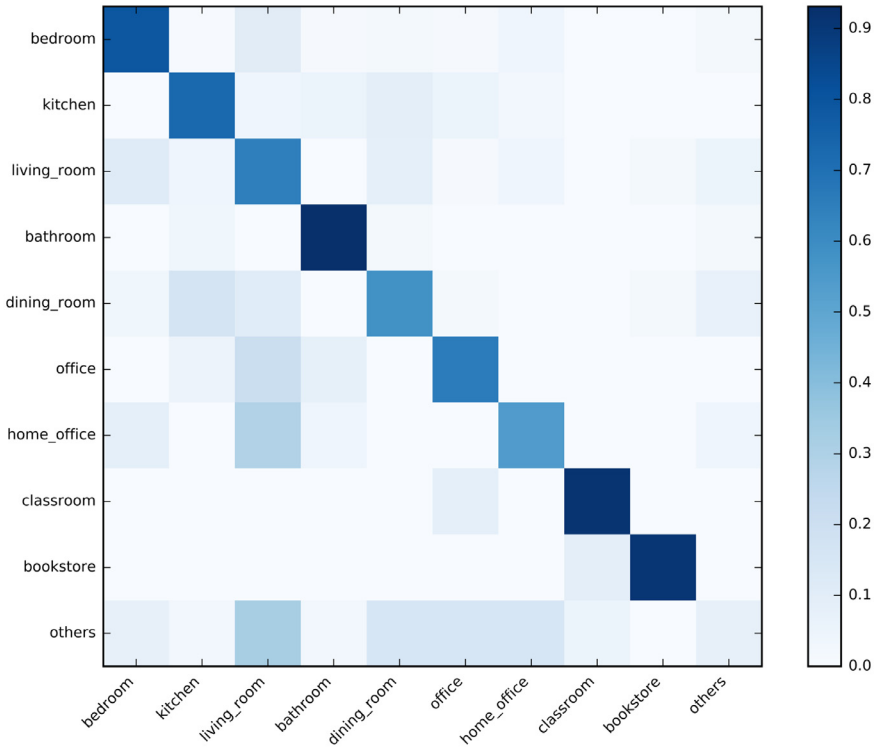


Fig. 9. The classification confusion matrix of MAPNet on the NYU Depth Dataset V2. The vertical axis shows the ground-truth classes. The classes in the horizontal axis are in the same order with those in the vertical axis. The horizontal axis shows the predicted classes.

high score as shown in the second row of Fig. 8. However, MAPNet can still correctly classify the scene by using other contextual cues, for example *curtain in the bedroom* and *sink in the bathroom*. When other cues are not discriminative, MAPNet will fail to classify the occluded scenario correctly. As shown in the third row of the Fig. 8, the *bathroom* and the *conference room* are classified as *bedroom*.

4.4. Results on NYU depth dataset V2

Table 5 shows the performance comparison on NYU Depth Dataset V2 with other state-of-the-art methods. The proposed MAPNet alone outperforms most of the existing methods except for the *Local + Global + OOR* in [50]. However, *Local + Global + OOR* in [50] uses both aggregated local information and global in-

Table 5
Performance comparison with the state-of-the-art methods on NYU Depth dataset V2.

Methods	Extra annotation	Accuracy(%)
(Gupta et al. 2013 [16])	No	45.4%
(Wang et al. 2016 [6])	No	63.9%
(Song et al. 2017 [48])	No	65.8%
(Song et al. 2017 [52])	No	66.7%
Local + OOR (Song et al. 2017 [50])	Yes	60.1%
MAPNet (ours)	**No**	**66.8%**
Local + Global + OOR (Song et al. 2017 [50])	Yes	66.9%
MAPNet + Global(ours)	**No**	**67.7%**

Table 6
Ablation study for MAPNet on NYU Depth dataset V2.

Methods	Accuracy(%)
Baseline RGB+D (Global)	60.6%
IAP(class-aware) for RGB+D	65.8%
CAP for RGB+D (classification loss only)	66.2%
CAP for RGB+D (classification & smoothness loss)	66.8%
CAP for RGB+D (classification loss only) + Global	66.5%
CAP for RGB+D (classification & smoothness loss) + Global	67.7%

formation for classification. When compared with *Local + OOR* in [50] that also only use local cues, our MAPNet achieves superior results with a large margin with *no* extra annotations. To fairly compare with the *Local + Global + OOR* model in [50], the global model is also combined with MAPNet. The performance further boosts to 67.7%, showing the advantage over [50]. The confusion matrix of our final test results on the NYU Depth Dataset V2 is shown in Fig. 9.

Table 6 shows the ablation study of MAPNet on the NYU Depth Dataset V2. Consistent conclusions can be obtained as those on the SUN RGB-D Dataset. One observation in our experiments is that the smoothness constraint plays a more important role on the NYU Depth Dataset V2. One possible reason we speculate is that the number of images in the NYU Depth Dataset V2 is smaller, the smoothness regularization is more important for small-scale datasets to learn the consistent cross-modal attention.

5. Conclusions

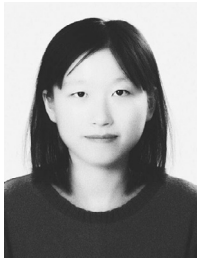
In this paper, we have proposed the Multi-modal Attentive Pooling Network (MAPNet) to aggregate discriminative semantic cues in RGB and Depth modalities for RGB-D indoor scene classification. Unlike other methods aggregating the semantic cues in scene images using Fisher vector (FV) or fully connected layer, the Intra-modality Attentive Pooling (IAP) block is proposed to select and aggregate discriminative semantic cues meanwhile maintain the spatial invariance in each modality. To exploit the complementary information fully in fusing the selected discriminative cues across the two modalities, we argue that the RGB and Depth should provide different contributions on different semantic cues for RGB-D indoor scene classification. The Cross-modality Attentive Pooling (CAP) block is thus proposed to learn the modality attention weights which are further employed to pool the discriminative local cues of each modality. We achieve the state-of-the-art results for RGB-D indoor scene classification on both SUN RGB-D Dataset and NYU Depth Dataset V2, which validates the effectiveness of the proposed framework.

One direction of future work is to explore the relations between various semantic cues. The relations can help represent scene images and discover new semantic cues. Another future work is to embed the region proposal process with the MAPNet learning. The highlighted semantic cues can be further utilized to generate more reliable region proposals.

References

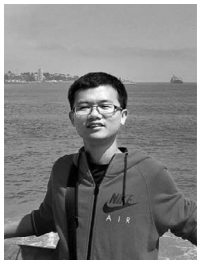
[1] L.-J. Li, H. Su, L. Fei-Fei, E.P. Xing, Object bank: a high-level image representation for scene classification & semantic feature sparsification, in: Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.
[2] F. Yang, W. Choi, Y. Lin, Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2129–2137.
[3] T.H.N. Le, C.N. Duong, L. Han, K. Luu, K.G. Quach, M. Savvides, Deep contextual recurrent residual networks for scene labeling, Pattern Recognit. 80 (2018) 32–41.
[4] C. Couprie, C. Farabet, L. Najman, Y. Lecun, Indoor semantic segmentation using depth information, in: International Conference on Learning Representations, 2013.
[5] S. Song, S.P. Lichtenberg, J. Xiao, Sun RGB-D: a RGB-D scene understanding benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.
[6] A. Wang, J. Cai, J. Lu, T.-J. Cham, Modality and component aware feature fusion for RGB-D scene classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5995–6004.
[7] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: theory and practice, Int. J. Comput. Vis. 105 (3) (2013) 222–245.
[8] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3304–3311.
[9] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 413–420.
[10] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, N. Vasconcelos, Scene classification with semantic fisher vectors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2974–2983.
[11] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, NetVLAD: CNN architecture for weakly supervised place recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.
[12] X. Cheng, J. Lu, J. Feng, B. Yuan, J. Zhou, Scene recognition with objectness, Pattern Recognit. 74 (2018) 474–487.
[13] P. Tang, X. Wang, B. Shi, X. Bai, W. Liu, Z. Tu, Deep FisherNet for object classification, (2016) arXiv:1608.00182.
[14] Y. Li, M. Dixit, N. Vasconcelos, Deep scene image classification with the MFAFVNet, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5746–5754.
[15] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: European Conference on Computer Vision, 2014, pp. 345–360.
[16] S. Gupta, P. Arbelaez, J. Malik, Perceptual organization and recognition of indoor scenes from RGB-D images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 564–571.
[17] X. Ren, L. Bo, D. Fox, RGB-(D) scene labeling: Features and algorithms, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2759–2766.
[18] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, 2012, pp. 746–760.
[19] A. Eitel, J.T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust RGB-D object recognition, in: IEEE International Conference on Intelligent Robots and Systems, 2015, pp. 681–687.
[20] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. (2018) inpress.
[21] R. Girshick, Fast R-CNN, in: IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[22] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[23] A. Oliva, A. Torralba, Building the gist of a scene: the role of global image features in recognition, Prog. Brain Res. 155 (2006) 23–36.
[24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893.
[25] L. Herranz, S. Jiang, X. Li, Scene recognition with CNNs: objects, scales and dataset bias, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 571–579.
[26] 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. (2017) inpress.
[27] R. Wu, B. Wang, W. Wang, Y. Yu, Harvesting discriminative meta objects with deep CNN features for scene classification, in: IEEE International Conference on Computer Vision, 2015, pp. 1287–1295.
[28] M. George, M. Dixit, G. Zogg, N. Vasconcelos, Semantic clustering for robust fine-grained scene recognition, in: European Conference on Computer Vision, 2016, pp. 783–798.

[29] 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.
[30] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 842–850.
[31] I. Ilievski, S. Yan, J. Feng, A focused dynamic attention model for visual question answering, (2016) arXiv:1604.01485.
[32] 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.
[33] M.D. Dixit, N. Vasconcelos, Object based scene representations using fisher scores of local subspace projections, in: Advances in Neural Information Processing Systems, 2016, pp. 2811–2819.
[34] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[35] A. Wang, J. Lu, J. Cai, T.-J. Cham, G. Wang, Large-margin multi-modal deep learning for RGB-Dobject recognition, IEEE Trans. Multimedia 17 (11) (2015) 1887–1898.
[36] H. Zhu, J.-B. Weibel, S. Lu, Discriminative multi-modal feature fusion for RGBD indoor scene recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2969–2976.
[37] 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.
[38] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: adaptive attention via a visual sentinel for image captioning, in: IEEE Conference on Computer Vision and Pattern Recognition, 6, 2017.
[39] F. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with LSTM, Neural Comput. 12 (10) (2000) 2451–2471.
[40] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, G. Hua, Neural aggregation network for video face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4362–4371.
[41] R. Girdhar, D. Ramanan, Attentional pooling for action recognition, in: Advances in Neural Information Processing Systems, 2017, pp. 33–44.
[42] K.E. Van de Sande, J.R. Uijlings, T. Gevers, A.W. Smeulders, Segmentation as selective search for object recognition, in: IEEE International Conference on Computer Vision, 2011, pp. 1879–1886.
[43] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[44] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2846–2854.
[45] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Attention-over-Attention Neural Networks for Reading Comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1, 2017, pp. 593–602.
[46] P.K. Nathan Silberman Derek Hoiem, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, 2012.
[47] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: ACM Multimidia, 2014, pp. 675–678.
[48] X. Song, L. Herranz, S. Jiang, Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs, in: AAAI Conference on Artificial Intelligence, 2017, pp. 4271–4277.
[49] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. Huang, Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3029–3037.
[50] X. Song, C. Chen, S. Jiang, RGB-D scene recognition with object-to-object relation, in: ACM Multimidia, 2017, pp. 600–608.
[51] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, Y. Liu, Understand scene categories by objects: a semantic regularized scene classifier using convolutional neural networks, in: IEEE International Conference on Robotics and Automation, 2016, pp. 2318–2325.
[52] X. Song, S. Jiang, L. Herranz, Combining models from multiple sources for RGB-Dscene recognition, in: International Joint Conferences on Artificial Intelligence, 2017, pp. 4523–4529.



Yabei Li received her B.S. degree in electrical engineer from Hefei University of Technology, Hefei, China, in 2014. She is currently working toward the Ph.D. degree in the Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation Chinese Academy of Sciences (CASIA), Beijing, China. Her research interests include computer vision, scene classification, semantic segmentation, and deep learning.

Zhang Zhang received the B.S. degree in computer science and technology from Hebei University of Technology, Tianjin, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China in 2009. Currently, he is an associate professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include activity recognition, video surveillance, and time series analysis. He has published 20s research papers on computer vision and pattern recognition, including IEEE TPAMI, CVPR, and ECCV etc.



Yanhua Cheng received the B.S. degree in automation from Huazhong University of Science and Technology in 2012, and the Ph.D. degree from the Institute of Automation Chinese Academy of Sciences (CASIA) in 2017. He currently works in Tencent Wechat AI, Beijing, China. He has published several leading international conferences such as CVPR, ICCV, IJCAI etc. His research interests include computer vision, object recognition, semantic segmentation, and deep learning.

Liang Wang received both the B.Eng. and M.Eng. degrees from Anhui University in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, and leading international conferences such as CVPR, ICCV, and ICDM. He is a senior member of the IEEE, and an IAPR Fellow.

Tieniu Tan received the B.Sc. degree in electronic engineering from Xian Jiaotong University, China, in 1984, and the M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He is currently a professor with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. His research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a fellow of CAS, TWAS, BAS, IAPR, and the U.K. Royal Academy of Engineering, and Past President of the IEEE Biometrics Council.