Cognitive Computation (2019) 11:825 840
https://doi.org/10.1007/s12559-018-9580-y

RGB-D Scene Classification via Multi-modal Feature Learning

Ziyun Cai[1,2] · Ling Shao[1,3]

Received: 13 February 2018 / Accepted: 12 July 2018 / Published online: 2 August 2018

Abstract

Most of the past deep learning methods which are proposed for RGB-D scene classification use global information and directly consider all pixels in the whole image for high-level tasks. Such methods cannot hold much information about local feature distributions, and simply concatenate RGB and depth features without exploring the correlation and complementarity between raw RGB and depth images. From the human vision perspective, we recognize the category of one unknown scene mainly relying on the object-level information in the scene which includes the appearance, texture, shape, and depth. The structural distribution of different objects is also taken into consideration. Based on this observation, constructing mid-level representations with discriminative object parts would generally be more attractive for scene analysis. In this paper, we propose a new Convolutional Neural Networks (CNNs)-based local multi-modal feature learning framework (LM-CNN) for RGB-D scene classification. This method can effectively capture much of the local structure from the RGB-D scene images and automatically learn a fusion strategy for the object-level recognition step instead of simply training a classifier on top of features extracted from both modalities. The experimental results on two popular datasets, i.e., NYU v1 depth dataset and SUN RGB-D dataset, show that our method with local multi-modal CNNs outperforms state-of-the-art methods.

Keywords Deep learning · Local fine-tuning · Convolutional neural networks · RGB-D scene classification

Introduction

Indoor scene classification has received increasing attention in both academia and industry over the past few years. It plays an important role for a wide range of practical applications,e.g., semantic recognition [1–4], content-based image indexing and retrieval [5–8], and remote sensing image classification [9, 10]. In the real-world situation, the intra-class variation of scenes is massive and the spatial layouts are vastly different. In addition, occlusion, low illumination, sophisticated background, and even different view angles can result in more challenges. Therefore, although much progress has been made, indoor scene recognition is still a challenging task.

Deep Convolutional Neural Networks (CNNs) obtain great success for high-level tasks, such as action recognition [11, 12], image classification [13–15], object detection [16], and emotion recognition [17, 18]. Li et al. [17] organized differential entropy features from different channels as two-dimensional maps to train the hierarchical convolutional neural network (HCNN) for the classification of the positive, neutral, and negative emotion states. Feng et al. [18] present a ranking-based multi-label convolutional neural network model (RM-CNN) which considers the order and relevance of labels to address emotion detection in microblogs. CNNs improve the state-of-the-art performance on many important datasets (e.g., the ImageNet dataset [19]), and even surpasses human performance on some datasets [20]. However, on the scene classification task, CNN features are still used rudimentarily. For example, Zhou et al. [21] simply collected a large-scale scene-centric dataset called “Places” to train Alexnet [22]. Then, they directly extracted holistic CNN features from the model. Additionally, they simply combined the training set of Places-CNN and the training set of ImageNet-CNN to train a Hybrid-CNN for Hybrid features. Although the

✉ Ling Shao
ling.shao@ieee.org

Ziyun Cai
caiziyun@163.com

[1] School of Automation, Northwestern Polytechnical University, Xi’an 710065, China

[2] College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China

[3] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates

published performance can be improved through scene-centric CNNs, it mainly depends on the abundant training data and very deep networks. Therefore, it is considered that scene classification with Deep Neural Networks is still in its infancy.

An indoor scene usually contains many different objects which can provide indirect clues for higher level tasks. Humans recognize the class of one unknown scene mainly relying on the object-level information. Given a new indoor image, for example, “Toilet,” we can quickly recognize this scene category when we find some representative objects such as “closestool,” “mirror,” and “hand sink” in it. In addition, we will enhance the recognition accuracy on this kind of scenes when we see other similar scenes. The hypothesis about how humans recognize a scene is illustrated in Fig. 1. Therefore, the scene categories represented by object information would reduce the variety among intra-class scene images and show larger discrimination among inter-class scene images. It makes constructing mid-level representations with discriminative object parts generally more useful than directly considering all pixels in the whole image for the scene classification task.

Recent developments in low-cost RGB-D sensors, for example, the Microsoft Kinect, have opened a new dimension which can generate depth information from the surrounding environment. Compared to RGB images which provide information about appearance and texture, depth images contain additional information about object shapes (see Fig. 1) and are robust to the variations in color, illumination, rotation angle, and scale [23, 24]. In the real-world scenario, the acquisition of various kinds of information including appearance, texture, and shape can improve humans’ accuracy rates in object recognition and scene classification. Though it has been proved that combining RGB and depth information in image/video classification can significantly improve the classification accuracy, it still needs a highly efficient method to fuse information from these two modalities to perform high-level reasoning.

There is no doubt that combining the advantages of Deep Convolutional Neural Networks, local scene features, and RGB-D image information can help researchers design more sophisticated scene classification algorithms. Most of the methods which are proposed for indoor scene image classification using local information and depth data have something in common: RGB-D CNN features are firstly extracted at different locations and scales of an image separately. And then, these learned features are simply concatenated as RGB-D features or encoded as a combined feature representation. At last, the feature representations are classified using a classifier such as SVM. Though results show that these kinds of feature representations are competitive and object-level information has the potential to improve scene classification, it still has two issues: (1) According to scene images, since spatial aggregation performed by pooling layers in CNNs is too simple and does not hold much information about local feature distributions, the fundamental architecture of CNNs is not supposed to be most suitable for classifying scene images. Local feature distributions in the aggregated features are neglected when critical inferences happen in the fully connected layers near the top of CNNs. (2) Simply concatenating RGB and depth features has not explored the correlation between raw RGB and depth images. Meanwhile, the complementary property between RGB and depth data is ignored, and the learning procedure cannot be adjusted mutually.

In order to address these two issues, in this paper, we propose an RGB-D local multi-modal feature learning method (LM-CNN) for scene classification. LM-CNN can effectively capture much of the local structure from the RGB-D scene images and automatically learn a fusion

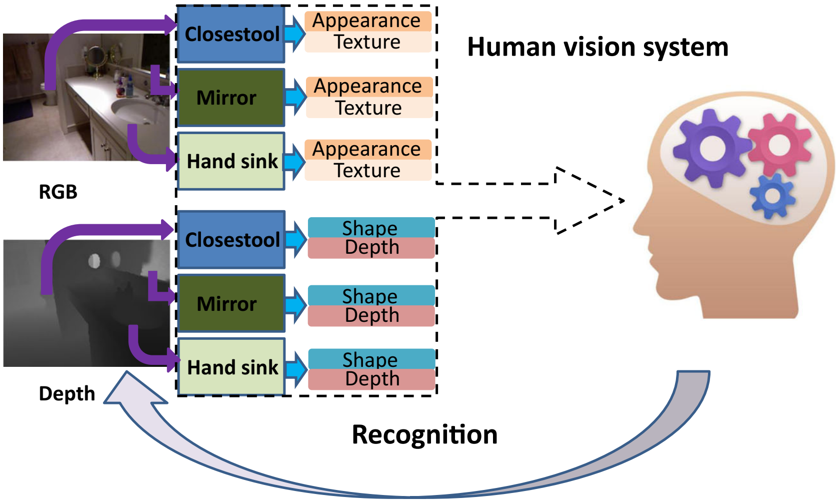

Fig. 1 The hypothesis about human vision system for scene classification in a natural environment. Humans firstly find some representative objects in these scenes. Then, humans learn these scene categories through the mid-level object features. Consequently, humans can show better recognition performance on the similar scenes

strategy for the object-level recognition step instead of simply training a classifier on top of features extracted from both modalities. We firstly utilize one region proposal extraction method over experimental RGB-D datasets and apply proposal screening on these generated region proposals to select the representative region proposals. Then, we group these selected proposals into clusters and encode selected depth proposals. Following the human way of scene recognition, we perform the CNNs to understand objects in the early stage. Our local fine-tuning multi-modal network automatically learns to combine these two processing streams on an additional layer in a late fusion approach. At last, the multi-level scene image representation is built from top of the probability distribution of the region proposals. LM-CNN is described in detail in the "Proposed RGB-D Local Multi-modal Classification Pipeline" section. The main contributions of this paper are a novel method for scene classification built on top of CNN features and a local fine-tuning multi-modal network using the representative proposals discovered from the target dataset.

The rest of this paper is organized in the following way. The "Related Work" section reviews related work on local deep feature learning and RGB-D fusion methods. In the "Proposed RGB-D Local Multi-modal Classification Pipeline" section, we give a detailed introduction of LM-CNN. Experimental setup and results on the verification details of our framework and the relevant experimental result analysis are comprehensively presented in the "Experimental Results" section. Finally, the conclusion and future work are given in the "Conclusion"section.

Related Work

As for the global deep feature learning methods, Place-CNN [21] is the most successful deep feature learning model in scene classification. Place-CNN is trained on a large-scale scene-centric dataset with 205 scene categories and 2.5 million images with category labels using the well-known architecture Alexnet [22]. Zrira et al. [25] proposed a novel approach which is based on restricted Boltzmann machines (RBMs) and the joint density model for recognition and representation of indoor environments. During the past few years, much work on classifying scene images using local deep learned information has been conducted. Feichtenhofer et al. [26] presented a novel ConvNet architecture based on transforming a purely spatial network to one that can encompass spacetime via hierarchical injection of temporal residuals. Gong et al. [27] presented a multi-scale orderless pooling scheme (MOP-CNN) which extracted CNN activations for local image patches at multiple scale levels. MOP-CNN performs orderless vector of locally aggregated descriptors (VLAD) pooling of these activations at each level separately and then concatenates the features as the final feature representation. Yoo et al. [28] presented a straightforward framework for better image representation by combining low-level local descriptors and mid-level deep neural activations of CNNs. The proposed multi-scale pyramid pooling method can perform better utilization of neural activations from pre-trained CNNs. Liao et al. [29] developed a scene classification model with regularization of semantic segmentation based on the Alexnet, called SS-CNN, where the features learned for scene classification in SS-CNN automatically contain object-level information. Although these methods have made great progress in scene recognition tasks, they do not provide a natural solution to fuse with the depth information.

There are also several methods proposed for RGB-D data fusion. For example, Gupta et al. [30] proposed one algorithm which generalizes the $gPb - ucm$ approach [31] through making effective use of depth information. Bo et al. [32] proposed Hierarchical Matching Pursuit (HMP) which builds feature hierarchies layer by layer with an increasing receptive field to obtain abstract representations of RGB-D data. Beyond this, Lai et al. [33] used 3D spin images and SIFT descriptors for depth features, and texton, color histogram, and standard deviation of each color channel for RGB features. Socher et al. [34] presented a model which is based on the combination of CNNs and Recursive Neural Networks (RNNs) [35], but this model extracted features of RGB-D images separately. Cai et al. [36] proposed a new RGB-D fusion method which first projects raw RGB-D data into a complex space and then jointly extracts features from the fused RGB-D images. Song et al. [37] proposed an approach which made a 3D volumetric scene from RGB-D images as inputs and a 3D object bounding box as output through Region Proposal Network (RPN) to learn objectness and a joint 2D + 3D object recognition network to extract geometric features in 3D and color features in 2D. However, the abovementioned methods have not explored the correlation and complementarity between raw RGB and depth images. Most of the methods just learn features from RGB and depth separately and then simply concatenate them as RGB-D features or encode these two kinds of features. The major disadvantage is that the correlation and complementary property between RGB and depth are ignored, and the learning procedure cannot be adjusted mutually.

In contrast, the proposed CNNs-based local multi-modal feature learning framework (LM-CNN) in this paper can effectively capture much of the local structure from the RGB-D scene images and automatically learn a fusion strategy for the object-level recognition step instead of simply training a classifier on top of features extracted from both modalities.

Proposed RGB-D Local Multi-modal Classification Pipeline

In this section, we introduce our LM-CNN in detail. The flow chart of our proposed method is illustrated in Fig. 2. It can be implemented as follows. As we mentioned in previous sections, our pipeline is built on the top of pre-trained CNNs. We firstly choose the state-of-the-art region proposal generating method which is most suitable for our pipeline to do region proposal extraction on our RGB-D datasets (in the "RGB-D Region Proposal Extraction" section). Then, we do proposal screening on these generated region proposals for the selection of representative region proposals (in the "Region Proposal Screening" section). After obtaining the discriminative region proposals, we group these proposals into clusters through an approach called Regularized Information Maximization (RIM) [38] (in the "Discriminative Region Proposals Clustering" section). Before we apply local fine-tuning of the multi-modal model on the above grouped region proposals, an RGB to depth encoding algorithm is performed over these proposals (in the "Depth Region Proposal Encoding" section). Our local multi-modal fine-tuning model is introduced in the "Local Fine-tuning of Multi-modal Architecture" section. At last, the multi-level scene image representation is built from top of the probability distribution of the region proposals (in the "Multi-level Representation from Region Proposals" section).

RGB-D Region Proposal Extraction

Generating region proposals is to obtain a small set of relative bounding boxes which try to cover all of the objects in the image. It has wide applications such as efficient object detection [39–41], weakly supervised learning [42, 43], and unsupervised representation learning [44]. Currently, many approaches have been proposed for generating region proposals including BING [45], MCG [46], and Edge Boxes [47]. BING trains a simple linear classifier over edge features, and then this classifier is applied in a sliding window manner. After this, a very fast agnostic detector can be obtained. MCG combines the advantages of two leading methods for generating proposals (gPbUCM [48] and CPMC [49]). Moreover, MCG proposes an improved hierarchical segmentation, a new method to generate proposals and a new ranking procedure. Edge boxes is similar to BING, but it uses object boundaries as features for the scoring. In our pipeline, the quality of the extracted region proposals plays an important role. The effective region proposals should satisfy three criteria: *high recall rate*, *few number of proposals*, and *tolerable evaluation speed*. One paper which evaluates ten publicly available detection proposal methods has proposed that only Edge Boxes maintains good performance on the above three criteria [50]. Therefore, we choose Edge Boxes to generate high-quality RGB region proposals in our

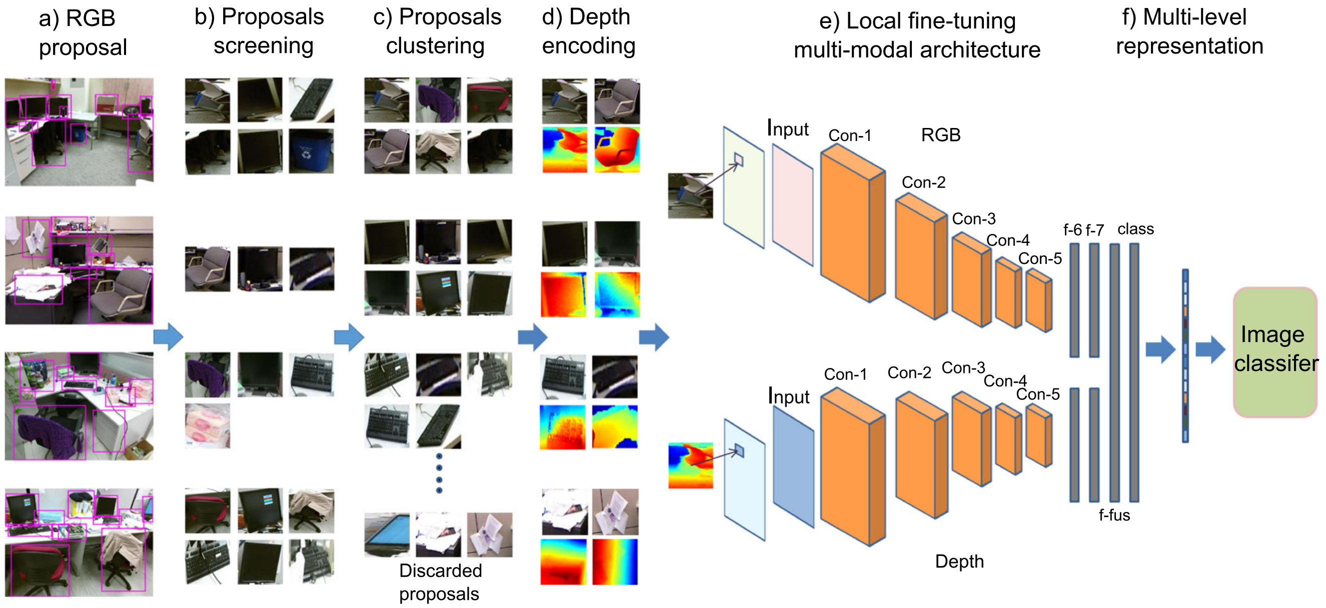

Fig. 2 The flow chart of the proposed LM-CNN. **a** High-quality RGB region proposals are generated through Edge Boxes. We choose the top-ranked proposals. **b** Unrepresentative region proposals are removed by one-class SVMs. **c** These proposals are grouped into clusters through RIM. The discarded region proposals are considered in an extra cluster. **d** Depth region proposals are then encoded into JET-style depth proposals. **e** Multi-modal local fine-tuning architecture is performed on Caffe pre-trained network, which can decide the category of a test region. **f** We train an image classifier for multi-level representation from region proposals (see text for more details)

pipeline. Meanwhile, region proposals from hierarchical image segmentation [31] are also used. The corresponding depth region proposals can be acquired through cropping the depth scene images into depth region proposals along the location of the RGB region proposals on the RGB scene images. Some example images about RGB-D region proposals can be found in Fig. 3. We choose ImageNet-CNN features [22] which are learned from the pre-trained Caffe model [51] on image classification dataset (i.e. ImageNet) for all the image region proposals. The feature dimension after CNNs in the FC7 layer is 4096.

Region Proposal Screening

In practice, we can consider that there exist some representative region proposals in each scene category. Some other unrepresentative region proposals may also appear in this scene category, but only few images contain these region proposals. We consider these unrepresentative region proposals as outliers. Motivated by [52], we use one-class SVM to remove these unrepresentative region proposals and then estimate the discriminative power among scene categories for each region proposal. A one-class SVM can separate the data from the origin to remove unrepresentative region proposals. Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ $(\mathbf{x}_i \in R^d)$ be the region proposals from one class, and kernel mapping φ: $X \rightarrow H$ maps original region features into another feature space. The outcome of the decision function only relies on the inner product of the vectors in the feature space H. Therefore, it is not necessary to perform the non-linear mapping explicitly by defining a simple kernel function that fulfills Mercers conditions $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$

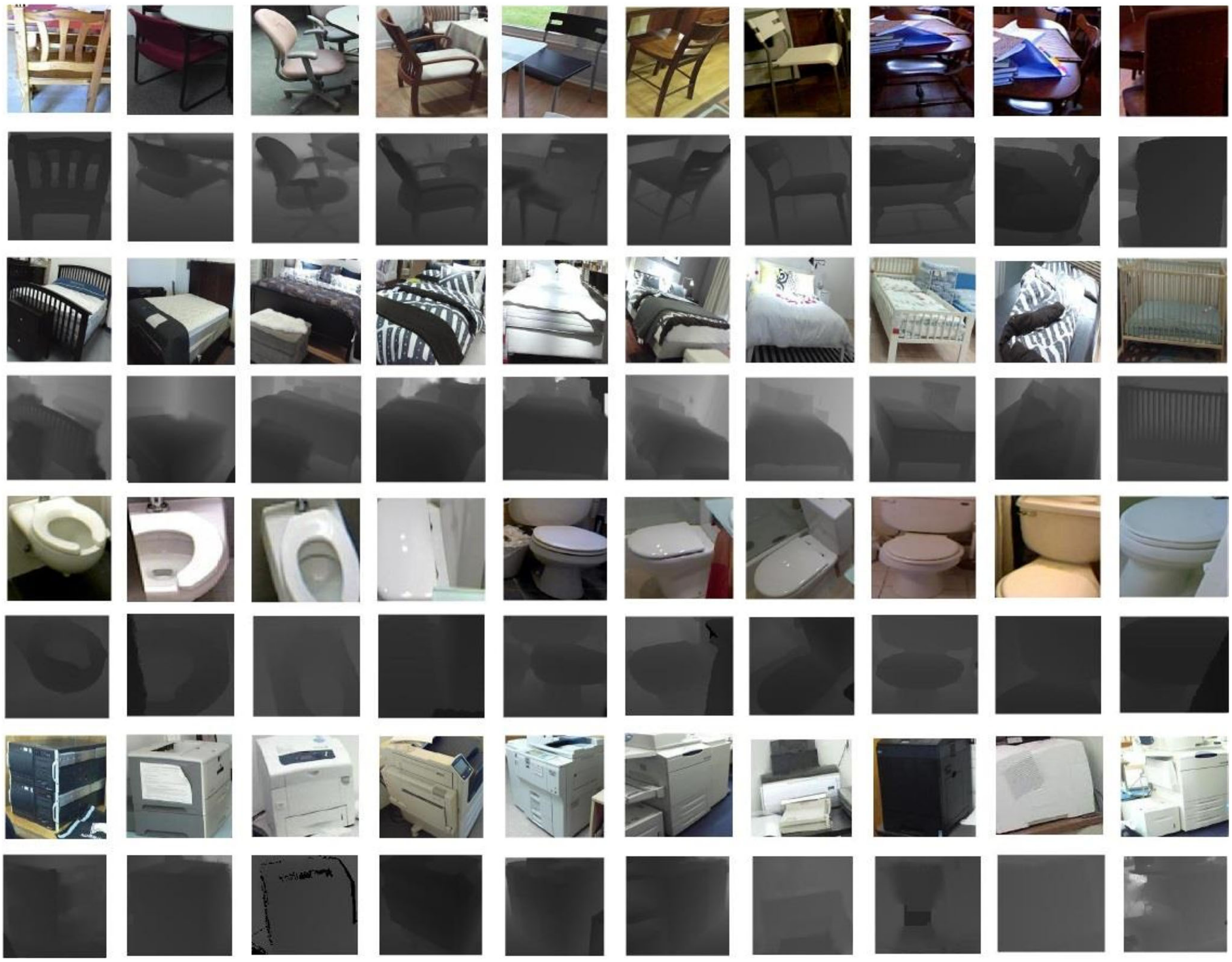

Fig. 3 Example images about RGB-D region proposals from the SUN RGB-D dataset and the NYU Depth v1 dataset. Images on the same row belong to the same cluster. First, third, fifth, and seventh rows are the RGB region proposals. Second, fourth, sixth, and eighth rows are the corresponding depth region proposals

[53]. The Gaussian Radial Base Function is the largely used kernel. We solve the following optimization:

$$\begin{aligned} &\min_{u,\lambda,\eta} \tfrac{1}{2}\|u\|^2 + \tfrac{1}{vn}\sum_{i=1}^{n}\lambda_i - \eta, \\ &s.t.\ (u \cdot \varphi(\mathbf{x}_i)) \geq \eta - \lambda_i, \lambda_i \geq 0, i \in (1, 2, \cdots, n), \end{aligned} \tag{1}$$

where $v \in (0, 1]$ controls the ratio of outliers. $u \in H$ and $\eta \in \mathbb{R}$ are a weight vector and an offset which can parameterize a hyperplane in the feature space. $\lambda \in \mathbb{R}^n$, λ_i are called the nonzero slack variables. The decision function can be obtained:

$$f(x) = \text{sign}(u \cdot \varphi(\mathbf{x}_i) - \eta), \tag{2}$$

which returns the positive sign when given the representative region proposals and returns the negative sign when given the outliers. Aiming to achieve better performance, we use three cascaded classifiers. We define that each of the classifiers labels 15% of the input region proposals as unrepresentative proposals and prune them.

According to our hypothesis, since each image I_i can be expressed as several region proposals, we define each region proposal from I_i as r_j^i, where i is the number of input images and j is the number of representative region proposals. Meanwhile, we use y_i to express the labels of the input scene images. In a natural world, a discriminative region proposal should usually appear in one scene category but unusually appear in other scene categories. It means that if we find K-nearest neighbors n_j^i of r_j^i from all image patches generated from all training images except I_i. Following this, we give the discriminative power for each region proposal among scene categories. The discriminative power is from 0 to 1. We can also consider it as the weight W_j^i of each region proposal. W_j^i can be expressed through the following class density estimator:

$$W_j^i = P(y_i \mid r_j^i) = \frac{P\left(r_j^i, y_i\right)}{P\left(r_j^i\right)} \approx \frac{K_y}{K}, \tag{3}$$

where K_y is the number of region proposals among the K-nearest neighbors that share the same scene label with r_j. We make the assumption that the K-nearest neighbors of r_j^i are almost the same as r_j^i. Then, K_y is used to estimate the joint probability between a patch r_j^i and its label y_i. We set K as 100 in all the experiments. Table 1 gives the distribution of region proposal weights after the screening.

Table 1 Region proposal weight distribution

Weight	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1]
Percentage	34.5	26.1	16.2	16.0	7.2

Discriminative Region Proposals Clustering

After obtaining the discriminative region proposals, we group these proposals into clusters. It can help us discover the relationship between scene category labels and region proposal labels. Krause et al. [38] proposed a framework called Regularized Information Maximization (RIM) which can simultaneously cluster the data and train a discriminative classifier. This algorithm contains the optimization of an intuitive information theoretic object function which strikes a balance among class separation, class balance, and cluster complexity. As we can see from Fig. 3, the region proposals after clustering have the similar appearance and semantic meaning. Till now, we can obtain the region proposal clusters of RGB images. To acquire the corresponding depth region proposals, we crop the depth scene images into depth region proposals according to the locations of the RGB region proposals on the RGB scene images. Figure 3 shows some RGB-D region proposal pairs from the SUN RGB-D dataset and the NYU Depth v1 dataset.

Depth Region Proposal Encoding

After we have obtained the depth region proposals, unlike RGB images, these depth images cannot be directly used as inputs for the CNNs. To solve this problem, one encoding method called HHA [54] has been proposed. HHA encodes the depth image into 0 to 255 range with three channels at each pixel and emphasizes complementary discontinuities in the image (depth, surface normal, and height). In this step, we employ the MATLAB jet colormap which is effective and computationally cheap. Jet firstly normalizes the depth values between 0 and 255; then, the normalized image is transformed from a single-channel to a three-channel image by a jet color map. For the pixels on the depth image, the distance to color values is mapped from red (near) over green to blue (far). Some examples of the encoded images from the dataset can be found in Fig. 4. In our experiments (see the "Without Region Proposal Screening" section), it proves that Jet encoding outperforms HHA encoding for our method.

Local Fine-tuning of Multi-modal Architecture

After we obtain the RGB region proposals and their corresponding depth region proposals, we choose to fine-tune the pre-trained CNNs on these RGB-D region proposals. During the training phase, all the original RGB/depth images are randomly cropped into 227 × 227 pixels. These cropped images are the inputs of the Caffe model. The used CNNs contain five fully convolutional layers, three fully connected layers, and 60 million

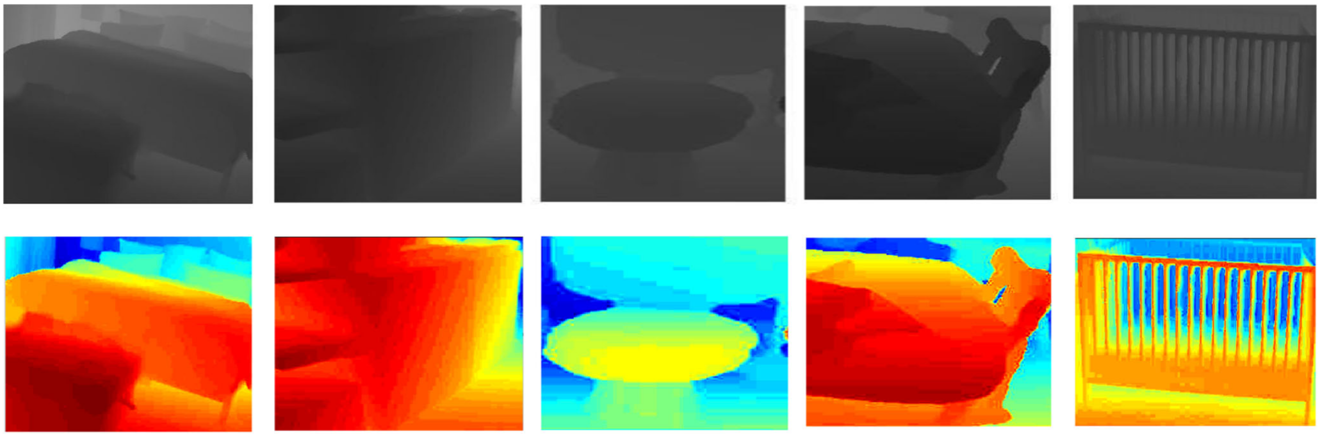

Fig. 4 Some examples of the Jet-encoded images from the dataset. Images on the first row are the depth region proposals. Images on the second row are the Jet-encoded depth region proposals

parameters. The proposed architecture is shown in Fig. 2. It contains two streams—which process the RGB data and depth data independently, and then are combined in a late fusion approach. Each stream contains one Caffe model implementation of the CNNs which are pre-trained for object classification on the ImageNet dataset. Then, we fine-tune the CNNs on our region proposal categories. The details about the CNNs architecture can be found in [22].

We have defined each RGB region proposal as Rr^i_j and each corresponding depth region proposal as Dr^i_j. The region proposal label can be expressed as Y_i. In each individual stream, we choose to use the pre-trained Alexnet on the large-scale image classification dataset (ImageNet). All parameters including the weights and biases P^R from RGB region proposals and P^D from depth region proposals are initialized from the Caffe-trained network on the ImageNet dataset. We then train the two streams separately through putting a randomly initialized softmax classification layer on the top of the RGB and depth layers. For the RGB or depth region proposal image stream network, we minimize the negative log likelihood $\mathcal{L}$ of the training data. They can be expressed as:

$$\min_{\mathbf{W}^R, P^R} \sum_{i=1}^{N} \mathcal{L}(\text{softmax}(\mathbf{W}^R g^R(\mathbf{r}^i; P^R)), Y_i), \tag{4}$$

$$\min_{\mathbf{W}^D, P^D} \sum_{i=1}^{N} \mathcal{L}(\text{softmax}(\mathbf{W}^D g^D(\mathbf{d}^i; P^D)), Y_i), \tag{5}$$

where $\mathbf{W}^R$ and $\mathbf{W}^D$ are the weights on the softmax layer, which map from $g(\cdot)$ to $\mathbb{R}^M$. $g^R(\mathbf{r}^i; P^R)$ and $g^D(\mathbf{d}^i; P^D)$ are the representations of RGB's last fully connected layer and depth's last fully connected layer separately. The loss is expressed as:

$$\mathcal{L}(s, y) = -\sum_k y_k \log s_k. \tag{6}$$

The softmax function is defined as:

$$\text{softmax}(\mathbf{z}) = \frac{\exp(\mathbf{z})}{\|\mathbf{z}\|_1}. \tag{7}$$

After these two stream networks are trained, we firstly discard the softmax weights and then concatenate them. We fine-tune the responses $g^R(\mathbf{r}^i; P^R)$ and $g^D(\mathbf{d}^i; P^D)$ as we mentioned in Eqs. 4 and 5. At last, a fusion stream $f([g^R(\mathbf{r}^i; P^R), g^D(\mathbf{d}^i; P^D)]; P^F)$ is used for the last-layer responses. Same with RGB and depth individual networks, RGB-D fusion network also ends at the softmax layer. All weights of the network are learned with a fixed momentum (set to 0.9). The dropout ratio of the fully connected layer is set to 0.5. To avoid over-fitting, we initialize the learning rate to 0.01 and make it reduce to 0.001 after $20k$ iterations. It finally stops at $50k$ iterations. For the fusion network, we train this through jointly optimizing the parameters to minimize the negative log likelihood:

$$\min_{\mathbf{W}^f, P^R, P^D, P^f} \sum_{i=1}^{N} \mathcal{L}(\text{softmax}(\mathbf{W}^f f([\mathbf{g}^R, \mathbf{g}^D]; P^f), Y_i), \tag{8}$$

where $\mathbf{g}^R$ and $\mathbf{g}^D$ are the representations of RGB-D's last fully connected layers. $\mathbf{W}^f$ is the weight of the softmax layer, and P^f is the parameter from RGB-D fusion region proposals.

Note that we obtain N (around several hundreds) region proposal classes during the clustering step, but we change the output layer of the ImageNet 1000-way classification into $(N + 1)$-way classification. Meanwhile, other layers remain unchanged. The extra way means the region proposals which are thrown away in the screening step. It can make the region proposal classifier robust to the noisy labels.

Multi-level Representation from Region Proposals

After N RGB-D region proposal clusters have been learned in the discriminative clustering step, the recognition of a given RGB-D image pair can be summarized as follows. We firstly perform the EdgeBoxes method on the given image pair to generate the region proposals. Each RGB-D region proposal can be classified into one pair of the clusters through the region proposal classifier. Since it has been proved that Spatial Pyramid Matching (SPM) [55] and modified Vector of Locally Aggregated Descriptors (VLAD) [56, 57] are successful, we decide to choose both of these two methods. According to SPM, we use three levels of SPM and choose the center of all region proposal clusters falling into a SPM region as the splitting center in our experiments. Then, we can obtain a hierarchical histogram of region proposal labels about the image. It can be used for the classification of the coming image. The difference between modified VLAD and VLAD is that modified VLAD does not use K-means clustering, and modified VLAD chooses discriminative region proposal clusters as the clusters for VLAD. Then we can obtain n clusters for RGB-D images separately. Each region proposal of the test image is assigned to its nearest cluster center resulting in a 4096-d vector per cluster. Then, each 4096-d is reduced into $4096/2n$-d vector through PCA. At last, we can obtain a 4096-d VLAD descriptor through concatenating these $4096/2n$-d vectors.

Another kind of feature we consider is hybrid Places feature which is first mentioned in [21]. It is learned from over 2.5 million labelled pictures of scenes and combines local and global information of these scene images. In our experiments, RGB hybrid Places image features and depth hybrid Places image features are obtained separately.

At last, we concatenate normalized VLAD/SPM features and the RGB-D hybrid Places features to train a network with two hidden layers. The image representation is expressed as the concatenation of all of the feature vectors. The weights between these two kinds of features are learned via cross validation on part of the training data. Our experiments were performed on the pre-trained Caffe model which is treated as the basis for our RGB-D fusion network. The model contains five convolutional layers (max-pooling is followed after the first, second, and fifth convolution layer), two fully connected layers and a softmax classification layer. Rectified linear function (ReLU) is the activation function of the neurons during the fully-connected hidden layers. Both of the stream networks are initialized with weights and biases of the first eight layers from the pre-trained network. The softmax layer is discarded. We then continue our stepwise training.

Experimental Results

We evaluate our LM-CNN along with current state of the art on the NYU Depth v1 dataset [58] and the SUN RGB-D dataset [59]. Both of these two datasets are derived from the publicly available RGB-D sensor-based scene database. In our experiments, we not only compare our method with hand-crafted methods such as GIST [60] but also compare deep feature learning models such as Alexnet, VGG [61], and some other representative models. The Places-CNN [21] scene features which are learned through Alexnet or VGG use the Places dataset for model pre-training. The Places2-CNN scene features use the Places2 dataset which has much more scene images than the Places dataset for model pre-training. Details of the datasets and experimental setup are provided below.

Datasets

NYU Depth v1

To evaluate our LM-CNN, we firstly conduct experiments on the NYU Depth v1 dataset created by Vision Learning Graphics (VLG) lab in New York University. It is collected from wide domains (the background is changing from one to another), facilitating multiple applications. For the indoor scene classification task, the NYU Depth v1 dataset consists of 7 different kinds of scene classes totally containing 2347 labelled frames. Since the standard classification protocol removes scene "cafe" from the dataset, we use the remaining 6 different scenes. Example images in the NYU Depth v1 dataset are shown in Fig. 5. Since there are so many objects in one scene and the correlation between images in one scene is low, it makes NYU Depth v1 a very challenging dataset. The baseline when only using RGB images is 55% [58].

SUN RGB-D

We also test LM-CNN on SUN RGB-D dataset. This dataset is captured by four different sensors (Intel RealSense, Asus Xtion, Kinect v1, and Kinect v2) and contains 10,335 RGB-D images. These images are organized into 19 scene categories such as bathroom, computer room and lecture theatre with more than 80 images in each category. Figure 6 shows some example images from this dataset. Every shown scene is from one of the 19 scene categories. Following the setup in [59], we split the dataset to ensure each sensor has around half for training and half for testing. These images which are captured from the same building or house with similar furniture styles are either all go into the training

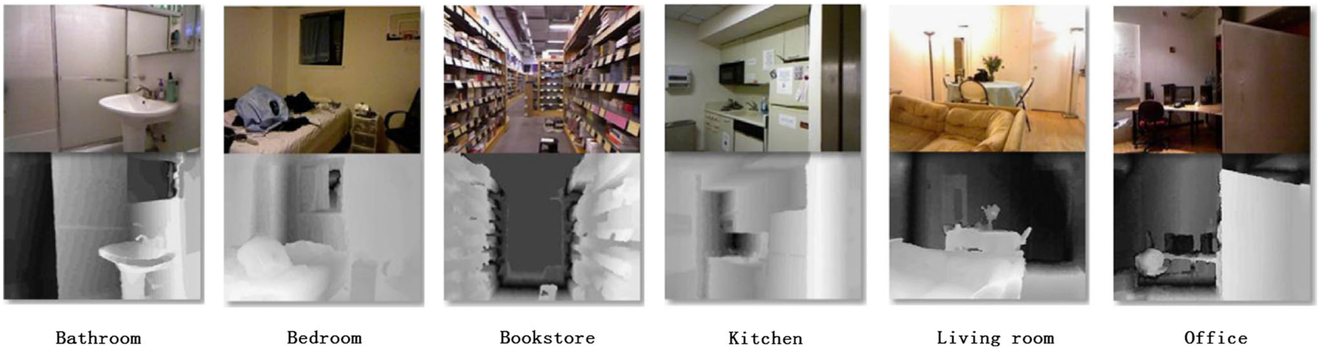

Fig. 5 Some example images in the NYU Depth v1 dataset. It includes 6 object classes (bathroom, bedroom, bookstore, kitchen, living room, and office). We can find 6 paired samples shown in this figure. In each pair, the RGB image is shown on the top and the corresponding depth image is on the bottom

set or the testing set and are not spread across both sets. To the best of our knowledge, the SUN RGB-D dataset is the largest and most challenging RGB-D scene dataset currently. The baseline when using GIST and RBF kernel SVM only gives a classification accuracy of 23%.

Experiment Setup

The first experiment about the proposed LM-CNN is on the NYU Depth v1 dataset. During the experiment, we obtain 131 top ranked RGB region proposals through Edge Boxes from each image. Meanwhile, we also generate 32(96) region proposals in the top (bottom) level by hierarchical image segmentation from each image. After region proposal screening, 15% region proposals are removed as unrepresentative region proposals. Then we do discriminative clustering on these screened region proposals resulting in 70(30) proposal classes. To make the region proposal classifier robust to the noisy labels, the discarded region proposals in the screening step are considered as the 71th proposal class. The corresponding depth region proposals can be acquired through cropping the depth scene images into depth region proposals along the location of these RGB region proposals on the RGB scene images. The whole local fine-tuning fusion procedure is performed on the famous public Caffe toolbox. We use the pre-trained model of the large-scale image classification dataset (ImageNet). The layers from both of the stream networks are initialized from the pre-trained eight layers. The softmax layer is discarded. Then we concatenate the softmax layer of the two individual networks. The output layer on the ImageNet 1000-way classification is changed into a new output layer of 71-way (31-way) classification. In our multi-level representation step, through modified VLAD and Hybrid CNN, each image can be expressed as concatenated normalized VLAD features and RGB-D Hybrid Places

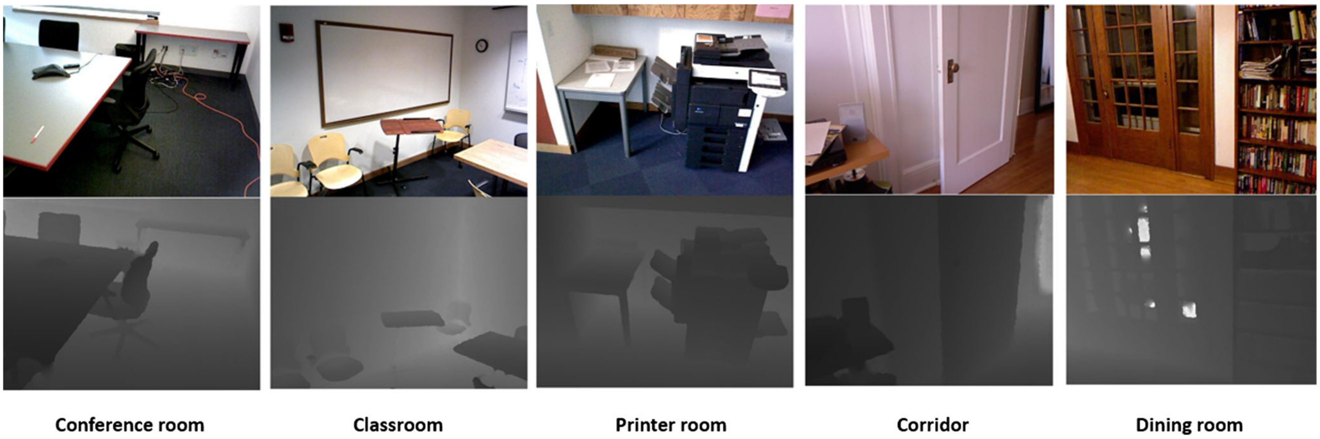

Fig. 6 Some example images from the SUN RGB-D dataset. It includes 19 scene categories (conference room, classroom, bookstore, printer room, corridor, dining room, and so on). Due to space limitation, we only show 5 paired scene samples in this figure. In each pair, the RGB image is shown on the top and the corresponding depth image is on the bottom

Table 2 The comparison results of LM-CNN and other published methods on the NYU Depth v1 dataset

Method	RGB (%)	Depth (%)	RGB-D (%)
CNN-RNN [34]	73.5	65.2	75.7
RICA [62]	74.5	64.7	74.5
LLC [63]	66.1	61.5	66.3
SPM [55]	52.8	53.2	63.4
HMP-S [32]	72.6	63.9	72.8
CNNs+LLC [64]	73.1	61.8	75.4
Places2-CNNs [21]	74.5	66.9	76.8
Hybrid-CNNs [21]	74.1	**68.2**	77.4
LM-CNN	**74.7**	67.8	**79.3**

Bold numbers indicate the best results

features of the whole image. At last, the image classification step is done over a neural network with two fully connected layers (200 nodes for each).

We compare LM-CNN with state-of-the-art methods including (1) R. Socher et al. [34], using the combination of CNNs and RNNs; (2) Le et al. [62], using robust soft reconstruction cost for ICA; (3) Wang et al. [63], using locality constraint to select similar basis of local image descriptors; (4) Bo et al. [32], using transfer learning-based method; (5) Jin et al. [64], using self-trained CNNs and LLC; (6) Zhou et al. [21], using Places2 dataset for pre-training, and fine-tuning on the training set of NYU depth v1 dataset; and (7) Zhou et al. [21], using both Places2 dataset and ImageNet dataset for pre-training, and fine-tuning on the training set of NYU depth v1 dataset. The comparison results are shown in Table 2. It can be seen that LM-CNN achieves the best performance and outperforms state-of-the-art methods in the RGB-D scene classification task. Figure 7 shows the confusion matrix about our method across six classes on the NYU Depth v1 dataset, whose diagonal elements represent the recognition accuracy for each category.

In Table 2, the performance of our LM-CNN achieves 79.3% on RGB-D image pairs, which outperforms published methods, i.e., RICA, LLC, SPM, CNNs + LLC, Places2-CNNs, and Hybrid-CNNs, by 4.8%, 13.0%, 15.9%, 3.9%, 2.5%, and 1.9% respectively. The main reason is that deep learned features are indeed able to achieve higher performance than hand-crafted features when the number of training samples is large enough. However, the individual RGB or depth performance through our method hardly shows any advantages compared to individual performance by Places2-CNNs and Hybrid-CNNs. It proves that our local fine-tuning multi-modal fusion architecture plays an important role in LM-CNN.

The second experiment is on the SUN RGB-D dataset. During the experiment, similar with the NYU Depth v1 dataset, we firstly obtain 103 top-ranked RGB region proposals through Edge Boxes from each image. 32(96) region proposals are generated in the top (bottom) level by hierarchical image segmentation from each image. Following region proposals screening, 15% region proposals are removed as unrepresentative region proposals. Then we do discriminative clustering on these screened region proposals resulting in 190(60) proposal classes. The discarded region proposals in the screening step are considered as the 191th proposal class. The corresponding depth region proposals are acquired through cropping the

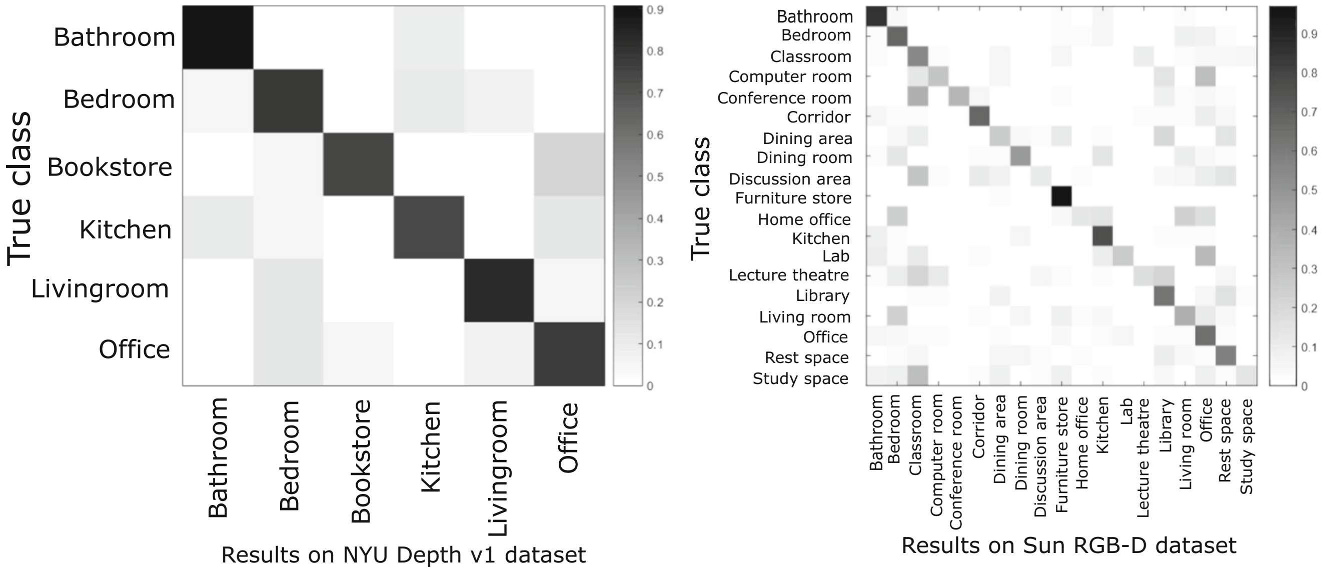

Fig. 7 Confusion matrixes about our method LM-CNN on NYU Depth v1 dataset (left) and SUN RGB-D dataset (right). The labels on the vertical axis express the true classes and the labels on the horizontal axis denote the predicted classes

depth scene images into depth region proposals along the location of these RGB region proposals on the RGB scene images. We use the pre-trained Caffe model of the large-scale image classification dataset (ImageNet). The layers from both of the stream networks are initialized from the pre-trained eight layers. The softmax layer is discarded. Then, we concatenate the softmax layer of the two individual networks. The output layer on the ImageNet 1000-way classification is changed into a new output layer of 191-way (61-way) classification. In our multi-level representation step, each image can be expressed as concatenated normalized VLAD features and RGB-D Hybrid Places features of the whole image. At last, a neural network with two fully connected layers (200 nodes for each) is trained for image-level classification.

LM-CNN is compared with state-of-the-art methods including: (1) R. Socher et al. [34], using the combination of CNNs and RNNs; (2) Oliva et al. [60], using GIST features and RBF kernel SVM; (3) Zhou et al. [21], using Places-CNN features and linear SVM; (4) Zhou et al. [21], using Places-CNN features and RBF kernel SVM; (5) Liao et al. [29], using the original Alexnet trained with only the SUN RGB-D dataset; (6) Liao et al. [29], using Places-CNNs and object-level information; and (7) Wang et al. [65], using high-dimensional Fisher vector (FV) features from RGB, HHA, and surface normal modalities. Moreover, we compare our method with two classical models: Places2-CNNs + softmax + Alexnet and Places2-CNNs + softmax + VGG [61]. However, according to the RGB-D image pairs, we simply concatenate RGB features and depth features. The comparison results are shown in Table 3. It can be seen that our method achieves the best performance and outperforms state-of-the-art methods in the RGB-D scene classification task. Figure 7 shows the confusion matrix about our method across 19 classes on the SUN RGB-D dataset, whose diagonal elements represent the recognition accuracy for each category.

Table 3 shows that the performance of LM-CNN achieves 48.7% on RGB-D image pairs. It outperforms published methods, i.e., GIST + RBF kernel SVM, Places-CNN + Linear SVM, Places-CNN + RBF kernel SVM, Alexnet, SS-CNN-R6, and Combine FV and CNN, by 25.7%, 11.5%, 9.7%, 18.0%, 7.4%, 6.4%, 7.4%, and 0.6%, respectively. Besides, our method also outperforms two popular global fine-tuned models: Alexnet and VGG, which are both pre-trained on the Place2 scene dataset. Similar with the NYU v1 depth dataset, the individual RGB or depth performance through our method hardly shows any advantages compared to individual performance by Places2-CNNs + softmax + Alexnet and Places2-CNNs + softmax + VGG. It proves that our architecture is efficient as a whole.

Global and Local Fine-tuning Discussions

To show the advantages of our local fine-tuning method, we perform some additional global fine-tuning experiments on the NYU v1 depth dataset and the SUN RGB-D dataset for image classification. These experiments use a pre-trained deep network and focus on global fine-tuning, which take the entire images as input and rely on the network itself to learn all the information within a coming new dataset. We set up these experiments with Alexnet, different pre-trained models and different classifiers. Studies on CNNs, such as GoogLeNet [66], indicate that using deeper models would improve classification performance more substantially than shallow ones. Since our method is only based on Alexnet, we ignore other deeper networks, i.e., VGG net [61] and GoogLeNet [66], and choose Alexnet as the network. Three different large-scale datasets are ImageNet, Places, and Places2 which are pre-trained on Alexnet respectively. At the last step, we choose CNN features + SVM or softmax for image classification. In total, there are six different combinations in our experiment: ImageNet-CNNs features + Alexnet + SVM, ImageNet-CNNs + Alexnet + softmax,

Table 3 The comparison results of our method and other published methods on the Sun RGB-D dataset

Method	RGB (%)	Depth (%)	RGB-D (%)
CNN-RNN [34]	35.6	26.1	39.2
GIST+RBF kernel SVM [60]	19.7	20.1	23.0
Places-CNN+Linear SVM [21]	35.6	25.5	37.2
Places-CNN+RBF kernel SVM [21]	38.1	27.7	39.0
SUN RGB-D+Alexnet [29]	24.3	–	30.7
SS-CNN-R6 [29]	36.1	–	41.3
Combine FV and CNN [65]	–	–	48.1
Places2-CNNs+softmax+Alexnet	41.7	32.1	42.3
Places2-CNNs+softmax+VGG [61]	43.5	**34.7**	45.1
LM-CNN	**46.4**	34.6	**48.7**

Bold numbers indicate the best results

Table 4 The comparison results of global methods and our method on the NYU Depth v1 dataset and the SUN RGB-D dataset

Methods		NYU Depth v1 dataset			SUN RGB-D dataset		
		RGB (%)	Depth (%)	RGB-D (%)	RGB (%)	Depth (%)	RGB-D (%)
Global	ImageNet-CNNs features + Alexnet + SVM	71.2	59.1	72.0	27.4	23.1	30.3
	ImageNet-CNNs + Alexnet + softmax	69.8	58.7	71.9	28.9	22.3	31.4
	Places-CNNs features + Alexnet + SVM	72.6	62.2	74.7	35.6	25.5	37.2
	Places-CNNs + Alexnet + softmax	72.4	64.3	73.8	36.2	24.6	38.7
	Places2-CNNs features + Alexnet + SVM	73.8	67.1	76.5	41.9	31.8	42.1
	Places2-CNNs + Alexnet + softmax	74.5	66.9	76.8	41.7	32.1	42.3
Local	**LM-CNN**	**74.7**	**67.8**	**79.3**	**46.4**	**34.6**	**48.7**

Bold numbers indicate the best results

Places-CNNs features + Alexnet + SVM, Places-CNNs + Alexnet + softmax, Places2-CNNs features + Alexnet + SVM, and Places2-CNNs + Alexnet + softmax. The comparison results are shown in Table 4.

As we can see from Table 4, LM-CNN outperforms other global fine-tuning methods by 4.9% at least on SUN RGB-D dataset. It indicates the advantages of our local fine-tuning pipeline of RGB-D region proposals and performing classification on top of them. According to other six global experiments, under the same dataset pre-trained Alexnet, different classifier choices result in little difference on classification performance. Therefore, the choice of the dataset for pre-training is mostly responsible for the classification accuracy. Following the same architecture as the network proposed in [22], the ImageNet dataset (ILSVRC 2012) for pre-training [19] contains 1.2 million widely various high-resolution images with 1000 different classes. Compared to the ImageNet dataset, Places is a scene-centric dataset with 205 scene categories and 2.5 million images with category labels, and Places2 is also a scene-centric dataset but with 8 million images from 401 scene categories. From the global results in Table 4, we can find that Places-CNN and Places2-CNN perform much better. It proves that a CNN network trained using a scene-centric dataset is able to achieve a significant improvement on a scene benchmark in comparison with a network trained using an object-centric dataset.

Ablation Study

In this section, we analyze the effectiveness of individual components in our pipeline. We discard one single component (i.e., region proposal screening, clustering, depth encoding, and local fine-tuning) and keep other components untouched. Table 5 shows a summary of the comparison results on NYU Depth v1 dataset and SUN RGB-D dataset. The final results of our full pipeline are also shown. We conduct a comprehensive analysis of these comparison results in the following sections.

Without Region Proposal Screening

We directly feed all the region proposals without any screening into the subsequent components (clustering, depth encoding and local fine-tuning). During our region proposal screening step, we discard the region proposals with lower weights (see Eq. 3). In this case, we define the screening ratio as the percentage of discarded region proposals in the screening step. It can be seen from Fig. 8 that when the

Table 5 Evaluation results of ablation studies on NYU Depth v1 and SUN RGB-D datasets

Configuration	NYU Depth v1 dataset			SUN RGB-D dataset		
	RGB (%)	Depth (%)	RGB-D (%)	RGB (%)	Depth (%)	RGB-D (%)
Without region proposal screening	73.9	67.4	77.8	41.9	33.7	42.8
Without discriminative region proposals clustering	73.7	67.2	77.3	41.7	33.4	42.5
Without depth region proposal encoding	74.7	57.2	74.9	44.3	21.2	44.6
Without local fine-tuning	72.5	66.4	76.9	40.1	29.7	41.6
Our full pipeline LM-CNN	**74.7**	**67.8**	**79.3**	**44.3**	**34.6**	**48.7**

Bold numbers indicate the best results

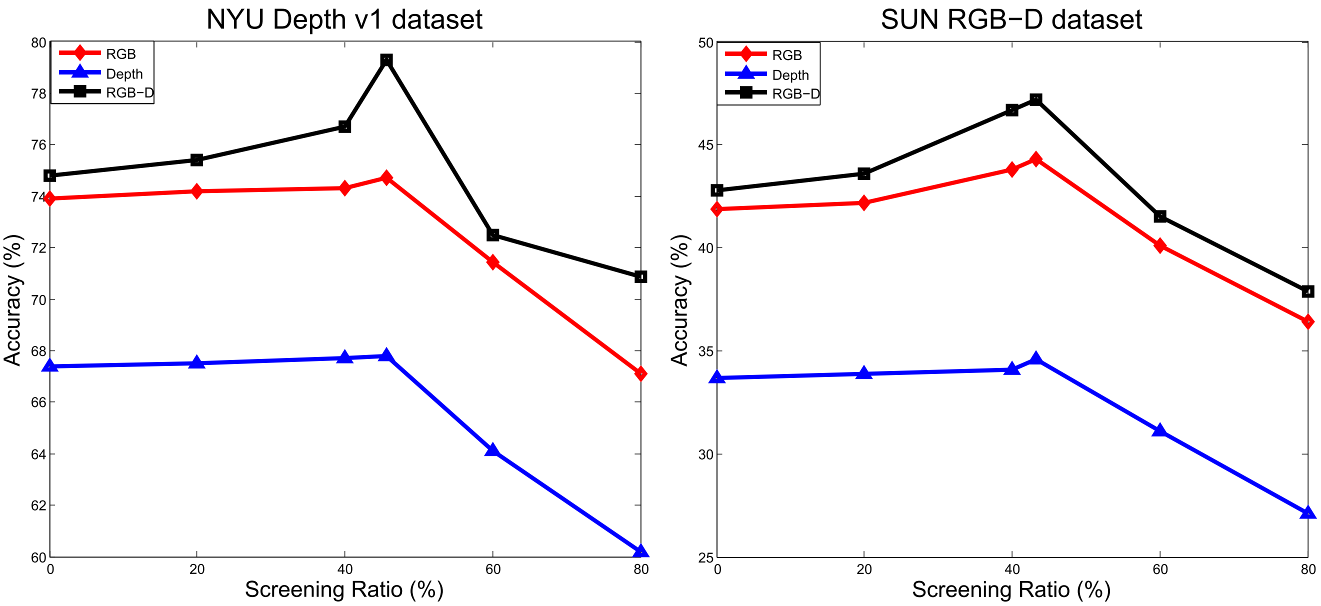

Fig. 8 Recognition accuracies with different screening ratios on NYU Depth v1 and SUN RGB-D

screening ratio is 0, the recognition accuracies of RGB, Depth, and RGB-D in NYU Depth v1 dataset are 73.9% 67.4%, and 77.8% and in SUN RGB-D dataset are 41.9% 33.7%, and 42.8%. All of the recognition accuracies without region proposal screening are lower than those with the full pipeline (Table 5). This is because that, although we have obtained reasonable region proposals through the Edge Boxes method, there still exist some proposals which are either false positives or unrepresentative objects shared by scene categories. On the contrary, a screening ratio that is too high will also result in a low recognition performance, because too high ratios will discard some discriminative region proposals. An optimal ratio can be obtained through cross validation on a subset of the training data.

Without Discriminative Region Proposals Clustering

We directly take the collection of screened region proposals as a large codebook and treat each region proposal as a visual word without region proposals clustering. Then, we utilize LSAQ [67] (with 100 nearest neighbors) coding and SPM pooling to build the imagelevel representation of RGB

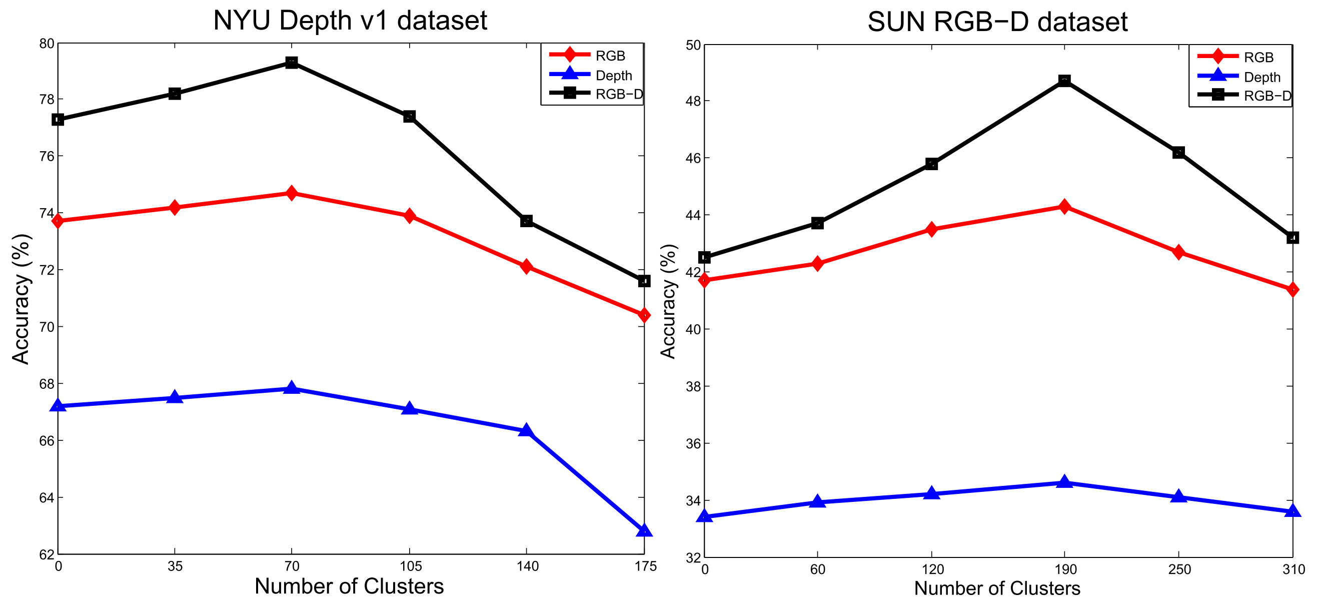

Fig. 9 Recognition accuracies with different number of clusters on NYU Depth v1 and SUN RGB-D

Table 6 Comparison of different depth encoding methods on NYU Depth v1 and SUN RGB-D datasets

Depth encoding	NYU Depth v1 dataset		SUN RGB-D dataset	
	Depth (%)	RGB-D (%)	Depth (%)	RGB-D (%)
Without encoding	57.2	74.9	21.2	44.6
HHA	66.9	78.3	34.2	47.1
Jet	**67.8**	**79.3**	**34.6**	**48.7**

Bold numbers indicate the best results

and depth images. The recognition performances on both of NYU Depth v1 and SUN RGB-D datasets are lower than the performances of our method (Table 5), which illustrates that the discriminative region proposals clustering step in our pipeline is crucial. In this component, the common semantic meaning shared among similar proposals is emphasized, and the less important differences among them are tolerated. Therefore, discriminative region proposals clustering can improve the generality and representativeness of the discovered region proposals. Figure 9 shows the recognition accuracies with different numbers of clusters on NYU Depth v1 and SUN RGB-D. It shows that an overly small number of clusters will assign different semantic meanings to the same discriminative region proposals, which results in a poor performance. On the other hand, too many clusters will also result in low recognition accuracies due to the poor generality of the semantic meanings of discriminative region proposals.

Without Depth Region Proposal Encoding

After obtaining the depth region proposals, we directly use these depth images as inputs for our multi-modal local fine-tuning architecture. It can be seen from Table 5 that the recognition accuracies of depth and RGB-D become low without depth encoding. In addition, we conduct experiments to compare two different depth encoding methods (HHA and Jet) described in the Discriminative Region Proposals Clustering section. Both of these two encoding methods produce colorized images. Compared to HHA encoding which requires additional image preprocessing, the colorizing depth process of Jet encoding has negligible computational overhead. From the results, presented in Table 6, it is clear that using Jet encoding method yields slightly better performance than the HHA encoding method.

Without Local Fine-tuning

We directly utilize the responses from the RIM clustering model for pooling without local fine-tuning. From the comparison results of our experiments on NYU Depth v1 dataset in Table. 5, the recognition rate of RGB-D without local fine-tuning is 76.9%, which is around 2.4% lower than that with local fine-tuning. Similarly, the recognition rate of RGB-D on SUN RGB-D dataset without local fine-tuning is 41.6%, which is around 7.1% lower than that with local fine-tuning. This illustrates that in our pipeline, local fine-tuning is consistent with the common sense, which can better define separation boundaries between clusters.

Conclusion

In this paper, we present a CNNs-based local multi-modal feature learning framework (LM-CNN) for RGB-D scene classification. LM-CNN is built on the top of the pre-trained CNNs model. We firstly perform a region proposal extraction method on an RGB-D dataset. Then, we apply proposal screening on these generated region proposals to select the representative region proposals and group these selected proposals into clusters through the RIM algorithm. Aiming to leverage large CNNs trained for proposal recognition on the ImageNet dataset, we use an effective encoding method from depth to image data. Our local fine-tuning multi-modal model consists of a two-stream convolutional neural network that can learn fusion information from both RGB and depth proposals before classification. At last, the multi-level scene image representation is built from top of the probability distribution of the region proposals. Experiments are conducted on both NYU v1 depth and SUN RGB-D datasets to test the performance of our method thoroughly. The experimental results show that LM-CNN outperforms state-of-the-art approaches.

We believe our work has the potential to significantly improve RGB-D scene understanding. An extended evaluation shows that our local fine-tuning method outperforms direct global fine-tuning methods. The experiments also show that CNNs trained using a scene-centric dataset are able to achieve an improvement on scene benchmarks compared to a network trained using an object-centric dataset. Our work opens up many possibilities for learning rich, expressive local features from raw RGB-D images. Exploring other local fine-tuning methods and RGB-D fusion architectures would be a popular direction in the future.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Lu X, Li X, Mou L. Semi-supervised multitask learning for scene recognition. IEEE Trans Cybern. 2015;45(9):1967–1976.
2. Zhuo W, Salzmann M, He X, Liu M. Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference. In: IEEE Conference on computer vision and pattern recognition, no. EPFL-CONF-227441. 2017.
3. Cong Y, Liu J, Yuan J, Luo J. Self-supervised online metric learning with low rank constraint for scene categorization. IEEE Trans Image Process. 2013;22(8):3179–3191.
4. Lu X, Wang B, Zheng X, Li X. Exploring models and data for remote sensing image caption generation, IEEE Transactions on Geoscience and Remote Sensing.
5. Yu J, Tao D, Rui Y, Cheng J. Pairwise constraints based multiview features fusion for scene classification. Pattern Recogn. 2013;46(2):483–496.
6. Gao Y, Wang M, Tao D, Ji R, Dai Q. 3-D object retrieval and recognition with hypergraph analysis. IEEE Trans Image Process. 2012;21(9):4290–4303.
7. Bian W, Tao D. Biased discriminant euclidean embedding for content-based image retrieval. IEEE Trans Image Process. 2010;19(2):545–554.
8. Lu X, Chen Y, Li X. Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features. IEEE Trans Image Process. 2018;27(1):106–120.
9. Cheng G, Zhou P, Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans Geosci Remote Sens. 2016;54(12):7405–7415.
10. Cheng G, Li Z, Yao X, Guo L, Wei Z. Remote sensing image scene classification using bag of convolutional features. IEEE Geosci Remote Sens Lett. 2017;14(10):1735–1739.
11. Wang P, Li W, Gao Z, Zhang Y, Tang C, Ogunbona P. Scene flow to action map: a new representation for RGB-D based action recognition with convolutional neural networks, IEEE Conference on Computer Vision and Pattern Recognition.
12. Ma S, Bargal SA, Zhang J, Sigal L, Sclaroff S. Do less and achieve more: training CNNS for action recognition utilizing action images from the web. Pattern Recogn. 2017;68:334–345.
13. Yang W, Jin L, Tao D, Xie Z, Feng Z. Dropsample: a new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten chinese character recognition. Pattern Recogn. 2016;58:190–203.
14. Cheng G, Yang C, Yao X, Guo L, Han J. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns, IEEE Transactions on Geoscience and Remote Sensing.
15. Luo Y, Wen Y, Tao D, Gui J, Xu C. Large margin multi-modal multi-task feature extraction for image classification. IEEE Trans Image Process. 2016;25(1):414–427.
16. Montserrat DM, Lin Q, Allebach J, Delp EJ. Training object detection and recognition CNN models using data augmentation. Electron Imaging. 2017;2017(10):27–36.
17. Li J, Zhang Z, He H. Hierarchical convolutional neural networks for EEG-based emotion recognition. Cognitive Computation. 2017;10:1–13.
18. Feng S, Wang Y, Song K, Wang D, Yu G. Detecting multiple coexisting emotions in microblogs with convolutional neural networks. Cognitive Computation. 2017;10:1–20.
19. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 248–255.
20. Wan L, Zeiler M, Zhang S, Cun YL, Fergus R. Regularization of neural networks using dropconnect. In: International Conference on Machine Learning; 2013. p. 1058–1066.
21. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. In: Neural Information Processing Systems; 2014. p. 487–495.
22. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems; 2012. p. 1097–1105.
23. Han J, Shao L, Xu D, Shotton J. Enhanced computer vision with Microsoft Kinect sensor: a review. IEEE Trans Cybern. 2013;43(5):1318–1334.
24. Cai Z, Han J, Liu L, Shao L. RGB-D datasets using Microsoft Kinect or similar sensors: a survey. Multimed Tools Appl. 2017;76(3):4313–4355.
25. Zrira N, Khan HA, Bouyakhf EH. Discriminative deep belief network for indoor environment classification using global visual features. Cognitive Computation. 2017;10:1–17.
26. Feichtenhofer C, Pinz A, Wildes RP. Temporal residual networks for dynamic scene recognition. In: IEEE Conference on computer vision and pattern recognition; 2017.
27. Gong Y, Wang L, Guo R, Lazebnik S. Multi-scale orderless pooling of deep convolutional activation features. In: European Conference on Computer Vision; 2014. p. 392–407.
28. Yoo D, Park S, Lee J-Y, Kweon IS. Fisher kernel for deep neural activations, arXiv:1412.1628.
29. Liao Y, Kodagoda S, Wang Y, Shi L, Liu Y. Understand scene categories by objects: a semantic regularized scene classifier using convolutional neural networks. In: IEEE International Conference on Robotics and Automation; 2016. p. 2318–2325.
30. Gupta S, Arbeláez P, Girshick R, Malik J. Indoor scene understanding with RGB-D images: bottom-up segmentation, object detection and semantic segmentation. Int J Comput Vis. 2015;112(2):133–149.
31. Arbelaez P, Maire M, Fowlkes C, Malik J. Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell. 2011;33(5):898–916.
32. Bo L, Ren X, Fox D. Unsupervised feature learning for RGB-D based object recognition. In: Experimental Robotics; 2013. p. 387–402.
33. Lai K, Bo L, Ren X, Fox D. A large-scale hierarchical multi-view RGB-D object dataset. In: IEEE International Conference on Robotics and Automation (ICRA); 2011. p. 1817–1824.
34. Socher R, Huval B, Bath B, Manning C, Ng AY. Convolutional-recursive deep learning for 3D object classification. In: Neural Information Processing Systems; 2012. p. 665–673.
35. Socher R, Lin CC, Manning C, Ng AY. Parsing natural scenes and natural language with recursive neural networks. In: International Conference on Machine Learning; 2011. p. 129–136.
36. Cai Z, Shao L. RGB-D data fusion in complex space. In: IEEE International Conference on Image Processing. Beijing; 2017. p. 1965–1969.

37. Song S, Xiao J. Deep sliding shapes for amodal 3D object detection in RGB-D images.
38. Krause A, Perona P, Gomes RG. Discriminative clustering by regularized information maximization. In: Advances in Neural Information Processing Systems; 2010. p. 775–783.
39. Wang X, Yang M, Zhu S, Lin Y. Regionlets for generic object detection. In: IEEE International Conference on Computer Vision; 2013. p. 17–24.
40. Uijlings JR, van de Sande KE, Gevers T, Smeulders AW. Selective search for object recognition. Int J Comput Vis. 2013;104(2):154–171.
41. Lu X, Zhang W, Li X. A hybrid sparsity and distance-based discrimination detector for hyperspectral images. IEEE Trans Geosci Remote Sens. 2018;56(3):1704–1717.
42. Siva P, Xiang T. Weakly supervised object detector learning with model drift detection. In: International Conference on Computer Vision; 2011. p. 343–350.
43. Deselaers T, Alexe B, Ferrari V. Localizing objects while learning their appearance. In: European Conference on Computer Vision; 2010. p. 452–466.
44. Lu X, Zheng X, Yuan Y. Remote sensing scene classification by unsupervised representation learning. IEEE Trans Geosci Remote Sens. 2017;55(9):5148–5157.
45. Cheng M.-M., Zhang Z, Lin W.-Y., Torr P. Bing: binarized normed gradients for objectness estimation at 300fps. In: IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 3286–3293.
46. Arbeláez P, Pont-Tuset J, Barron J, Marques F, Malik J. Multiscale combinatorial grouping. In: IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 328–335.
47. Zitnick CL, Dollár P. Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision; 2014. p. 391–405.
48. Gu C, Lim JJ, Arbeláez P, Malik J. Recognition using regions. In: IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 1030–1037.
49. Carreira J, Sminchisescu C. Constrained parametric min-cuts for automatic object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition; 2010. p. 3241–3248.
50. Hosang J, Benenson R, Schiele B. How good are detection proposals, really?, arXiv:1406.6962.
51. Jia Y, Shelhamer E, Donahue J, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: International Conference on Multimedia; 2014. p. 675–678.
52. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–1471.
53. Vapnik V. The nature of statistical learning theory. 2013.
54. Gupta S, Girshick R, Arbeláez P, Malik J. Learning rich features from RGB-D images for object detection and segmentation. In: Europen Conference on Computer Vision; 2014. p. 345–360.
55. Yang J, Yu K, Gong Y, Huang T. Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 1794–1801.
56. Arandjelovic R, Zisserman A. All about VLAD. In: IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 1578–1585.
57. Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition; 2010. p. 3304–3311.
58. Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops); 2011. p. 601–608.
59. Song S, Lichtenberg SP, Xiao J. Sun rgb-d: A RGB-D scene understanding benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 567–576.
60. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis. 2001;42(3):145–175.
61. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556.
62. Le QV, Karpenko A, Ngiam J, Ng AY. ICA with reconstruction cost for efficient overcomplete feature learning. In: Advances in Neural Information Processing Systems; 2011. p. 1017–1025.
63. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y. Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition; 2010. p. 3360–3367.
64. Jin L, Gao S, Li Z, Tang J. Hand-crafted features or machine learnt features? Together they improve RGB-D object recognition. In: International Symposium on Multimedia; 2014. p. 311–319.
65. Wang A, Cai J, Lu J, Cham TJ. Modality and component aware feature fusion for RGB-D scene classification. In: IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 5995–6004.
66. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 1–9.
67. Liu L, Wang L, Liu X. In defense of soft-assignment coding. In: IEEE International Conference on Computer Vision; 2011. p. 2486–2493.