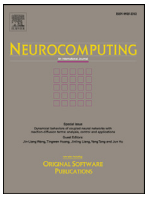
Neurocomputing 373 (2020) 81–89

Contents lists available at ScienceDirect

Neurocomputing

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



MSN: Modality separation networks for RGB-D scene recognition

Zhitong Xiong, Yuan Yuan*, Qi Wang

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 20 May 2019
Revised 20 August 2019
Accepted 18 September 2019
Available online 25 September 2019

Communicated by Dr Bo Du

Keywords:
Scene recognition
Deep learning
RGB-D
Multi-modal feature learning

ABSTRACT

RGB-D image based indoor scene recognition is a challenging task due to the complex scene layouts and cluttered objects. Although the depth modality can provide extra geometric information, how to better learn the multi-modal features is still an open problem. Considering this, in this paper we propose the modality separation networks to extract the modal-consistent and modal-specific features simultaneously. The motivations of this work are from two aspects: 1) The first one is to learn what is unique to each modality and what is common between the two modalities explicitly; 2) The second one is to explore the relationship between global/local features and modal-specific/consistent features. To this end, the proposed framework contains two branches of submodules to learn the multi-modal features. One branch is used to extract the individual characteristics of each modality by minimizing the similarity between two modalities. Another branch is to learn the common information between two modalities by maximizing the correlation term. Moreover, with the spatial attention module, our method can visualize the spatial positions where different submodules focus on. We evaluate our method on two public RGB-D scene recognition datasets, and new state-of-the-art results are achieved with the proposed framework.

1. Introduction

Visual recognition has been improved dramatically thanks to the development of deep learning methods [1]. Since deep neural networks can learn to extract high semantic level visual features, many computer vision tasks [2–5] have benefited from them, including scene recognition task. Although using deep convolutional neural networks (CNN) can boost the performance of scene classification, it is still a challenging task due to the complex spatial layout and large intra-class variation. Scene images are usually not object-centric, which is different from traditional image classification task. Thus directly using global CNN features for scene image classification is suboptimal. Considering this difference, large scene classification dataset Places [6] is released to provide enough data for scene classification research.

Although the performance can be improved using the scene dataset Places instead of object-centric dataset for pre-training, the global CNN features are still not flexible enough to capture the spatial variability of scene image. The large intra-class variation makes the scene recognition difficult.

As the global CNN features are not effective enough to represent the scene, several methods [7–11] have been proposed to leverage the local features. These approaches usually sample CNN features densely at different locations and scales of a scene image and encoding them via Fisher vectors (FV) [12] or Vector of Locally Aggregated Descriptors (VLAD) [13]. Nevertheless, densely sampling patches may introduce irrelevant noise to the learned features. To get rid of the noise, another type of method employs the object detection task to sample meaningful local patches. However, accurately detecting the cluttered objects in indoor scene is also a difficult task. Besides, some detected objects may be irrelevant for the scene classification.

Since RGB-D image can provide extra geometric information with the depth modality, scene recognition performance can be improved significantly with the extra depth modality. Nevertheless, how to effectively learn multi-modal features is still an open problem. To combine the features of RGB and depth modalities, the most popular method is to learn RGB and depth representations separately and then combine them together by concatenation or summation. However, directly concatenating these multi-modal features ignores the correlation between two modalities. Other methods propose to guarantee the consistency of multi-modal features (modal-consistent features), and these methods can indeed improve the RGB-D scene classification performance. Nevertheless, they neglect the modal-complementary (modal-specific) features between two modalities.

* Corresponding author.
E-mail addresses: xiongzhitong@gmail.com (Z. Xiong), y.yuan1.ieee@gmail.com (Y. Yuan), crabwq@gmail.com (Q. Wang).

https://doi.org/10.1016/j.neucom.2019.09.066

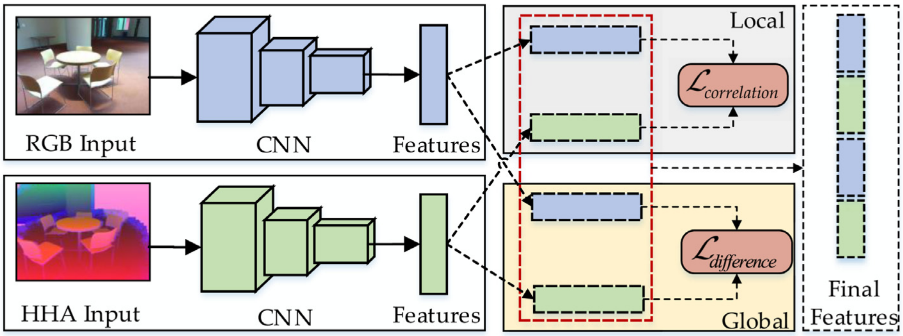


Fig. 1. The motivation of the proposed method. We find that it is more effective to extract modal-consistent information from local features, and modal-specific information from global features.

As aforementioned, global and local features are both important for scene recognition task. Since the global layout information and local object-level information are both useful to classify a scene. Besides that, modal-specific and modal-consistent features are also crucial for improving the RGB-D scene recognition performance. Although these features have been explored by prior research, the relationship between them are neglected. Previous works focus on either the global and local feature learning or the modal-specific and modal-consistent feature learning. However, the relationship between the global/local features and modal-specific/consistent features has not been explored. Thus how to integrate the global–local feature learning and modal-specific/consistent feature learning into an unified framework still needs more research efforts.

Considering the issues depicted above, in this work, we propose a novel modality separation network. Our motivation is to learn the modal-consistent and modal-complementary features simultaneously to extract more discriminative representations for RGB-D scene classification. As illustrated in Fig. 1, the proposed framework consists of two branches. The first branch is to learn the local modal-consistent features and another one is used to learn the global modal-specific features. Finally, these four feature vectors are concatenated together to represent the RGB-D scene image. Specifically, we make the following contributions:

(1) We propose a two-branch modality separation network to explicitly learn modal-specific and modal-consistent features. Global Modal-Specific (GMS) and Local Modal-Consistent (LMC) feature learning module are designed to learn modal-specific and consistent features simultaneously.
(2) We propose a local modal-consistent (LMC) feature learning module which consists of a spatial attention module and a key feature selection module to learn the local modal-consistent features. The local feature learning module selects local part features with the response map and triplet ranking loss. With this module, important local object-level features can be learned with no need of extra annotations.
(3) We explore the relationship between local/global and modal-specific/consistent features. To the best of our knowledge, the proposed method is the first work to explore whether the local or global features are more suitable for learning the modal-consistent or -specific information.

The remainder of this paper is organized as follows. In Section 2, the related works about scene recognition and multi-modal feature learning are reviewed. In Section 3, the proposed framework and sub-modules are presented in detail. In Section 4, the experimental results are discussed and analyzed. Some insights about this work is also presented in this section. Finally, in Section 5, the conclusion of this work is drawn.

2. Related work

We will review the previous work from two aspects: RGB-D scene classification methods and RGB-D multi-modal feature learning methods.

2.1. RGB-D scene classification

Global CNN features based methods. Zhou et al. [6] introduced a large scale scene classification dataset named Places and achieved significant performance improvement on this dataset than fine-tuning the ImageNet pre-trained models. Although training the deep CNN with large scale scene datasets can capture richer and more flexible scene representations, global CNN features are still not adequate to handle the great geometric and appearance variability of complex scene images.

Local features based methods. Gong et al. [8] designed a multi-scale CNN framework and aggregated the densely sampled multi-scale features using VLAD [13] to represent the scene image. Similarly, Yoo et al. [9] introduced Fisher Vectors to encode the sampled local features. Song et al. [14] proposed a two-stage training method for RGB-D scene classification. Zuo et al. [7] proposed a framework to learn a discriminative and shareable feature transformation filter bank for local image patches, and showed the effectiveness of complementary local features. Nevertheless, densely sampling image patches may introduce noise into learned models, which decreases the scene classification performance.

Object detection based methods. Wang et al. [15] proposed a framework to learn modality and component aware features for RGB-D scene classification. CNN region proposals were used as local features, and FV was used to combine these local features. Object detection was employed in the work of [16] to extract accurate object-level features, and the object-to-object relation was also modeled in their work. Although better classification performance can be obtained with more accurate local descriptors, these methods deeply rely on the performance of object detection. However, how to accurately detect cluttered objects in indoor scenes is still difficult. The error accumulation problem and extra computation cost are also limitations of these methods.

2.2. RGB-D multi-modal feature learning

Multi-modal feature learning has been investigated by considerable works, Yu et al. [17] proposed a multimodal hypergraph learning based sparse coding method for the click prediction. Multimodal Distance Metric Learning based method was proposed in [18]. Liu and Tao [19] proposed a promising multiview Hessian regularization algorithm to combine multi-modal features for semi-supervised learning. They further extended the local structure preserving method from Hessian to p-Laplacian [20] and achieved sig-

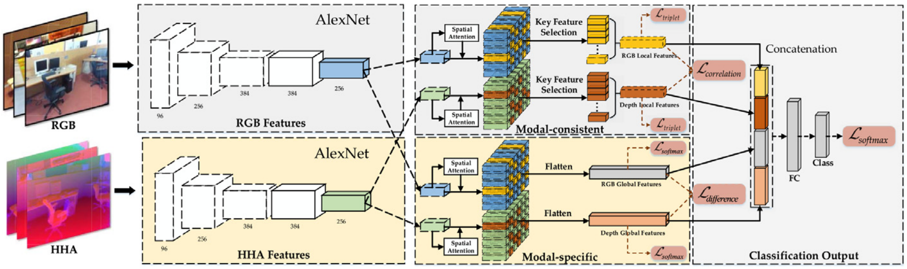


Fig. 2. The whole network architecture of the proposed framework. There are two branches of submodules for multi-modal feature learning: LMC module (upper branch) and GMS module (lower branch).

nificant performance. Additionally, varieties of strategies have been proposed to fuse the multi-modal features effectively [21]. These fusion methods can be roughly divided into four categories. 1) *Image level multi-modal fusion.* Couprie et al. [22] constructed the RGB-D Laplacian pyramid with the RGB modality and depth modality. 2) *Feature-level multi-modal combination.* Song et al. [23] concatenated the two-stream CNN features and fed them into one fully connected layer to fuse the multi-modal features. Song et al. [24] exploited a three-stream CNN to combine RGB branch and two depth branches of features with the element-wise summation. 3) *Modal correlative feature fusion.* To learn modal-consistent features, Wang et al. [25] enforced the RGB features to be close to the depth features with a correlation term. 4) *Modal correlative and distinctive features fusion.* Li et al. [26] attempted to learn the correlative and distinctive features simultaneously between the RGB and depth modality, and they achieved improved scene classification performance. Nevertheless, enforcing the consistency between multiple modalities obstructs the model from learning the modal-complementary features. Besides, these multi-modal feature fusion methods ignore the flexible local features.

3. Our method

The detailed framework will be introduced in this section. The whole proposed network is illustrated in Fig. 2. In this work, we use triplets of RGB-D images as input, and employ a margin ranking loss to select key local features. Firstly, three RGB and HHA [27] encoded image pairs including two samples $\{a, p\}$ with the same class label and one sample $\{n\}$ with different class label are sampled randomly to form a triplet $\{a, p, n\}$ input (a indicates the anchor sample, p is the positive sample and n is the negative one). Then the RGB CNN branch and Depth CNN branch are used for the first stage feature encoding. The last layer features of the two CNN branches are input to the proposed two submodules: LMC module and GMS module. With these two feature learning modules, local modal-consistent features and global modal-specific features are extracted for the final scene classification. The details of the LMC module and GMS module will be described in the following sections.

3.1. Local modal-consistent feature learning

The proposed local feature selection module consists of spatial attention, response map selection and triplet ranking loss. However, it is not a simple combination of existing works. 1) The response map selection is proposed to rank and select important local activations, which can filter out features of background image patches. 2) The spatial attention module is used to encourage the model to focus on important local features. 3) The irrelevant patch features can be further filtered out by the triplet ranking loss. The triplet ranking loss in this work is used for regularizing the local feature selection, which is quite different from other works.

After extracting CNN features for the triplet input $\{a, p, n\}$ with the two CNN branches f_{RGB} and f_d, we can get three RGB feature maps $\{F_{RGB}x|x \in \{a, p, n\}\}$ and three depth feature maps $\{F_d x|x \in \{a, p, n\}\}$. These six feature maps are input to the LMC and GMS modules respectively. The LMC module consists of a spatial attention module and a key feature selection module. Additionally, an extra loss function is proposed for training the LMC module. We will describe this in detail in this section.

3.1.1. Spatial attention module

Different from existing methods, to learn the modal-consistent features, we do not enforce the two modal features to be correlative directly. We opt to propose a spatial attention module to enable the network to learn to focus on different spatial positions. Then we maximize the correlation loss function between the attended multi-modal features. Our intuition is to find which feature vectors on the input feature maps are important for different modalities to learn the modal-consistent features.

For simplicity, we take for example one input pair in the triplet: $F_{RGB} \in \mathbb{R}^{(N,C,H,W)}$ and $F_d \in \mathbb{R}^{(N,C,H,W)}$, where N, C, H, W are the batch size, channel number, height and width respectively. As shown in Fig. 3, in our work, the last layer feature maps are with 256 channels and 6×6 spatial size. The spatial attention module is adapted from the non-local neural networks [28]. We adopt the dot product as the similarity function, and the input features are firstly embedded by three convolution networks with 1×1 kernel. The dot-product similarity can be represented as

$$f = \theta(F_{RGB})^T \phi(F_{RGB}), \tag{1}$$

where θ and ϕ are two convolution operations. Then a softmax activation is applied to normalize the attention maps. The attended features can be obtained by

$$F'_{RGB} = softmax(\theta(F_{RGB})^T \phi(F_{RGB}))g(F_{RGB}), \tag{2}$$

where g is a 1×1 convolution layer and its output channel number is the same with the input channel number as shown in Fig. 3. Finally, a residual connection is connected to the attended features to output the final spatial attention results.

3.1.2. Key feature selection

Our motivation for key feature vectors selection is that a $C \times 1 \times 1$ feature vector in the last feature maps can represent a small patch corresponding to the original image. We aim to select the important feature vectors from the feature maps to get rid of the irrelevant local features.

Different from the patch sampling based and object detection based methods, our work select key region features from the high semantic level CNN feature maps, which can filter out irrelevant features and need less computation cost than object detection. Specifically, we first sum up the input feature maps along the channel axis, and a response map with shape (N, H, W) is obtained. Then we sort the response map and select the K highest

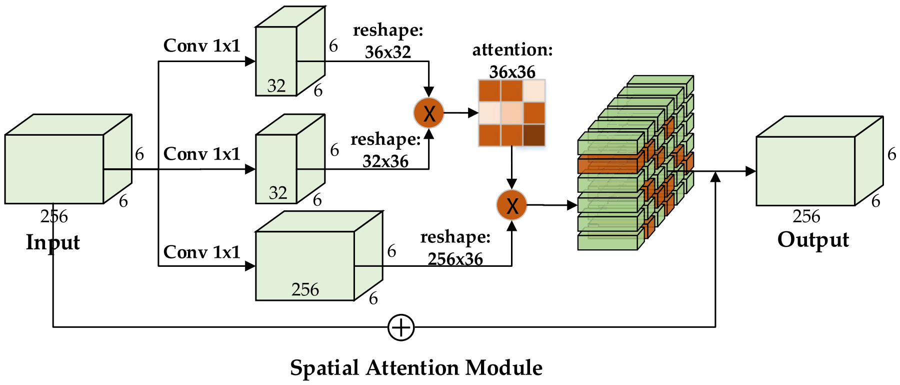


Fig. 3. The spatial attention architecture of the proposed framework.

response indexes *ind*. For each response map $F_{resp} \in \mathbb{R}^{(H\times W)}$ in the input batch, this selection process can be formulated as

$$ind = sort(F_{resp}),$$
$$ind = ind[1:K]. \tag{3}$$

With the selected indexes $ind \in \mathbb{R}^K$, we can index K feature vectors from the attended feature maps $F'_{RGB}x$ as the local features $\{E_x \in \mathbb{R}^{(N,C,K)} | x \in a, p, n\}$. However, some of the selected high response feature vectors may be irrelevant to the scene classification task. Thus we further propose a triplet correlation loss to select local features which are correlated between different images with the same scene class label. In this work, the triplet margin ranking loss is exploited to supervise the local feature selection module, and this loss function can be formulated as

$$\mathcal{L}_{\mathrm{RGB-triplet}} = \max\{\mathrm{d}(\mathrm{E_a}, \mathrm{E_p}) - \mathrm{d}(\mathrm{E_a}, \mathrm{E_n}) + \alpha, 0\},$$
$$d(x, y) = ||x - y||_2, \tag{4}$$

where α is the margin parameter and it is greater than zero [29]. For the selection of depth modal features, the loss computation is similar to the RGB modality depicted above. The total triplet loss for the two modalities is presented as

$$\mathcal{L}_{\mathrm{triplet}} = \mathcal{L}_{\mathrm{RGB-triplet}} + \mathcal{L}_{\mathrm{depth-triplet}}. \tag{5}$$

To sum up, we propose the response map based key feature selection module which is regularized by the triplet ranking loss. With the response map, we can filter out some background image patches, and the irrelevant patch features can be further filtered out by the triplet ranking loss.

3.1.3. Modal-consistent feature learning

With the selected local features for each modality, we further propose to encourage the multi-modal features for the same image to be close. To learn the local modal-consistent features, we employ the cosine similarity to measure the correlation between RGB local features and depth local features. This can be formulated as

$$\mathcal{L}_{\mathrm{correlation}} = \sum_{\mathrm{X}\in\{\mathrm{a,p,n}\}} 1 - \cos(\mathrm{E_{RGB}X}, \mathrm{E_d X}),$$
$$cos(x, y) = \frac{\langle x, y\rangle}{||x||\ ||y||}, \tag{6}$$

where $E_{RGB}x$ and $E_d x$ are the selected local features for RGB modality and depth modality respectively. For the LMC feature learning module, the total loss function can be represented as

$$\mathcal{L}_{\mathrm{LMC}} = \mathcal{L}_{\mathrm{triplet}} + \mathcal{L}_{\mathrm{correlation}}. \tag{7}$$

3.2. Global modal-specific feature learning

Although local features are flexible for representing the scene images, global features are also useful for describing the scene layouts. Considering that the RGB modality and depth modality have different characteristics from each other, we aim to learn the modal-complementary features through the global modal-specific feature learning module.

Our motivation for the GMS module is that the global scene features contain the full-image scene layout and appearance information. The RGB modality is more useful for providing the appearance information and the depth modality is more effective to provide the scene layout information. Thus we aim to encourage the different modalities to concentrate on different information and learn to extract the modal-complementary features.

To make full use of the information in each modality, the GMS module is designed to extract the modal-specific features. This module begins with a spatial attention module to enable different modalities to focus on different feature vectors. The spatial attention module is the same as the one introduced in the LMC module, and the output of it in GMS module can be denoted as F'_{RGB} and F'_d. Then the two modal features are flattened to be input to the fully connected layers. This can be represented as

$$V_{RGB} = SpatialAttention(F'_{RGB}),$$
$$V_{RGB} = SpatialAttention(F'_d), \tag{8}$$

where *SpatialAttention* is the spatial attention module, and V_{RGB} and V_d are the attended and flattened feature vectors of the two modalities. We use these two feature vectors for global modal-specific feature learning. In order to extract the individual characteristics for each modality, we use the auxiliary cross-entropy loss to force the different modalities input to have the same classification output. Specifically, the total cross-entropy loss with softmax is represented as

$$\mathcal{L}_{\mathrm{CE}}(\mathrm{X_{RGB}}, \mathrm{X_d}, \mathrm{y}) = -\log\frac{\mathrm{e}^{\mathrm{X_{RGBy}}}}{\sum_{\mathrm{j=1}}^{\mathrm{C}} \mathrm{e}^{\mathrm{X_{RGBj}}}} - \log\frac{\mathrm{e}^{\mathrm{X_{dy}}}}{\sum_{\mathrm{j=1}}^{\mathrm{C}} \mathrm{e}^{\mathrm{X_{dj}}}}, \tag{9}$$

where X_{RGB} and X_d are the output of the auxiliary fully connected layers, and y is the class label of the input RGB-D image.

Moreover, to enforce the model to learn the modal-specific features, we further propose to maximize the similarity between the global features of the two modalities. Specifically, the cosine similarity is employed to measure the correlation between the two modal features. The loss function can be represented as follows.

$$\mathcal{L}_{\mathrm{difference}} = \sum_{\mathrm{X}\in\{\mathrm{a,p,n}\}} \max(0, \cos(\mathrm{V_{RGB}X}, \mathrm{V_d X}) - \sigma), \tag{10}$$

where σ is the margin and it is set to 0.5 in this work. With the auxiliary cross-entropy loss and the modal-specific loss term, the total loss function for the GMS module can be formulated as

$$\mathcal{L}_{\text{GMS}} = \mathcal{L}_{\text{difference}} + \mathcal{L}_{\text{CE}}. \tag{11}$$

3.3. Multi-modal global and local feature learning

After the LMC and GMS feature learning module, we concatenate the learned local modal-consistent and global modal-specific feature together for the final scene classification.

$$F_{mm} = concat(E_{RGB}, E_d, V_{RGB}, V_d), \tag{12}$$

where F_{mm} is the final multi-modal global and local feature vector. E_{RGB} and E_d are the local modal-consistent features for two modalities. G_{rgb}, G_d are the global modal-specific features for the two modalities.

Then the final features are input to a fully connected layer and a softmax layer is used to output the final classification result. The parameters of the whole framework can be optimized jointly and the proposed framework can be trained in an end-to-end manner. The overall loss of our framework consists of the loss of LMC and GMS module and the final classification cross-entropy loss function. Thus the total loss function can be formulated as

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{LMC}} + \lambda_2 \mathcal{L}_{\text{GMS}}, \tag{13}$$

where λ_1 and λ_2 are the balancing weights to the different loss terms.

To sum up, the proposed framework takes triplet training samples as input and learn to extract the local modal-consistent features and global modal-specific features simultaneously for RGB-D scene training. All the modules can be optimized jointly and the whole network can be trained in an end-to-end manner.

4. Experiments

To evaluate the effectiveness of our proposed modules, we conduct RGB-D scene classification experiments on two public datasets: SUN RGB-D [23] and NYU Depth v2 (NYUD v2) [34]. **SUN RGB-D** dataset contains 10,355 RGB-D images, and these images are divided into 19 categories. To compare with the previous methods, we adopt the following experimental settings: 4845 images are used for training and 4659 images for testing. **NYUD v2** contains 1449 images and are divided into 10 categories: 9 common indoor scene types and one ‘others’ category. We use 795 images for training and 654 images for testing as the same with the previous work [35].

4.1. Parameters setup

We implement the proposed framework using the Pytorch deep learning platform. The HHA encoding is computed with the code from [27]. To further enhance the classification performance and prevent the deep model from overfitting, data augmentation is used in our work. All three images in the triplet input are firstly resized to 224 × 224, and then random erasing [36], random flip and random rotation are used for RGB and depth modality at a probability of 0.5. To compare with current state-of-the-art methods, AlexNet [1] pre-trained on Places dataset is employed as the backbone network. The initial learning rate is set to 1e-4 and is reduced by a fraction of 0.9 every 50 epochs. The proposed method is trained with no more than 300 epochs. Adam [37] optimizer is adopted for the network training, and the batch size is set to 64. To jointly train all the modules, we set the parameters λ_1 and λ_2 to 1 in all of our experiments. For local feature selection, K is set to 16 in our work.

Since the proposed framework includes two global feature learning subnetworks and two local feature selection networks, we will describe the architecture details in this section. For different CNN backbones, the sizes of final feature maps are different. The final feature maps of AlexNet are with the size of $256 \times 6 \times 6$. As for ResNet18, the feature maps are with the size of $512 \times 7 \times 7$. The local feature selection module employs a *stride* 1 convolution layer with kernel size 1×1, and the number of output channels is the same with the input feature maps (256 for AlexNet and 512 for ResNet18). K is set to 16 for selecting local features, so the final local features are with the size of 256*16*2 for two modalities. By concatenating global and local features of two modalities, the final multi-modal features are with the size of $256 * 16 * 2 + 256 * 2$. Then the final features are input to the final classification networks to output the scene recognition results.

4.2. Evaluation measurement

To compare with previous works on RGB-D scene recognition, we use mean class accuracy as the evaluation metric. The mean class accuracy is the mean accuracy over categories, which is computed by averaging the accuracies of all the categories.

$$MeanAcc = \frac{1}{C} \sum_{c=1}^{C} \frac{correct_c}{Num_c}, \tag{14}$$

where $correct_c$ is the number of correctly predicted samples of class c, and Num_c is the total number of samples of class c.

4.3. SUN RGB-D dataset

Eight state-of-the-art methods are compared on SUN RGB-D dataset. The results are shown in Table 1. We have also summarized the type of local features and multi-modal feature learning for each method. Among them, Song et al. [23] concatenated RGB and HHA features for scene classification. Liao et al. [30] used the image-level fusion to combine the RGB channels with depth and normal channel. Moreover, they further built a multi-task learning task for the scene recognition and semantic segmentation. Zhu

Table 1
Experimental results on SUN RGB-D dataset.

	Methods	Local features	Multi-modal fusion	Accuracy(%)
State-of-the-art	[23]	No	Feature-level Concatenation	39.0%
	[30]	No	Image-level Concatenation	41.3%
	[31]	No	inter- & intra- modality correlation	41.5%
	[15]	CNN Proposals	Local & Global Features Concatenation	48.1%
	[16]	Object Detection	Local & Global Features Concatenation	54.0%
	[32]	Patches Sampling	Feature-level Concatenation	53.8%
	[33]	Feature Selection	Local & Global features	55.9%
	[26]	No	Modality Distinction & Correlation	54.6%
Proposed	Our method	Key Feature Selection	Global & Local Features Modality Distinction & Correlation	**56.2**%

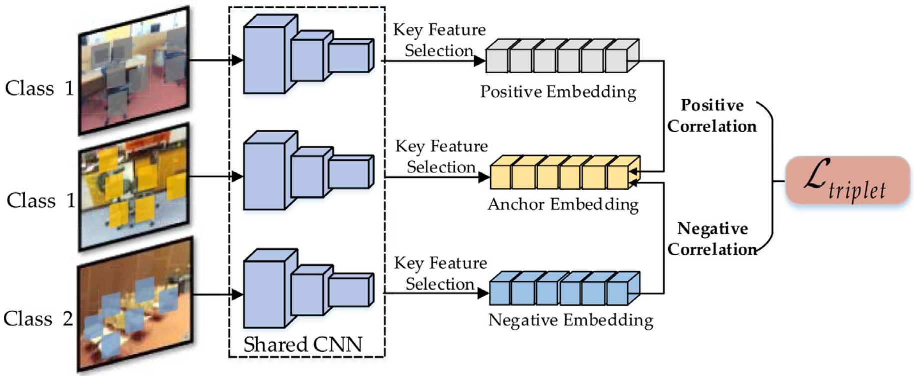


Fig. 4. The whole network architecture of the proposed framework.

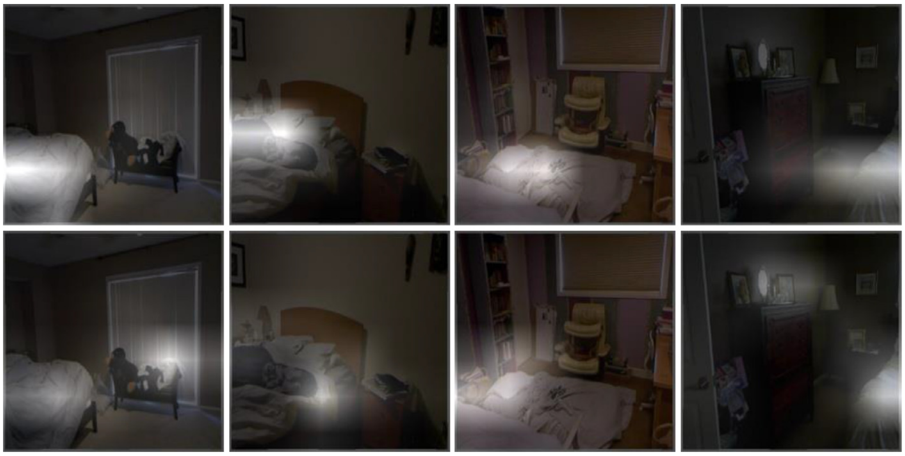


Fig. 5. The illustration of the attention map for LMC module. The first row is the RGB attention map and the second row is the HHA attention map (We use RGB image instead of HHA for more clear comparison). We can see that they focus on similar positions to learn modal-consistent features.

et al. [31] modeled the intra-class and inter-class modality correlations for scene classification [15]. and [16] are object detection based methods. Li et al. [26] designed a framework to learn distinctive and correlative multi-modal features simultaneously. As shown in Table 1, the proposed method outperforms all existing methods and achieves new state-of-the-art performance 56.2%.

From the experimental results we can see that object detection based methods [15,16] are more effective than other methods with no use of local features. Moreover, considering the correlation and distinction for multi-modal feature learning [26] can also improve the classification performance. Our method needs no object detection or proposals for local feature learning, and it can still outperform object-detection based methods. This indicates the effectiveness of the proposed framework.

4.4. NYUD V2 dataset

Similar to SUN RGB-D dataset, seven current state-of-the-art methods are compared on the NYU v2 dataset. Table 2 shows the experimental results of comparison methods, the proposed method achieves significant performance improvement (68.1%) to existing methods on NYUD v2 dataset. Moreover, the proposed method is more efficient than object detection based methods and can still outperform them in classification performance.

4.5. Ablation study

To evaluate the proposed modules more comprehensively, we conduct ablation studies on NYU v2 dataset to show the effect of the proposed submodules. As shown in Table 3, using features of single modality is not discriminative enough for indoor scene classification. We also conduct experiments to show the effect of LMC module and GMS module. The LMC module can improve the performance of RGB-D baseline significantly from 61.5% to 66.1%, and the GMS module can improve the performance over baseline method by 3.8%.

Table 2
Experimental results on NYUD v2 Dataset .

	Methods	Local features	Accuracy(%)
State-of-the-art	[38]	No	45.4%
	[15]	CNN Proposals	63.9%
	[26]	No	65.4%
	[14]	Local Patches	65.8%
	[33]	Feature Selection	67.8%
	[32]	Patches Sampling	67.5%
	[16]	Object Detection	66.9%
Proposed	Our method	Key Feature Selection	**68.1**%

We also do ablation study to explain why we choose to learn local modal-consistent features and the global modal-specific. We use GMC and LMS to represent the global modal-consistent and local modal-specific feature learning module. The results in Table 3 shows that 'GMS & LMC' achieves the best performance, which supports the motivation of the proposed method and shows the effectiveness of the framework.

Some visualization examples of the GMS and LMC attention maps are presented in Fig. 5 and Fig. 6. From these figures we can see that similar features are selected for LMC module and different features are focused for GMS module.

Table 3
Ablation study on NYUD v2 Dataset .

Methods	Accuracy(%)
RGB	53.5%
Depth(HHA)	51.1%
RGB-D(HHA)	61.5%
RGB-D *Triplet*$^+$	64.3%
RGB-D Spatial Attention	65.1%
RGB-D GMS	65.3%
RGB-D LMC	66.1%
RGB-D GMC & LMC	67.3%
RGB-D GMC & LMS	67.1%
RGB-D GMS & LMS	66.5%
RGB-D GMS & LMC	**68.1%**

4.6. Insights for multi-modal feature learning

The proposed modality separation networks learn to extract the local modal-consistent features and global modal-specific features simultaneously. As the LMC and GMS submodules employ spatial attention module to learn to focus on important CNN features, we can visualize the attention map to see what the submodules have learned.

4.6.1. Visualization of GMS attention maps

The illustration of the attention map for GMS module is shown in the upper two rows of Fig. 7. There are four pairs of attention maps. Similar to the LMC module, the up one is the RGB attention map (channel #0) and the bottom one is the attention map (channel #0) for HHA. We can see that the attention positions of the two modalities are different to learn the local modal-consistent features.

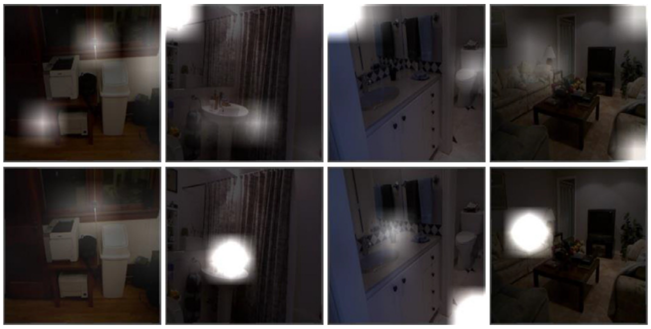


Fig. 6. The illustration of the attention map for GMS module. The first row is the RGB attention map and the second row is the HHA attention map (We use RGB image instead of HHA for more clear comparison). We can see that they focus on different positions to learn modal-specific features.

4.6.2. Visualization of LMC attention maps

The illustration of the attention map for LMC module is shown in the lower two rows Fig. 7. There are eight pairs of attention maps. For one image pair (up and bottom),the up one is the RGB attention map (channel #0) and the bottom one is the attention

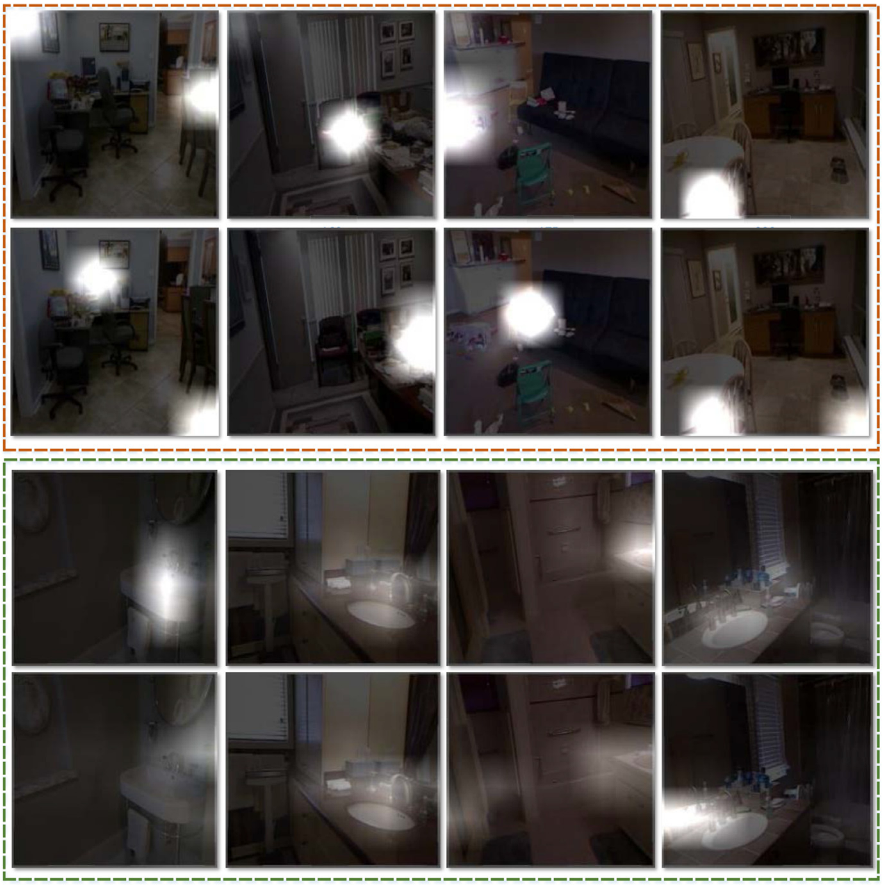


Fig. 7. The illustration of the attention map for GMS (upper two rows) and LMC module (lower two rows). The first rows of each module visualization are attention maps of RGB image, and the second rows are attention maps of depth image. (We still use the RGB image for more clear comparison.).

map (channel #0) for HHA. We can see that the attention maps of the two modalities focus on similar positions to learn the local modal-consistent features.

From the experimental results we find that local features are more suitable for modal-consistent feature learning and global features are more useful to extract modal-specific representations. The selected local features mainly represent the foreground objects, and the object-level features are more likely to be similar in different modalities. While the global features mainly contain the global scene layout information, the RGB appearance features of global scene may be quite different from the geometric depth representations. By exploring these relationships, we propose a better multi-modal global & local feature learning framework, which outperforms existing methods.

5. Conclusion

In this work, we present a RGB-D scene classification framework which consists of the LMC and GMS submodule to learn modal-consistent and modal-specific features simultaneously. The LMC module can learn to adaptively focus on important features to maximize the correlation between two modalities. Moreover, the GMS module learns to find important features to minimize the similarity between two modalities. We further explore the relationship between global/local and modal-specific/consistent feature learning, and the results provide interesting insights for the research of multi-modal feature learning. The proposed method achieves new state-of-the-art results on two public datasets, which indicates the effectiveness of the proposed method.

Declaration of Competing Interest

None.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant U1864204 and 61773316, State Key Program of National Natural Science Foundation of China under Grant 61632018, and Project of Special Zone for National Defense Science and Technology Innovation.

References

[1] 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.
[2] B. Zhao, X. Li, X. Lu, Hsa-rnn: hierarchical structure-adaptive rnn for video summarization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7405–7414.
[3] Q. Wang, J. Wan, X. Li, Robust hierarchical deep learning for vehicular management, IEEE Trans. Veh. Technol. 68 (5) (2018) 4148–4156.
[4] Q. Wang, J. Gao, X. Li, Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes, IEEE Trans. Image Process. 28 (9) (2019) 4376–4386.
[5] Y. Xu, B. Du, L. Zhang, Beyond the patchwise classification: spectral-spatial fully convolutional networks for hyperpsectral image classification, IEEE Trans. Big Data (2019), doi:10.1109/TBDATA.2019.2923243.
[6] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.
[7] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, X. Jiang, Learning discriminative and shareable features for scene classification, in: ECCV, 2014, pp. 552–568.
[8] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: ECCV, 2014, pp. 392–407.
[9] D. Yoo, S. Park, J.-Y. Lee, I.-S. Kweon, Fisher kernel for deep neural activations, 2014. abs/1412.1628
[10] Y. Xu, B. Du, F. Zhang, L. Zhang, Hyperspectral image classification via a random patches network, ISPRS J. Photogramm. Remote Sens. 142 (2018) 344–357.
[11] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, X. Li, Hierarchical feature selection for random projection, IEEE Trans. Neural Netw. Learn. Syst. 30 (5) (2018) 1581–1586.
[12] J. Sánchez, F. Perronnin, T. Mensink, J.J. Verbeek, Image classification with the fisher vector: theory and practice, Int. J. Comput. Vis. 105 (3) (2013) 222–245.
[13] H. Jegou, 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.
[14] X. Song, L. Herranz, S. Jiang, Depth CNNs for RGB-d scene recognition: learning from scratch better than transferring from RGB-CNNs, in: AAAI, 2017, pp. 4271–4277.
[15] 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.
[16] X. Song, C. Chen, S. Jiang, Rgb-d scene recognition with object-to-object relation, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 600–608.
[17] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.
[18] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. 47 (12) (2016) 4014–4024.
[19] W. Liu, D. Tao, Multiview hessian regularization for image annotation, IEEE Trans. Image Process. 22 (7) (2013) 2676–2687.
[20] W. Liu, X. Ma, Y. Zhou, D. Tao, J. Cheng, *p*-Laplacian regularization for scene recognition, IEEE Trans. Cybern. 49 (8) (2018) 2927–2940.
[21] Q. Wang, M. Chen, F. Nie, X. Li, Detecting coherent groups in crowd scenes by multiview clustering, IEEE Trans. Pattern Anal. Mach. Intell. (2018), doi:10.1109/TPAMI.2018.2875002.
[22] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor semantic segmentation using depth information, in: 1st International Conference on Learning Representations, {ICLR} 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Conference Track Proceedings, 2013.
[23] 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.
[24] X. Song, S. Jiang, L. Herranz, Combining models from multiple sources for RGB-d scene recognition, in: IJCAI, 2017, pp. 4523–4529.
[25] A. Wang, J. Cai, J. Lu, T.-J. Cham, Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition, in: IEEE International Conference on Computer Vision, 2015, pp. 1125–1133.
[26] Y. Li, J. Zhang, Y. Cheng, K. Huang, T. Tan, Df^{2}net: discriminative feature learning and fusion network for RGB-D indoor scene classification, in: the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[27] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-d images for object detection and segmentation, in: ECCV, Springer, 2014, pp. 345–360.
[28] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[29] V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, Learning local feature descriptors with triplets and shallow convolutional neural networks, in: BMVC, 2016, p. 3.
[30] 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: ICRA, IEEE, 2016, pp. 2318–2325.
[31] 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.
[32] X. Song, S. Jiang, L. Herranz, C. Chen, Learning effective RGB-d representations for scene recognition, IEEE Trans. Image Process. 28 (2) (2018) 980–993.
[33] Z. Xiong, Y. Yuan, Q. Wang, Rgb-d scene recognition via spatial-related multi–modal feature learning, IEEE Access 7 (2019) 106739–106747.
[34] P.K. Nathan Silberman Derek Hoiem, R. Fergus, Indoor segmentation and support inference from RGBD images, ECCV, 2012.
[35] 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.
[36] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, CoRR (2017) arXiv preprint arXiv:1708.04896.
[37] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
[38] S. Gupta, P. Arbeláez, R. Girshick, J. Malik, Indoor scene understanding with rgb-d images: bottom-up segmentation, object detection and semantic segmentation, Int. J. Comput. Vis. 112 (2) (2015) 133–149.

Zhitong Xiong received the M.E. degree in Northwestern Polytechnical University and is currently working toward the Ph.D. degree with the School of Computer Science and Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xin, China. His research interests include computer vision and machine learning.

Yuan Yuan (M’05-SM’09) is currently a Full Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xin, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (M’15-SM’15) received the B.E. degree in automation and the Ph.D. degree in pat- tern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xin, China. His research interests include computer vision and pattern recognition.