ASK: Adaptively Selecting Key Local Features for RGB-D Scene Recognition

Zhitong Xiong, Yuan Yuan*, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—Indoor scene images usually contain scattered objects and various scene layouts, which make RGB-D scene classification a challenging task. Existing methods still have limitations for classifying scene images with great spatial variability. Thus, how to extract local patch-level features effectively using only image label is still an open problem for RGB-D scene recognition. In this paper, we propose an efficient framework for RGB-D scene recognition, which adaptively selects important local features to capture the great spatial variability of scene images. Specifically, we design a differentiable local feature selection (DLFS) module, which can extract the appropriate number of key local scene-related features. Discriminative local theme-level and object-level representations can be selected with DLFS module from the spatially-correlated multi-modal RGB-D features. We take advantage of the correlation between RGB and depth modalities to provide more cues for selecting local features. To ensure that discriminative local features are selected, the variational mutual information maximization loss is proposed. Additionally, the DLFS module can be easily extended to select local features of different scales. By concatenating the local-orderless and global-structured multi-modal features, the proposed framework can achieve state-of-the-art performance on public RGB-D scene recognition datasets.**

***Index Terms*—RGB-D recognition, Local feature selection, Multi-modal feature learning**

I. Introduction

Scene recognition is a fundamental task for computer vision. Recent progress made in deep convolutional neural networks (CNNs) has greatly boosted the performance of various computer vision tasks, such as image classification [1], [2], object detection [3], [4], semantic segmentation [5], [6], [7] and video understanding [8], [9], [10] on large-scale benchmarks. Modern deep CNN architectures such as ResNet [11] and DenseNet [12] are well designed for high semantic-level image representation learning. However, directly applying these deep CNNs for scene recognition still suffers from a limitation: global image features are not flexible enough to represent the indoor scene image with cluttered objects and complex spatial layouts. Considering the difference between image classification and scene recognition, Zhou et al. [13] released a large scale scene classification dataset named *Places*. They showed the effectiveness of pre-training CNN parameters on Places instead of the object-centric dataset *ImageNet*.

The local object-level intermediate features are complementary to global CNN features [14]. Thus, selecting local features can be effective for taming the great geometric variability of scene image [15]. The local-feature based methods can be roughly divided into two categories. 1) Local patch-sampling based methods; 2) Object detection based methods. Several works [16], [17], [18] opted to extract features from different scales and locations densely and combined them via the fisher vector (FV) [19]. However, two disadvantages exist in these methods, which limit the further performance improvement. The first one is that global layout features are neglected. The second one is that densely-sampled local features may induce scene-irrelevant noise, which makes the learned features less discriminative. Object detection can also be used to extract more accurate object-level local features. However, the performance of these methods greatly relies on the accuracy of object detection. Unfortunately, detecting the cluttered objects accurately in complex indoor scenes is also nontrivial.

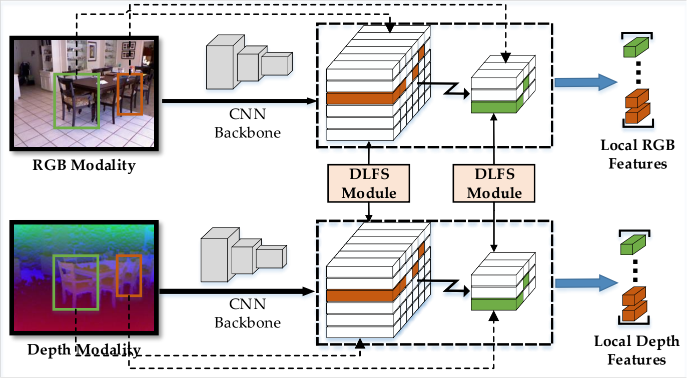


Fig. 1. Motivation of the proposed method. We aim to select local object-level feature vectors from multi-modality deep intermediate feature maps in an unsupervised manner.

Moreover, bottom-up local feature learning methods may also suffer from the following problems: Not all local object-level features contribute to the discriminative scene representations. For example, the 'chair' features in dining room and the 'chair' features in classroom are not discriminative for recognizing these two scenes. Merely using the local features may suffer from the ambiguity for recognizing different scenes. Some theme-level features are also important for scene classification, while object detection based methods may neglect them. For example, 'floor' and 'curtains' are background theme-level features, but they are also critical for recognizing scenes.

Considering the aforementioned limitations, how to extract scene-related local features for RGB-D indoor image in a weakly supervised manner is still under explored. Since RGB-D image contains spatially-aligned multi-modality information, making use of the multi-modal feature learning process

Corresponding author: Yuan Yuan, y.yuan.ieee@gmail.com.

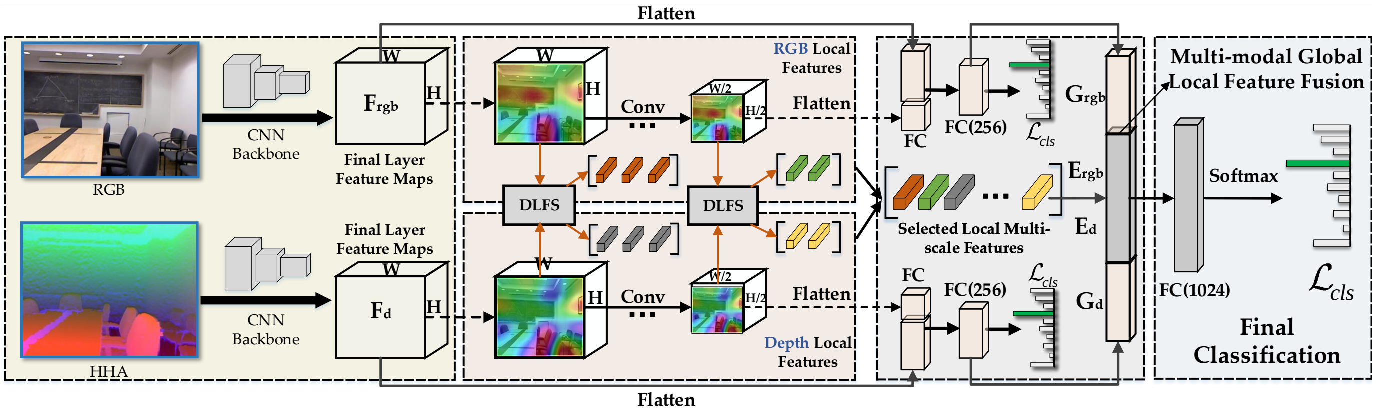


Fig. 2. Main architecture of the proposed method. 1) **Global modal-specific feature learning.** The final layer feature maps of two modalities are input to two fully connected (FC) layers to learn global discriminative features with the supervision of auxiliary cross-entropy loss separately. 2) **Local modal-specific feature learning.** Meanwhile, the final layer feature maps are also used to construct a multi-scale feature pyramid, and then DLFS modules adaptively select multi-scale intermediate CNN features for scene recognition.

is helpful for local feature extraction. In this work, we find that the correlation between local objects of RGB and depth modalities is stronger than background regions. Therefore, exploiting the correlation distribution of multi-modality feature can help to extract semantic local features, which has not been investigated by previous works. Based on this finding, a differentiable local feature selection (DLFS) module is designed to adaptively extract deep intermediate local features. Additionally, modality correlation loss and mutual information maximization loss are proposed for training the DLFS module discriminatively.

In this paper, we design a multi-modal feature learning framework for RGB-D scene recognition, which adaptively selects multi-scale key local features for scene recognition. As shown in Fig. 1, our motivation is to select multiple mid-level CNN feature vectors to represent local patches. To deal with the ambiguity problem caused by local features, the proposed framework also learns to extract global features which are important for describing the scene layout. The main contributions are as follows.

1) This work is the first to utilize the correlation between RGB and depth modalities to provide more cues for selecting local features. We design an effective differentiable local feature selection module based on the spatial-related correlation of multi-modal features.
2) We introduce a mutual information maximization loss for training the DLFS module, which encourages the discrimination of selected local features.
3) We design a compact global and local multi-modal feature learning framework to learn more discriminative representations for RGB-D scene recognition.

The rest of the paper is organized as follows. Related works about RGB-D scene recognition is shown in Section II. The detail of our method is presented in Section III. The experiments section IV demonstrated the performance of the proposed method. We compared the proposed method with its counterparts in detail in section V. Finally the conclusion of this paper is given in Section VI.

II. Related Work

In this section, the related works will be reviewed briefly. For local feature learning methods, patch-based, object detection based and CNN intermediate feature based methods are summarized and reviewed. Moreover, multi-modality feature learning methods for RGB-D data are also surveyed.

Additionally, more detailed comparison between the proposed method and related works are provided in section V.

Patch-sampling based methods. These approaches extract local features from the patch-based CNN intermediate representations. Gong et al. [16] designed a multi-scale CNN framework to sample the local patch features densely, and then encoded them via VLAD [20]. Some other methods [17], [21] represented the scene image with multi-scale local activations via the FV encoding. Depth image patches were exploited in the work of Song et al. [22]. They first trained the model with densely sampled depth patches in a weakly-supervised manner, and then fine-tuned the model with the full image. Nevertheless, densely sampled patch features may contain noise, which limits the scene recognition performance.

Object-detection based methods. To discard the irrelevant local features, several methods employed object detection for more accurate object-level features. Wang et al. [23] exploited the local region proposals to extract component representations, and encoded the local and global features together via fisher vector. Song et al. [24] employed Faster RCNN to detect objects on both RGB and depth images. More accurate object-level local features could be obtained with the object detection sub-module. They further modeled the object-to-object relation and achieved state-of-the-art scene recognition results [25]. Although improved performance could be obtained by [23], [24], the two-stage pipeline methods suffered from the error accumulation problem. The higher computational complexity is also a limitation of these methods. Moreover, detecting small objects in clutter in indoor scenes is nontrivial itself.

Selecting intermediate CNN representations is useful for many applications. KeypointNet [26] presented an end-to-end geometric reasoning framework to learn latent category-

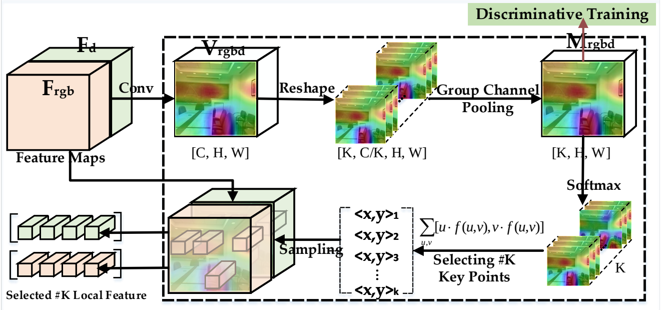


Fig. 3. Illustration of the proposed DLFS module.

specific 3D keypoints. KeypointNet can discover geometrically and semantically consistent keypoints adaptively without extra annotations. Zheng et al. [27] proposed a multi-attention convolutional neural network, which consisted of convolution, channel grouping and part classification sub-networks. Wang et al. [28] showed that intermediate CNN representations could be enhanced by learning a bank of convolutional filters that captured class-specific discriminative patches without extra bounding box annotations. Xu et al. [29] proposed deep regionlets for object detection, which could select non-rectangular regions within the detection framework.

Multi-modality feature learning is critical for RGB-D scene recognition, and various methods have been proposed [30]. Song et al. [31] fused the two modal features by concatenating them to one fully connected layer. Wang et al. [32] proposed to learn modal-consistent features between RGB and depth images. Li et al. [33] learned the modal-consistent and modal-distinctive embeddings between two modalities simultaneously. Spatial correspondence of local objects in RGB and depth modalities was exploited by [34] for multi-modal learning. The work in [35] employed cross-modal translation to explicitly regularize the training of scene recognition, which improved the generalization ability of the model.

III. Our Method

The whole network architecture of the proposed method is shown in Fig. 2. Depth image is firstly transformed to HHA (Horizontal disparity, Height above ground, Angle of the pixel's local surface normal with gravity direction) encoding [36]. RGB and HHA images are input to two branches of CNNs for feature extraction. Then, the final layer feature maps of the two modalities are used for global and local multi-modal feature learning. Finally, the global and local features of two modalities are concatenated together to form the final scene representation.

A. *Differentiable Local Feature Selection*

CNNs can learn to extract high semantic-level features with stacked convolution layers. Layer visualization shows that the intermediate CNN features are with high abstract-level and can represent object parts. Moreover, the work of [14] find that the global scene feature and the object-level local feature are complementary for scene recognition. Based on this insight, our motivation is to select multiple 1×1 mid-level CNN feature vectors to represent the local patches. Specifically, our model predicts K keypoints with the final layer feature maps of RGB and depth modalities, as illustrated in Fig. 3.

Suppose that the final layer feature map of RGB and depth modalities are F_{rgb} and F_d respectively. We then concatenate these two feature maps as F_{rgbd} to exploit multi-modal information for feature selection. The DLFS module takes F_{rgbd} as input, and uses a 1×1 convolution layer to learn the class-specific feature map for estimating K keypoints. This can be formulated as

$$V_{rgbd} = conv_{1\times1}(F_{rgbd}), \tag{1}$$

where $V_{rgbd} \in \mathbb{R}^{(C,H,W)}$ is the discriminative feature map used for predicting keypoints. C is the number of channels. H and W are the height and width of the feature maps respectively. $conv_{1\times1}$ is a convolution layer with 1×1 kernel size and output channel C. Although the channels of CNN feature maps are spatially-correlated, one separated channel is not enough to have strong part response [27]. Thus, the channels of V_{rgbd} are further grouped to be more spatially-correlated. Then we reshape feature V_{rgbd} to $V_{rgbd} \in \mathbb{R}^{(K,C/K,H,W)}$ for the following group channel pooling operation. To make the part response stronger, we sum up the channel groups of V_{rgbd} as

$$M^j_{rgbd} = \sum_{i=1}^{C/K} V_{rgbd_i}, \tag{2}$$

where $M_{rgbd} \in \mathbb{R}^{(K,H,W)}$ is the cross-channel grouped feature map for predicting K selected keypoints. $j \in \{1, 2, ..., K\}$ is the index of the channel groups.

If there are no accurate keypoint annotations, directly training a mapping function from the input feature maps to keypoints is difficult. The reason is that the mapping model is hard to converge without keypoint annotations for supervised training. Inspired by [26], we make our model output an attention map $h_j(u, v)$ to represent the probability of keypoint j occurring at position (u, v). In this work, we use the cross-channel grouped discriminative feature maps M_{rgbd} to output the attention map. Specifically, a 2D softmax layer is employed to produce the map h, which can be represented as

$$h_j = softmax(M^j_{rgbd}) \in \mathbb{R}^{(H,W)}, \tag{3}$$

where $h \in \mathbb{R}^{(K,H,W)}$ is the probability distribution map, and j is the index of predicted keypoints. Then the coordinates of the keypoints are computed by taking the expected values of the spatial distributions:

$$[x_j, y_j] = \sum_{u}^{H} \sum_{v}^{W} [u \cdot h_j(u, v), v \cdot h_j(u, v)], \tag{4}$$

where $[x_j, y_j]$ is the coordinate of the j^{th} predicted keypoint.

To extract the local features in an end-to-end manner, the feature vector-sampling module should be differentiable. Without loss of generality, let $E_{rgb} \in \mathbb{R}^{(K,C)}$ be the selected local features for RGB modality. Similarly, we denote E_d as the selected local features of the depth modality. Inspired

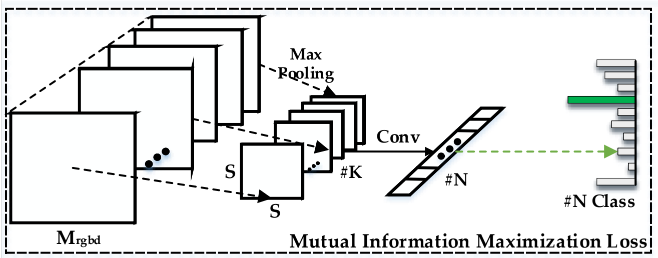


Fig. 4. Illustration of the proposed variational information maximization loss.

by [37], the differentiable bilinear feature sampling can be formulated as

$$E^c_{rgb_j} = \sum_u^H \sum_v^W F^c_{rgb}(u,v) max(0, 1-|x_j - v|) \\ \cdot max(0, 1-|y_j - u|), \tag{5}$$

where $j \in \{1, 2, ..., K\}$ is the index of K sampled local feature vectors, and $c \in [1...C]$ is the channel index. The coordinates (x_j, y_j) and (u, v) are normalized in the range of [-1,1]. Since the DLFS module takes the keypoint index and the feature maps as input, we need to compute the partial derivatives of F_{rgbd} as well as the predicted coordinates to allow the backpropagation through this module. The partial derivative w.r.t feature maps F_{rgb} and (x_j, y_j) are presented as follows. The partial derivative of feature maps F_{rgb} can be formulated as:

$$\frac{\partial E^c_{rgb}}{\partial F^c_{rgb}} = \sum_u^H \sum_v^W max(0, 1-|x_j - v|) \\ \cdot max(0, 1-|y_j - u|). \tag{6}$$

The partial derivatives of y_j is

$$\frac{\partial E^c_{rgb}}{\partial y_j} = \sum_{u,v}^{H,W} F^c_{rgb}(u,v) max(0, 1-|x_j - v|) g(v, y_j), \tag{7}$$

where $g(v, y_j)$ is a piecewise function, and it can be formulated as

$$g(v, y_i) = \begin{cases} 0, & \text{if} |v - y_j| \geq 1 \\ 1, & \text{if} v \geq y_j \\ -1, & \text{if} v \text{<} \mathrm{y}_j \end{cases} . \tag{8}$$

As for the depth modality F_d, the differentiable feature selection procedure is similar to the RGB modality.

For multi-scale local feature selection, we can construct feature pyramid with stride 2 convolutional layers as shown in Fig. 2. For each scale CNN feature maps, the local feature selection process is similar.

B. Discriminative Training for DLFS

Although the DLFS module can be trained in an end-to-end manner, the feature map M_{rgbd} is not guaranteed to be discriminative enough for selecting different local patch features. To encourage the channels of M_{rgbd} to be sensitive to different semantic parts, we propose a novel loss function to train the DLFS module.

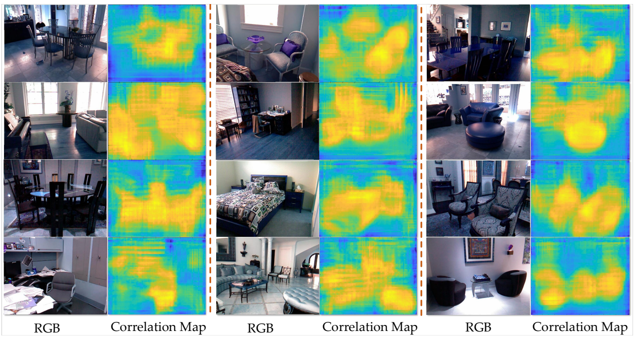


Fig. 5. Pixel-wise correlation map between deep features of RGB and depth modalities. RGB images are on the left, and the corresponding correlation-maps are displayed on the right. More visualization examples are displayed in Fig. 10.

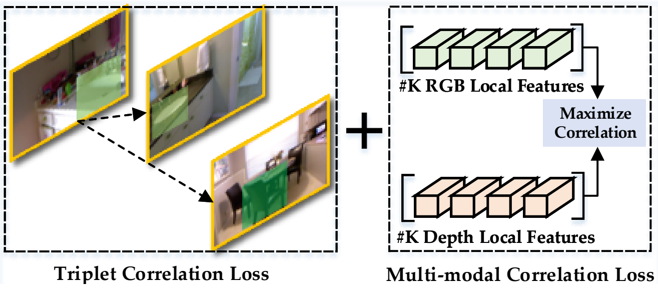


Fig. 6. Illustration of the proposed loss function for training DLFS module. The first term is the triplet-correlation loss, and the second one is the multi-modal local feature correlation loss.

In order to make M_{rgbd} more discriminative, we choose to retain high mutual information between M_{rgbd} and the scene class label. Since we aim to select feature vectors which are important for classifying different scenes, the feature map M_{rgbd} should be highly correlated to different scene classes. By maximizing the mutual information between M_{rgbd} and the class label L, the channels of M_{rgbd} is enforced to be class-correlated. As the exact computation of mutual information is intractable, variational lower bound is used for the approximation in this work.

Suppose that there are N classes, M_{rgbd} is firstly pooled by a max-pooling layer to get $T_{rgbd} \in \mathbb{R}^{S \times S \times K}$. Then a $S \times S$ convolution layer is used to transform T_{rgbd} into N scalar features. Finally, we get N features $U \in \mathbb{R}^N$ corresponding to N scene categories. The mutual information $I(L; U)$ can be defined as

$$\begin{aligned} I(L; U) &= H(L) - H(L|U) \\ &= H(L) + \mathbb{E}_{L,U}[\log q(L|U)] + KL \\ &\geq H(L) + \mathbb{E}_{L,U}[\log q(L|U)], \end{aligned} \tag{9}$$

where $H(\cdot)$ is the entropy, and KL denotes the Kullback-Leiber divergence. Maximizing the mutual information is equivalent to minimizing the following loss function $\mathcal{L}_{\text{VI}} = -\mathbb{E}_{L,U}[\log q(L|U)]$. $\mathcal{L}_{\text{VI}}$ can also be interpreted as the reconstruction error. In this work, we choose the variational distribution as a Gaussian distribution with heteroscedastic mean:

$$\mathcal{L}_{\text{VI}} = \sum_{n=1}^{N} \log \sigma_n + \frac{(L_n - U_{\mu_n})^2}{2\sigma_n^2} + c, \tag{10}$$

where c is a constant, and U_μ is the transformed features by FC layer as shown in Fig. 4. σ is a learnable parameter to be optimized. By minimizing $\mathcal{L}_{\text{VI}}$, the feature M_{rgbd} is encouraged to be discriminative for selecting key local features.

C. Unevenly distributed Multi-modal Correlation

We find that local objects contribute highly to the multi-modal correlation. To verify this, we compute the pixel-wise correlation map $P \in \mathbb{R}^{H,W}$ between the deep features $F_{rgb} \in \mathbb{R}^{C,H,W}$ of RGB modality and the depth modal deep features $F_d \in \mathbb{R}^{H,W}$. Then the pixel-wise correlation map can be defined as:

$$\begin{aligned} P^{ij} &= \rho({F_{rgb}}^{ij}, {F_d}^{ij}), \\ i &= 1, ..., H; j = 1, ..., W, \end{aligned} \tag{11}$$

where ρ is the cosine similarity function. ${F_{rgb}}^{ij} \in \mathbb{R}^C$ is a feature vector at position (i, j) of feature F_{rgb}. As shown in Fig. 5, we have visualized the pixel-wise correlation map P. It can be clearly seen that the correlation between two modalities is unevenly distributed. More specifically, large correlation values cluster on local objects. This indicates that multi-modal features of local objects have higher correlation than other spatial positions.

Based on this idea, local objects can be located by finding the largest correlation regions. Thus, we propose to maximize the correlation between RGB and depth local features to supervise the local feature selection module. The multi-modal correlation loss $\mathcal{L}_{Cm}$ can be computed as:

$$\mathcal{L}_{Cm} = \rho(E_{rgb}, E_d), \tag{12}$$

where ρ is the cosine similarity function. To further enhance the DLFS module with class-specific local features, triplet correlation loss is employed. As illustrated in Fig. 6, we aim to select local features that have larger correlation between the same class and smaller correlation between different classes.

For each sample in the triplet input $\{\mathbf{a}, \mathbf{p}, \mathbf{n}\}$, the corresponding selected local features are $\{E_{rgb}a, E_{rgb}p, E_{rgb}n\}$. $E_{rgb}p$ and $E_{rgb}a$ are positive and anchor features with the same class label, and $E_{rgb}n$ is the negative one with different scene class label. The triplet correlation loss is formulated as follows.

$$\mathcal{L}_{Crgb} = max\{\rho(E_a, E_p) - \rho(E_a, E_n) + \alpha, 0\}, \tag{13}$$

where $\mathcal{L}_{Crgb}$ is the triplet correlation loss for the RGB modality, and the loss computation for depth modality $\mathcal{L}_{Cd}$ is similar. α is the margin value and it is set to 1.0 in this work. The whole correlation loss for two modalities can be formulated as $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{Crgb} + \mathcal{L}_{Cd} + \mathcal{L}_{Cm}$.

D. Joint Global and Local Feature Representation

Though local features are effective for representing the scene images, merely using local features may suffer from the ambiguity problem. Thus in this work, we choose to learn the global and local features simultaneously to obtain more robust representations for scene recognition.

The global features are learned by the FC layers connected to the final feature maps F_{rgb} and F_d. As shown in Fig. 2, two auxiliary cross-entropy loss functions are employed for learning global modal-specific feature separately. This loss function can be formulated as

$$\mathcal{L}_{aux} = \mathcal{L}_{CE}(G_{rgb}, y) + \mathcal{L}_{CE}(G_d, y), \tag{14}$$

where L_{CE} represents the cross-entropy loss.

Finally, we concatenate the multi-modal global and local features together for the final scene classification. This can be denoted as

$$H_{mmgl} = concat(G_{rgb}, G_d, E_{rgb}, E_d), \tag{15}$$

where H_{mmgl} is the multi-modal global and local feature vector. Then H_{mmgl} is input to a fully connected layer, and the final classification result $\hat{y}$ is output through a softmax layer. We denote the final cross-entropy classification loss as $\mathcal{L}_{cls}$.

The overall loss function for training the proposed framework is a multi-task loss, which consists of four terms: 1) the global modal-specific auxiliary loss $\mathcal{L}_{aux}$; 2) the mutual-information maximization loss $\mathcal{L}_{\text{VI}}$; 3) the triplet correlation and multi-modal correlation loss $\mathcal{L}_C$; 4) the final classification loss $\mathcal{L}_{cls}$. This can be represented as

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{aux} + \lambda_2 \mathcal{L}_{\text{VI}} + \lambda_3 \mathcal{L}_C. \tag{16}$$

where λ_1, λ_2, and λ_3 are balancing weights of loss components for the multi-task loss function.

IV. Experiments

Two popular RGB-D scene recognition datasets are used to evaluate the proposed framework. One is the **SUN RGB-D** [31] dataset, which contains 10,355 RGB and depth image pairs. They are captured from different depth sensors including Kinect v1, Kinect v2, Asus Xtion and RealSense. These images are divided into 19 scene categories. To compare with existing methods, we follow the same experimental settings with [31]. For this dataset, 4,848 image pairs are used for training and 4,659 pairs for testing. Another dataset is the **NYUD v2** [38] dataset, which includes 1,449 RGB and depth image pairs divided into 10 categories. Following the experimental setting in [39], 795 image pairs are used for training and 654 pairs for testing.

A. Implementation Details

HHA encodings is computed with the code released by [36]. We use ResNet18 as the network backbone. Pre-trained parameters on **Places** are used for fine-tuning. Data augmentation is used in our work including random flip, cutout and random erasing [41].

TABLE I
ABLATION STUDY ON NYUD V2 DATASET

Feature Types	Methods	Mean-class Accuracy(%)
Single Modality	RGB / Depth / HHA	61.2 / 54.1 / 58.2
Multi-modality & Global	RGB-D (HHA Encoding)	64.2
	RGB-D Global ($\mathcal{L}_{aux}$)	65.5
	RGB-D Global (Spatial Attention [40])	66.1
Multi-modality & Local	RGB-D Local	64.1
	RGB-D Local ($\mathcal{L}_{\mathcal{C}}$)	66.2
	RGB-D Local ($\mathcal{L}_{\mathrm{VI}}$)	66.7
	RGB-D Local (Full DLFS)	67.3
Multi-modality & Global & Local	RGB-D Global & Local ($\mathcal{L}_{\mathcal{C}}$)	67.1
	RGB-D Global & Local ($\mathcal{L}_{\mathrm{VI}}$)	68.2
	RGB-D Global & Local (Full DLFS)	**68.9**
	RGB-D Global & Local (Two-Scale Full DLFS)	**69.3**

For optimization, Adam [42] is employed with an initial learning rate of 1e-4, and the learning rate is reduced by a fraction of 0.9 every 80 epochs. The batch size is set to 64 with shuffle. For all the experiments, 300 epochs are used to train the model. As for the multi-task training, we set the parameters λ_1 to 1, and λ_2, λ_3 are set to 0.1 for all experiments. Multi-scale DLFS is evaluated with the ResNet18 backbone. The first scale feature maps are with the size of 7×7. The second scale is obtained with a 1×1 kernel convolution layer. By setting the stride to 2, the second scale feature maps are with the size of 3×3. For the first scale, K is set to 16 and it is set to 4 for the second scale.

We randomly select 20% training samples for each scene category to form a validation set for both datasets. With this setting, the dataset is split into training/validation/test sets. Training set is used for model training, and validation set for model selection. Then the test set is used for model evaluation and performance comparison.

B. Ablation Study and Discussions

To evaluate the proposed framework comprehensively, we do the following ablation studies with ResNet18 backbone to explore the effect of different sub-modules. Additionally, to show the superiority of multi-modality RGB-D data and local feature, we also conduct experiments to compare the performance of different feature types.

*1) **Single Modality**:* The results are shown in Table I. The first row displays the accuracy of using single modality: including RGB image, depth image and HHA image. "RGB" denotes RGB image. "Depth" denotes the original depth image, which is not transformed to HHA encoding. While "HHA" denotes that HHA image is used as depth modality input instead of the original depth image. Since RGB image contains richer texture and appearance information than depth modality, using single RGB modality data can achieve better accuracy. As for the depth modality, "HHA" can obtain better performance than "Depth". This indicates the effectiveness of HHA encoding.

*2) **Multi-modality & Global**:* As shown in the table, "RGB-D (HHA Encoding)" is the method using global RGB and HHA features for scene classification, which can achieve obvious performance improvement compared with methods using single modality. This demonstrates that both RGB and depth modalities are useful for recognizing scenes. Additionally, we also study the effect of using auxiliary classification loss for multi-modality global features. "RGB-D Global ($\mathcal{L}_{aux}$)" improves the baseline method by 1.3%, which shows the effectiveness of the auxiliary loss for learning global modal-specific features.

Spatial attention mechanism can be used to focus on important local parts and extract more representative deep features. In this work, we use the non-local neural networks [40] to focus on different spatial regions with assigned weights. Although spatial attention can be used to learn local-sensitive deep features, the final features of method "RGB-D Global(Spatial Attention)" are still global features processed by fully connected layer. However, the proposed method selects key local features and abandon other features, which can be viewed as a kind of "hard-attention". Using spatial-attention to focus on local features softly obtains a performance of 66.1%, which is lower than the proposed local feature selection method.

*3) **Multi-modality & Local**:* We have also evaluated the performance of the proposed method when only using local features extracted by DLFS module. As shown in Table I, by training DLFS module with the proposed loss functions $\mathcal{L}_{\mathcal{C}}$ and $\mathcal{L}_{\mathrm{VI}}$, "RGB-D Local (Full DLFS)" can achieve better results than the baseline methods. This demonstrates that multi-scale discriminative local features are useful for scene recognition.

Since global RGB and depth features contain important scene layout information, which are complementary to local features. The proposed method exploits both local and global multi-modality features simultaneously, and can achieve better results than methods using global or local features alone.

*4) **Multi-modality & Global & Local**:* To study the effect of different loss functions of DLFS module, we do experiments to evaluate our framework with only $\mathcal{L}_{\mathcal{C}}$ and only $\mathcal{L}_{\mathrm{VI}}$. As shown in the table, "RGB-D Global & Local ($\mathcal{L}_{\mathcal{C}}$)" denotes the model with single-scale DLFS module and training loss

TABLE II
EXPERIMENTAL RESULTS ON SUN RGB-D DATASET

Methods	Local Features	Multi-Modality Learning	Mean-class Accuracy(%)
Song et al.[31]	No	Feature-level Fusion	39.0
Liao et al.[44]	No	Image-level Fusion	41.3
Zhu et al.[45]	No	Inter- & Intra- modality correlation	41.5
Wang et al.[23]	CNN Proposals	Local & Global Features Fusion	48.1
Song et al.[24]	Object Detection	Local & Global Features Fusion	54.0
Song et al. [46]	No	Global features	52.3
Song et al. [22]	Local Patches	Local & Global features Fusion	52.4
Li et al.[33]	No	Modality Distinction & Correlation	54.6
Song et al.[47]	Patches Sampling	Feature-level Fusion	53.8
Xiong et al.[34]	Feature Selection	Local & Global features Fusion	55.9
Song et al. [48]	Object Detection	Local & Global features Fusion	55.5
ASK (K=16 & K=4)	Local Feature Selection	Global & Local Features Modality Distinction	**57.3**

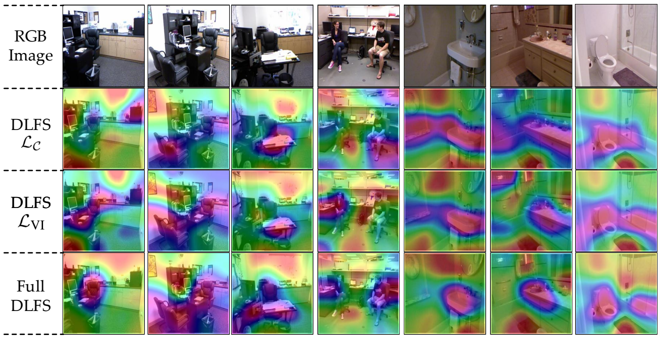


Fig. 7. Class-specific activation map (CAM) [43] visualization of feature maps F_{rgb}. RGB images are shown in the first row. The second row shows the CAM of "RGB-D Global & Local($\mathcal{L}_C$)", the third row shows the CAM of "RGB-D Global & Local($\mathcal{L}_{\text{VI}}$)" and the fourth row shows the CAM of "Full DLFS".

$\mathcal{L}_C$. This indicates that using $\mathcal{L}_C$ to learn local features can improve the scene classification performance. The main reason is that local deep grid features are complement to global scene features. The loss $\mathcal{L}_{\text{VI}}$ is also effective for improving the performance, as it can encourage the selected local features to be correlated with the scene class. The results indicate that both loss functions are useful for improving the DLFS module.

Moreover, we have visualized the class-specific activation map of the learned features with $\mathcal{L}_C$ loss, $\mathcal{L}_{\text{VI}}$ loss and both losses, i.e. "Full DLFS". From Fig. 7 we can see that the features of "Full DLFS" are more spatially-correlated.

C. SUN RGB-D Results

The comparison results on SUN RGB-D dataset are displayed in Table II. Among the compared methods, Song et al. [31], Liao et al. [44] and Zhu et al. [45] did not use local features. Wang et al. [23] and Song et al. [24] employed object detection for scene recognition.

Generally, the experimental results reveal that methods with local features can obtain better performance than those without local features. Different from existing methods, the proposed method selects multi-scale local features adaptively for scene recognition, and achieves even better performance.

We also summarize the multi-modality feature learning types in Table II. Feature-level fusion are commonly used multi-modality feature learning methods, while considering the correlation and distinction between different modalities can achieve better results than simply combining multi-modality features. Different from existing methods, the proposed method exploits local features to learn modal-correlated representations. Additionally, the proposed method exploits the spatial-distribution of multi-modality feature correlation to enhance the local feature mining process. Meanwhile, the learned local features are also encouraged to be more modality-correlated by the proposed loss. By using this mechanism cleverly, the proposed framework with DLFS module achieves state-of-the-art scene recognition result.

D. NYUD v2 Results

Since NYUD v2 dataset is relatively small and the training data in NYUD v2 dataset is heavily imbalanced, we use the weights pretrained on SUN RGB-D dataset for model initialization on NYUD v2 dataset. The comparison results on NYUD v2 dataset are displayed in Table III. Similar to SUN RGB-D dataset, from the results we can see that methods using local features can achieve better performance. Although Li et al.[33] did not use local features, they can still obtain competitive result by taking advantage of better multi-modality learning method. Compared with feature selection based method [34], this work can achieve better performance by the differentiable feature selection module and the effective training loss functions.

The main advantages of the proposed method is two-fold: 1) local features extracted by DLFS module is more effective than patch-sampling and object detection methods; 2) Global modality-distinctive and local modality-correlated features are jointly exploited in this work. To sum up, the comparison results on this dataset indicate the effectiveness of the proposed framework. Additionally, we have also done experiments to

TABLE III
EXPERIMENTAL RESULTS ON NYUD V2 DATASET

	Methods	Local Features	Mean-class Accuracy(%)
State-of-the-art Methods	Gupta et al. [49]	No	45.4
	Wang et al.[23]	CNN Proposals	63.9
	Li et al.[33]	No	65.4
	Song et al.[22]	Local Patches	65.8
	Du et al.[35]	No	66.5
	Song et al.[24]	Object Detection	66.9
	Song et al.[47]	Patches Sampling	67.5
	Xiong et al.[34]	Feature Selection	67.8
Proposed Method	ASK (K=16 & K=4)	Local Features Selection	**69.3**

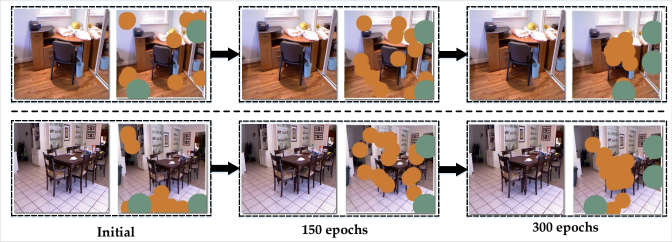


Fig. 8. Illustration of the selected multi-scale keypoints during the training stage. Semantic-meaningful local features are selected when the modal converges. Different colors represent different scales. (Best viewed in color.)

study the effect of selecting different number of local features. The results are displayed in Table IV. *'K=0'* denotes that no local features are used. Generally, the classification performance increases with more selected local features. However, when K is set to larger than 16, the performance decreases. When we select more features (*'K=36'*), the performance also decreases, which is only slightly better than *'K=0'*. The reason is that when more local features are selected, more noise will also be introduced, and the global features may be suppressed. Object or theme-level features will be affected by the irrelevant noisy features. The selected multi-scale keypoints are visualized in Fig. 8. Different colors represent different scales. It can be clearly seen that semantic local features are selected when the modal converges.

V. COMPARISONS WITH EXISTING METHODS

To show more details and give more comprehensive evaluations of the proposed work, we compare our work to other related existing methods and highlight the differences and similarities with them. The highlight of the proposed framework is that it selects multi-scale intermediate CNN features instead of sampling patches densely or using the extra object detection procedure. Several other works have also proposed unsupervised local-part localization methods. Here we compare our method with them and point out the main differences with them.

Spatial-related Multi-modal Feature Learning [34] found that similar spatial-attention maps are gained with attention mechanism for RGB and depth modalities, and enforcing similar spatial-attention map can boost the performance. This indicates that the same local objects are important features for both RGB and depth modalities. Inspired by this, we find that local objects contribute highly to the multi-modal correlation. Thus, we propose to maximize the correlation between RGB and depth local features, which provides more cues and supervision for local object-level feature selection.

KeypointNet [26] KeypointNet is an end-to-end geometric reasoning framework for learning latent category-specific 3D keypoints. KeypointNet can discover geometrically and semantically consistent keypoints adaptively with no extra annotations. KeypointNet stacks 13 layers of dilated convolutions to output $2N$ probability maps for predicting the $3D$ coordinates of N points. However, our work uses only one extra convolution layer with discriminative loss supervision for predicting probability maps, which is more time-efficient. Moreover, we use *stride* 2 convolution layer instead of dilated convolution to select the multi-scale CNN features, which is also more time-efficient. Additionally, a novel variational mutual-information maximization loss term is proposed in this work for discriminative training.

MA-CNN [27] MA-CNN consists of convolution, channel grouping and part classification sub-networks, which defines the parts as multiple attention areas. This work also employs the discriminative feature channels for part localization. However, the main difference is two-fold. 1) the local-part features are pooled from the attention areas in their work; 2) Multi-scale features are neglected. Different from their work, the proposed framework can learn to select multi-scale feature vectors in a totally differentiable manner.

Deep Regionlets [29] The architecture of deep regionlets includes a region selection network, which can learn more fine-grained features by selecting sub-regions adaptively. The local part feature selection is based on a spatial transformation and a gating network. STN [37] is employed to select local regions as local-part features. However, the proposed method aims to select multi-scale mid-level CNN feature vectors (keypoints in CNN feature maps) as the local discriminative representations.

Discriminative Filter Bank [28] This work also exploits mid-level CNN representations by learning a bank of convolutional filters that capture class-specific discriminative patches without extra part or bounding box annotations. Specifically, this work uses Global Max Pooling (GMP) to select only one

TABLE IV
NUMBER OF SELECTED FEATURES ON NYUD v2 DATASET

Number of Selected Features (K)	K=0	K=1	K=3	K=9	**K=16**	K=25	K=36
Mean-class Accuracy (%)	65.4	66.2	67.1	67.9	**69.1**	67.7	66.1

local feature vector from the intermediate CNN feature maps, which is different from our method. Moreover, multi-scale local features are also not considered in this work.

Spatial Attention [50] To select important local features from mid-level CNN feature maps, spatial attention is the most commonly employed method. Although non-local networks [50] can be used to focus on important local regions of feature maps, the softly attended feature maps still contain irrelevant features. However, our hard selection based method can select the important local feature vectors and discard the irrelevant ones. Moreover, the proposed method can select multiple multi-scale local features, which are more discriminative and representative than the spatial attention method. The visualization of the attention maps and the selected local feature vectors of our method are displayed in Fig. 9. The upper rows show the results of spatial attention based method, and the lower rows show the results of the proposed method. As we can see from Fig. 9, the spatial attention maps mainly focus on one local object regions, while our method can select multi-scale object-level ('chair' or 'wash basin') and theme-level ('curtain' or 'floor') feature vectors.

As displayed in Fig. 10, it can be clearly seen that the correlation between RGB and depth modalities is highly spatially related. The correlation of local objects are higher than other positions. Inspired by this intriguing finding, we can maximize the correlation between selected local multi-modal features to locate the positions of local objects.

VI. CONCLUSION

In this work, we present a multi-modal global and local feature learning framework for RGB-D scene classification. The differentiable local feature selection (DLFS) module is proposed to select important local object and theme-level features adaptively for RGB-D scene images. A novel loss function is proposed to supervise the training of DLFS module. Discriminative local object and theme-level representations can be selected with DLFS module from the spatially-correlated multi-modal RGB-D features. We take advantage of the correlation between RGB and depth modalities to provide more cues for selecting local features. Additionally, we further enhance the DLFS module with the multi-scale feature pyramid to select object-level features of different scales. Evaluations on SUN RGB-D and NYU Depth version 2 (NYUD v2) datasets have shown the effectiveness of the proposed framework.

VII. ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Project under Grant 2020YFB 2103902, National Natural Science Foundation of China under Grant 61632018, 61825603, U1801262, and 61871470.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Yuan Yuan, Jie Fang, Xiaoqiang Lu, and Yachuang Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1779–1792, 2018.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[4] Yuan Yuan, Zhitong Xiong, and Qi Wang, "VSSA-NET: vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Processing*, vol. 28, no. 7, pp. 3423–3434, 2019.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] Qi Wang, Junyu Gao, and Xuelong Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, 2019.

[7] Yuan Yuan, Jie Fang, Xiaoqiang Lu, and Yachuang Feng, "Spatial structure preserving feature pyramid network for semantic image segmentation," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3, pp. 1–19, 2019.

[8] Bin Zhao, Xuelong Li, and Xiaoqiang Lu, "CAM-RNN: co-attention model based RNN for video captioning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, 2019.

[9] Junyu Gao, Ieee Please Verify Yuan Yuan, and Qi Wang, "Feature-aware adaptation and density alignment for crowd counting in video surveillance.," *IEEE Transactions on Cybernetics*, 2020.

[10] Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu, "Structured dictionary learning for abnormal event detection in crowded scenes," *Pattern Recognition*, vol. 73, pp. 99–110, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[12] Li Da, Li Lin, and Li Xiang, "Classification of remote sensing images based on densely connected convolutional networks," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[13] Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[14] Hyo Jin Kim and Jan-Michael Frahm, "Hierarchy of alternating specialists for scene recognition," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, 2018, pp. 471–488.

[15] Yuan Yuan, Jie Fang, Xiaoqiang Lu, and Yachuang Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1779–1792, 2019.

[16] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014, pp. 392–407.

[17] Donggeun Yoo, Sunggyun Park, Joon-Young Lee, and In-So Kweon, "Fisher kernel for deep neural activations," *CoRR*, vol. abs/1412.1628, 2014.

[18] Zhen Zuo, Gang Wang, Bing Shuai, Lifan Zhao, Qingxiong Yang, and Xudong Jiang, "Learning discriminative and shareable features for scene classification," in *ECCV*, 2014, pp. 552–568.

[19] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

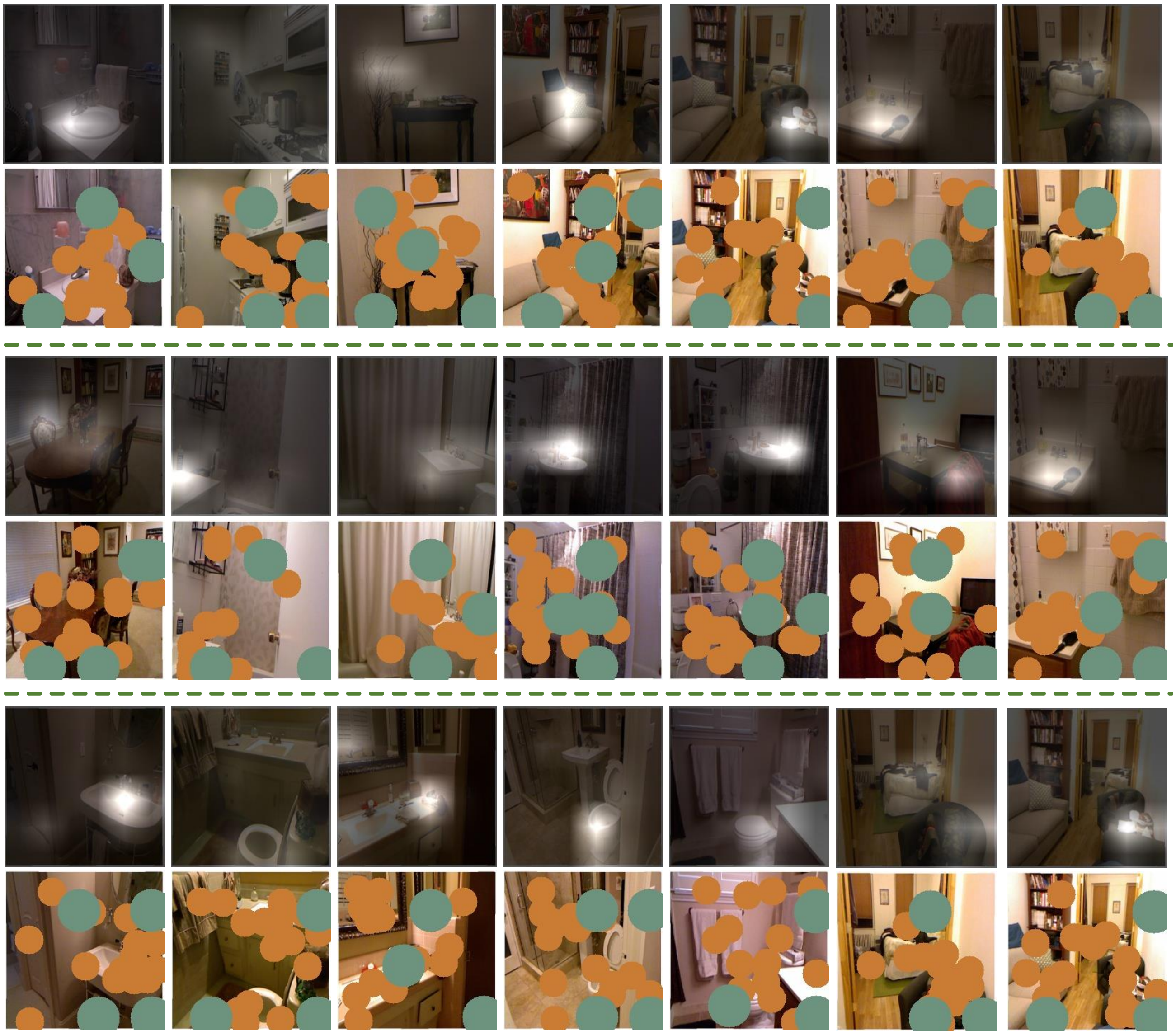


Fig. 9. The attention map visualization of spatial attention based method [50] and the selected keypoints visualization of the proposed method. The images of the first, third, fifth and seventh rows are the visualization for spatial attention based method. The second, fourth, sixth and eighth rows are the visualization of the selected keypoints of the proposed method. The upper and lower image pair is corresponding to the same image in NYUD v2 dataset [38].

[20] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[21] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos, "Scene classification with semantic fisher vectors," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 2974–2983.

[22] Xinhang Song, Luis Herranz, and Shuqiang Jiang, "Depth cnns for rgb-d scene recognition: Learning from scratch better than transferring from rgb-cnns.," in *AAAI*, 2017, pp. 4271–4277.

[23] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen 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.

[24] Xinhang Song, Chengpeng Chen, and Shuqiang Jiang, "Rgb-d scene recognition with object-to-object relation," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 600–608.

[25] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen, "Image representations with spatial object-to-object relations for RGB-D scene recognition," *IEEE Trans. Image Processing*, vol. 29, pp. 525–537, 2020.

[26] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi, "Discovery of latent 3d keypoints via end-to-end geometric reasoning," in *Advances in Neural Information Processing Systems*, 2018, pp. 2063–2074.

[27] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217.

[28] Yaming Wang, Vlad I Morariu, and Larry S Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4148–4157.

[29] Hongyu Xu, Xutao Lv, Xiaoyu Wang, Zhou Ren, Navaneeth Bodla, and Rama Chellappa, "Deep regionlets for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 798–814.

[30] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[31] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d:

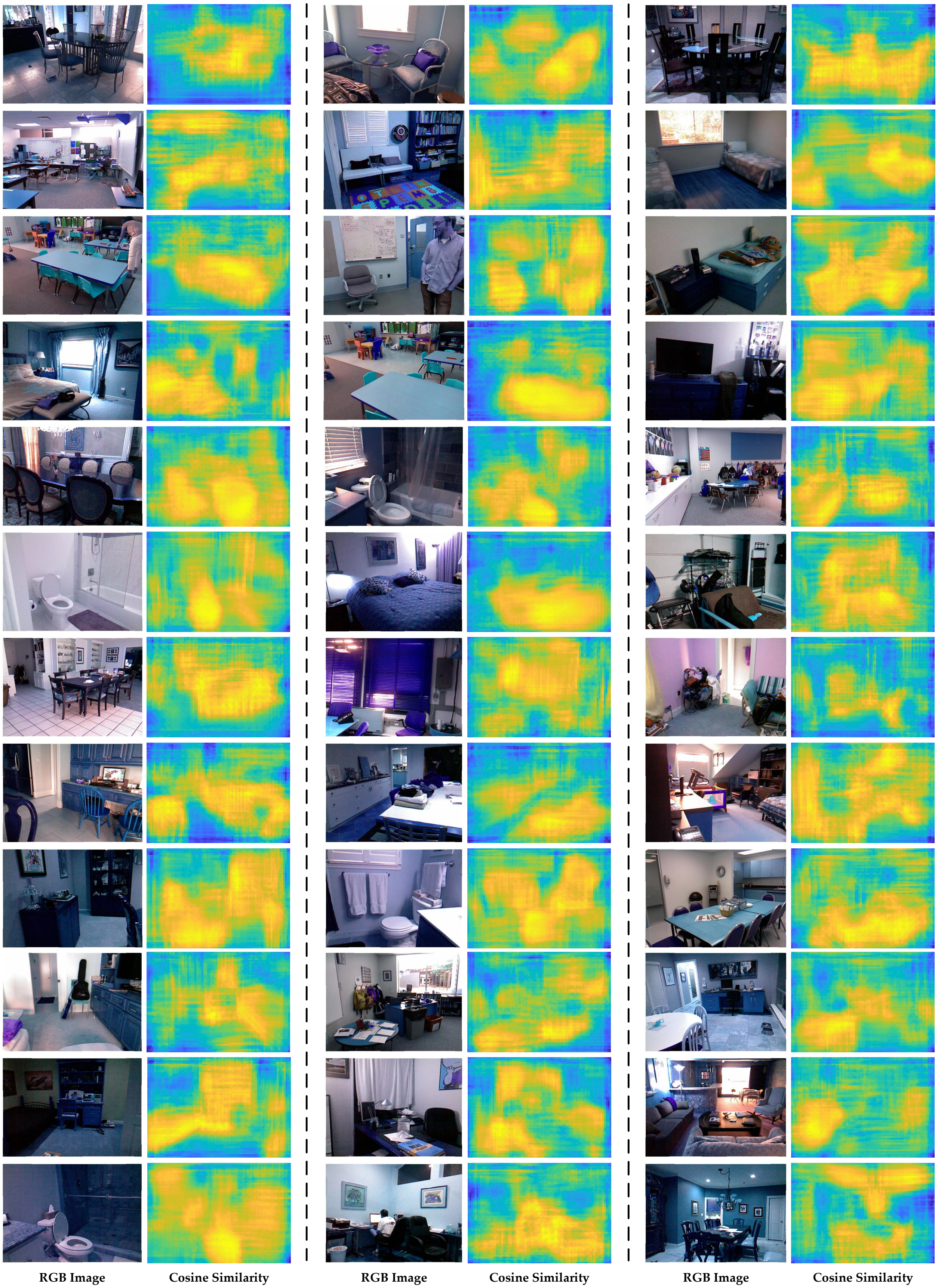


Fig. 10. Spatial cosine-similarity map visualization. Images of the second, fourth and sixth columns visualize cosine-similarity maps.

A rgb-d scene understanding benchmark suite," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[32] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen 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.

[33] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu 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.

[34] Zhitong Xiong, Yuan Yuan, and Qi Wang, "Rgb-d scene recognition via spatial-related multi-modal feature learning," *IEEE Access*, 2019.

[35] Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu, "Translate-to-recognize networks for rgb-d scene recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11836–11845.

[36] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*. Springer, 2014, pp. 345–360.

[37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 2017–2025.

[38] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[39] Saurabh Gupta, Pablo Arbelaez, and Jitendra 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.

[40] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[41] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[42] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 2921–2929.

[44] Yiyi Liao, Sarath Kodagoda, Yue Wang, Lei Shi, and Yong Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *ICRA*. IEEE, 2016, pp. 2318–2325.

[45] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu, "Discriminative multi-modal feature fusion for rgbd indoor scene recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2969–2976.

[46] Xinhang Song, Shuqiang Jiang, and Luis Herranz, "Combining models from multiple sources for rgb-d scene recognition," *IJCAI*, pp. 4523–4529, 2017.

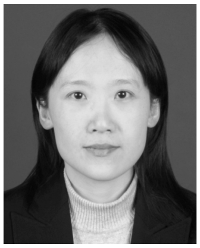
[47] Xinhang Song, Shuqiang Jiang, Luis Herranz, and Chengpeng Chen, "Learning effective rgb-d representations for scene recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 980–993, 2018.

[48] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen, "Image representations with spatial object-to-object relations for rgb-d scene recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 525–537, 2019.

[49] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[50] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

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, Xi'an, 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, Xi'an, 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 pattern 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, Xi'an, China. His research interests include computer vision and pattern recognition.