Scene Recognition With Prototype-Agnostic Scene Layout

Gongwei Chen, Xinhang Song, Haitao Zeng, and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*— Exploiting the spatial structure in scene images is a key research direction for scene recognition. Due to the large intra-class structural diversity, building and modeling flexible structural layout to adapt various image characteristics is a challenge. Existing structural modeling methods in scene recognition either focus on predefined grids or rely on learned prototypes, which all have limited representative ability. In this paper, we propose Prototype-agnostic Scene Layout (PaSL) construction method to build the spatial structure for each image without conforming to any prototype. Our PaSL can flexibly capture the diverse spatial characteristic of scene images and have considerable generalization capability. Given a PaSL, we build Layout Graph Network (LGN) where regions in PaSL are defined as nodes and two kinds of independent relations between regions are encoded as edges. The LGN aims to incorporate two topological structures (formed in spatial and semantic similarity dimensions) into image representations through graph convolution. Extensive experiments show that our approach achieves state-of-the-art results on widely recognized MIT67 and SUN397 datasets without multi-model or multi-scale fusion. Moreover, we also conduct the experiments on one of the largest scale datasets, Places365. The results demonstrate the proposed method can be well generalized and obtains competitive performance.**

***Index Terms*— Scene classification, convolution neural networks, graph neural networks, scene layout.**

I. Introduction

SCENE images (e.g., "classroom," "bedroom") are usually composed of specific semantic regions (e.g., "desk," "bed") distributed in certain spatial structures. Exploring the local regions and their spatial structures has been a long-standing research direction and plays a crucial role in scene recognition [1]–[3]. Due to the size and location changes of semantic regions (see Fig. 1), the spatial structures of images have great diversity, which makes it very difficult to represent them so as to adapt various image characteristics. Thus, how to build and model such structural layout into image representations is an obstacle problem.

Most existing methods [3]–[6] model spatial structural information based on predefined grid regions or densely sampled regions. These regions are in fixed sizes located in a grid, forming a simple and constant structure as a common prototype for all images, which results in rigid layout even with the extension of multi-scale setting. Some earlier works [1], [2], [7] have attempted to learn several prototypes for each category images with different models, such as constellation model in [1], Deformable Part-based Model (DPM) in [7] and DPM's variant in [2]. These prototypes can be regarded as templates with fixed topological structures for each scene category, where the geometric relations of the components are obtained through statistic learning. The spatial structure of each image is constructed by conforming to the prototypes. Although more than one prototype is usually used to characterize one scene category, such limited variety is not comprehensive enough to cover the large intra-class structural diversity of scene images. In contrast, our motivation is to design a layout modeling framework to flexibly capture the unconstrained spatial structures and effectively obtain discriminative patterns from them.

In this paper, we propose Prototype-agnostic Scene Layout (PaSL) construction method, which builds spatial structure for each single image without conforming to any prototype. Given an image, PaSL is constructed with the locations and sizes of discriminative semantic regions, which are detected by only using the convolutional activation maps of this image. Thus, PaSLs will vary from image to image and can flexibly express different spatial characteristics of the images. Considering the natural property of the graph to preserve diverse and free topological structures, we frame the structural modeling process as a graph representation learning problem. More specifically, we propose Layout Graph Network (LGN) where regions in PaSL are defined as nodes and two kinds of relations between nodes are encoded as edges. Through the graph convolution [8] and mapping operations of LGN, the topological structure and region representations can be transformed into a discriminative image representation.

Manuscript received September 5, 2019; revised February 14, 2020; accepted March 28, 2020. Date of publication April 14, 2020; date of current version April 28, 2020. This work was supported in part by the National Key Research and Development Project of New Generation Artificial Intelligence of China under Grant 2018AAA0102500, in part by the National Natural Science Foundation of China under Grant 61532018 and Grant 61902378, in part by the Beijing Natural Science Foundation under Grant L182054 and Grant Z190020, in part by the National Program for Special Support of Eminent Professionals and National Program for Support of Top-Notch Young Professionals, in part by the Lenovo Outstanding Young Scientists Program, and in part by the National Postdoctoral Program for Innovative Talents under Grant BX201700255. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Liang Wang. *(Corresponding author: Shuqiang Jiang.)*

Gongwei Chen, Xinhang Song, and Shuqiang Jiang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: gongwei.chen@vipl.ict.ac.cn; xinhang.song@vipl.ict.ac.cn; sqjiang@ict.ac.cn).

Haitao Zeng is with the School of Mechanical, Electronic and Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China, and also with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: haitao.zeng@vipl.ict.ac.cn).

Digital Object Identifier 10.1109/TIP.2020.2986599

Fig. 1. Image examples from MIT67 dataset. Here shows the images from three scene categories ("bathroom," "bedroom," and "classroom"). It can be seen that the objects in each scene can vary greatly in size and location, like "bathtub" in scene "bathroom," "bed" in scene "bedroom," "desk" in scene "classroom."

The main idea of PaSL construction method is inspired by the ability of pretrained CNNs to localize the meaningful semantic parts [9]. We make use of the convolution activation maps extracted from pretrained CNNs to detect semantic regions, and aggregate them to generate discriminative regions and form PaSL in an unsupervised way. The advantages of our method is two folds. One is that the whole process is performed on each image independently and can be easily extended to large scale datasets. Another is that PaSLs derived from different pretrained CNNs can yield comparable performances with same LGN, which demonstrates they have considerable generalization ability. Besides constructing PaSL, modeling it in graph structure is also an important contribution in this paper. Conventional structural models in scene recognition either have difficulty of optimization [2], [7] on large scale datasets or simplify the structural information [3], [10]. In contrast, we build Layout Graph Network upon PaSL by reorganizing it as a layout graph containing two subgraphs. These two subgraphs aim to capture different kinds of relations, spatial and semantic similarity relations between regions, respectively. Thanks to the independence between these two kinds of relations, we can explore structural information in a higher order space and easily encode it into more discriminative features. Furthermore, the application of graph convolution makes our model effectively handle various topological structures and easy to be optimized with large amounts of data.

We evaluate our model on three widely recognized scene datasets, MIT67 [1], SUN397 [11], and Places365 [12]. The ablation study shows that our method obtains up to 5% improvements over baselines that neglect structural information. Compared to current works on MIT67 and SUN397 that benefit from multi-model or multi-scale fusion methods, our model even outperforms them and obtains state-of-the-art results with single model in single scale. When extending our model to one of the largest scale dataset, Places365, it still shows competitive performance.

II. Related Work

A. Scene Recognition

Scene recognition is an important and challenging problem and has been extended to diverse research directions, like natural scene image recognition [13], RGB-D scene image recognition [14], and dynamic scene video recognition [15]. In this section, we mainly focus on natural scene image recognition methods and discuss their differences.

In early works, handcrafted features (like SIFT [16], RHOG [17]) have been regarded as the fundamental component in image classification. Based on these features, bag-of-feature methods (like VLAD [18], Fisher Vector [19]) have demonstrated great power on scene recognition. However, these methods incorporate local information in an orderless way, which loses spatial dependencies between local regions. Then, some works further explore the spatial contextual dependencies within bag-of-feature methods. Lazebnik *et al.* [20] proposed Spatial Pyramid Matching to exploit approximate spatial information in a predefined grid. Parizi *et al.* [21] used a reconfigurable model operated on a grid to capture the spatial information among regions.

Beyond this simple and fixed spatial information based on grids, some works explore the complex and flexible spatial structures formed by scene components in different ways. These works [1], [2], [7] construct the scene structures in a similar way to Deformable Parts Model (DPM) [7], DPM's variant [2], or a constellation model [1]. Based on these structural models, a fixed number of structures for each scene category, which can be named as scene prototypes, are learned. Then the spatial structural information of each image is discovered by conforming to the default structure of the most matched scene prototype. The spatial structures derived from scene prototypes have difficulty covering the intra-class variety of scenes. Differently, our approach can model the structural layout for each image following its own characteristics.

Recently, Deep Learning methods, especially Convolution Neural Networks, have been widely used in scene recognition. Some works [6], [22]–[24] combine bag-of-feature methods (like VLAD, Fisher Vector) or dictionary-based method with CNN to explore discriminative local information in an orderless way. To model spatial contextual dependencies, the works [3], [10] learn a sequential model (like LSTM [3]) or a graphical model (MRF [10]) on fixed size regions. Furthermore, the multi-scale strategy is adopted to capture more precise local information. However, these works either encounter the problem of noise regions caused by predefined grids, or simplify the spatial structural information, while our method can explore the complex spatial structural layouts and reform them in graphs to generate discriminative representations.

B. Discriminative Region Discovery

To discover the discriminative regions has been a long-standing study in visual recognition. Singh *et al.* [25] use an iterative optimization procedure to alternately cluster and train discriminative classifier on densely sampled patches. Juneja *et al.* [26] first propose an initial set of regions based on low-level segmentation cues, and then learn detectors on top

of these regions. However, these works all use the handcrafted features as region representations.

Benefited from Deep Learning methods and large scale datasets, there has been significant progress in image classification [13], [27], object detection [28], semantic segmentation [29], and many other vision tasks [30]–[32]. Some researchers try to use object detection or semantic segmentation models for localizing discriminative regions (like objects or stuff). Wang *et al.* [33] use an existing object proposal extractor to obtain region proposals, and apply Fisher Vector methods to encode these region features. López-Cifuentes *et al.* [34] propose to use semantic segmentation model to extract semantic regions, and encode these region representations with attention mechanism to enhance the image feature maps. The main issue of this kind of works is that extra models trained on different tasks are required and need other expensive label annotations.

Recently, Some works take advantage of CNN activations as region descriptors for discriminative region discovery. Wu *et al.* [4] obtain region proposals by performing MCG, and screen the regions by using one-class SVM and RIM clustering. Cheng *et al.* [35] sample a set of local patches in a uniform grid with their object scores extracted from ImageNet-CNN, then discard the patches containing non-discriminative objects by applying Bayes rules. One common characteristic of these works is that they generate the candidate regions independently of the CNN classifiers, which will incur much additional computational cost.

Besides these aforementioned approaches, some recent works explore the convolutional responses from CNNs to directly discover discriminative regions for fine-grained object recognition. Zheng *et al.* [36] group the convolutional channels to localize object parts in the well constrained spatial configurations. Zhang *et al.* [37] propose to integrate LSTM to CNN for extracting local age-sensitive regions. Wei *et al.* [38] use a simple thresholding method to discover object parts and select the largest component to represent the desired foreground object. In contrast, we formulate the discovery procedure for scene recognition, where more complex semantic regions and unconstrained spatial structures exist. Similarly, the work of [5] also uses a pretrained CNN classifier to generate discriminative regions for scene images. However, it needs extra scene category cue for each image and the CNNs with a specific architecture.

C. Graph Neural Networks in Computer Vision

Graph Neural Networks (GNNs) are designed to deal with the graph structured data, which were first proposed in [39]. Recently, some variants have been applied in program verification [40], molecular property prediction [41], document classification [8] and made significant progress. Inspired by the success of GNNs on graph structured data, some researches apply them in computer vision task, like multi-label classification [42], situation recognition [43], scene graph generation [44], zero-shot recognition [45], and etc. These works apply GNNs to natural graph data (like knowledge graph [42], [43], [45]), or constructed graph data with the supervision of annotated object regions (like scene graph [44]). In contrast to them, we perform GCN [8], a variant of GNN, on the structural layouts in scene images without external knowledge or object annotations.

III. Our Approach

In this section, we first introduce how to construct Prototype-agnostic Scene Layout (PaSL) from pretrained CNNs in an unsupervised way. Then we build Layout Graph Network upon PaSL to integrate structural information into visual representations. The whole process is shown in Fig. 2. In the following, we will go into details about our approach.

A. Prototype-Agnostic Scene Layout Construction

PaSL is constructed by the locations and sizes of discriminative regions (including objects, object-parts, and other visual patterns) in each image. To form PaSL, we first need to discover discriminative regions. Unlike previous works that use many selected image patches (from manual annotation [2] or region proposal [4]) to train region detectors, we only need the convolutional units from a pretrained CNN, without detector training.

Recently, Zhou *et al.* [46] have shown the convolutional units from a CNN pretrained on Places [13] dataset can be used as object detectors. And Bau *et al.* [9] extend this conclusion to more pretrained CNNs and more visual concepts. They demonstrated the individual convolutional units in CNN can be aligned with semantic concepts across a range of objects, parts, textures, scenes, materials, and colors. Inspired by these works, we utilize the convolutional units in pretrained CNNs as region detectors. In practice, given an image, we feed it into a pretrained CNN to extract the convolutional activation maps $\mathcal{A}$ ($\mathcal{A} \in R^{H\times W\times C}$) from the last convolutional layer (For VGG16, max pooling need to be employed). The c-th activation map in $\mathcal{A}$ is represented as $\mathcal{A}_c \in \mathbb{R}^{H\times W}$, while $c \in \{1, \ldots, C\}$. For instance, if the resolution of the input image is 224×224, we obtain $7 \times 7 \times 512$ activation maps as $\mathcal{A}$, where $H = W = 7$ and $C = 512$, by adopting a pretrained VGG16 model.

Based on the same assumption of [9], [46] that the desired regions (e.g., semantic regions) in feature maps have high response values, we propose an adaptive threshold T in Eq.1 to detect the candidates of discriminative regions.

$$T = \frac{1}{C}\sum_{c=1}^{C}\dot{\mathcal{A}}_c, \quad \dot{\mathcal{A}}_c = \max\left(\mathcal{A}_c\right) \tag{1}$$

For efficient computing, any activation map whose maximum value is under T is discarded, then a subset $\widetilde{A}$ of activation maps $\mathcal{A}$ is produced. Each activation map in $\widetilde{A}$ is scaled up to the input image resolution and then thresholded into a binary map B by using the threshold T. We take the connected components in B as the candidates of discriminative regions. The algorithm from [47] is adopted to generate bounding boxes of the connected components in each binary map. By performing the same operations on all activation maps in $\widetilde{A}$, we obtain bounding box set M

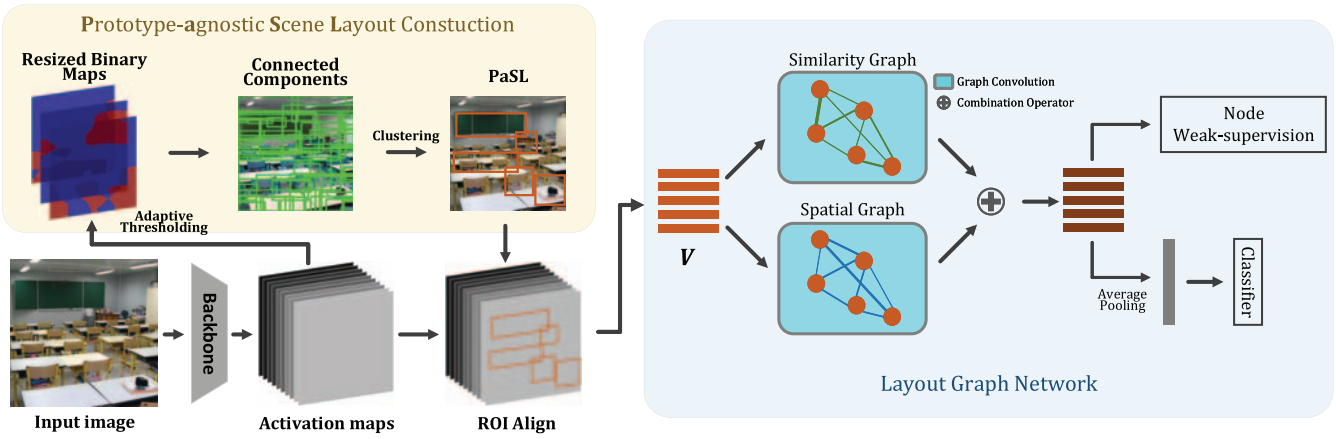


Fig. 2. Overview of our approach. A pretrained CNN model (like VGG16 or ResNet50) is used as the backbone to extract the candidates of discriminative regions. The desired discriminative regions are clustered from these candidates, and fed into ROI Align Layer to generate the node representations. Two subgraphs are constructed by treating regions as nodes and designing spatial or similarity relation as edge. Then, we perform graph convolution on two subgraphs and combine them to obtain the final node representations. Finally, node weak-supervision mechanism makes each node predict global image category by feeding it into a fully-connected layer. Meanwhile, the averagely pooled node representations, regarded as the global image representations, are exploited for scene recognition by a fully-connected layer.

of all candidates of discriminative regions. The element m in M is composed of the left-up and right-bottom coordinates, e.g., $m = \{x_{min}, y_{min}, x_{max}, y_{max}\}$, where (x_{min}, y_{min}) denotes the coordinate of left-up point in bounding box, and (x_{max}, y_{max}) is the coordinate of right-bottom point.

In practice, the number of elements in M is large, e.g., $\sim$ 500 for VGG16 and $\sim$ 1500 for ResNet50. If we construct PaSL with all regions from M, it will cause expensive computational cost in the later process. Meanwhile, the regions from M have two characteristics. One is that although adaptive thresholding can discard some small noise parts, there also have several wrong detected results imposed by the unsupervised process. Another one is that discovering from each activation map independently may bring many visually similar regions. In order to avoid the wrong or similar regions, we choose a simple yet effective way, e.g., clustering, to find the most representative regions in M as the desired discriminative regions. Accordingly, the discriminative regions D could be obtained by:

$$L = \mathbb{C}(M, N) \tag{2}$$

$$D = Aggre(M, L) \tag{3}$$

where $\mathbb{C}(\cdot)$ denotes hierarchical clustering method. N stands for the number of clusters, which also means the number of discriminative regions. L corresponds to the cluster labels of the elements in M. Given cluster labels L, we perform mean pooling method ($Aggre$) on bounding boxes of elements in the same cluster to obtain bounding boxes of discriminative regions $D = \{d_1, \ldots, d_N\}$. Specifically, clustering method can be k-means or spectral clustering. If choosing k-means, cluster centers can be directly used as discriminative regions.

Given discriminative regions, we define Prototype-agnostic Scene Layout (PaSL) as a collection of the locations and sizes of these regions in each image. The spatial structure, that is implicit in PaSL, requires to be represented in a certain form. To form the diverse and free topological structure of PaSL of each image, the graph is adopted as data structure. Following the common setting of graph structured data, we define the discriminative regions as nodes and encode two kinds of independent relations between regions as edges. The details will be described in the following section.

B. Layout Graph Network

For modeling the spatial structure of PaSL, we reorganize it as a layout graph, which is better for incorporating the structural information into visual representations. Given PaSL with discriminative regions, a layout graph $\mathcal{G} = \{V, A^{sp}, A^{sim}\}$ is constructed, which contains a node set V, and two adjacency matrices A^{sp}, A^{sim}. For clarity, we decompose the layout graph into two subgraphs with the same nodes but different adjacency matrices: spatial subgraph $\mathcal{G}^{sp} = \{V, A^{sp}\}$ and similarity subgraph $\mathcal{G}^{sim} = \{V, A^{sim}\}$. More specifically, these two subgraphs share the same node set $V = \{v_1, \ldots, v_N\}$, where v_i corresponds to the representation of discriminative region i. We apply RoIAlign [48] to extract the representation of each region from a pretrained CNN as the initial state vector of v_i. This pretrained CNN can be regarded as a feature extraction model, which is the same as the pretrained model for generating PaSL, unless otherwise stated.

1) Spatial Subgraph: The spatial information is vital in PaSL, because it implies the functions or properties of regions. One way to take advantage of this information is to exploit the spatial relations between regions. Specifically, we define a kind of spatial representation to encode this relation and then generate the spatial edge to form the adjacency matrix. As mentioned above, each discriminative region i has a bounding box $d_i = \{x_{min}, y_{min}, x_{max}, y_{max}\}$. Inspired by [49], we extract the spatial feature of each region as follows:

$$d_i^{sp} = \left[\frac{x_{min}}{W_{img}}, \frac{y_{min}}{H_{img}}, \frac{x_{max}}{W_{img}}, \frac{y_{max}}{H_{img}}, \frac{Ar}{Ar_{img}}\right] \tag{4}$$

where Ar and Ar_{img} are the areas of the region i and the image respectively. W_{img} and H_{img} denotes the width and

height of the image. We concatenate d_i^{sp} and d_j^{sp} to obtain the spatial representation d_{ij}^{sp} of spatial relation between the regions i and j. After generating the spatial representation, we employ an edge function F_e, implemented as one-layer fully connected network, to generate the spatial edge a_{ij}^{sp} as:

$$e_{ij}^{sp} = F_e(d_{ij}^{sp}) \tag{5}$$

$$a_{ij}^{sp} = \frac{\exp(e_{ij}^{sp})}{\sum_{j \neq i} \exp(e_{ij}^{sp})} \tag{6}$$

Then the spatial adjacency matrix A^{sp} is obtained to form spatial subgraph $\mathcal{G}^{sp}$. The diagonal values in A^{sp} are zero.

2) Similarity Subgraph: To explore the spatial information in PaSL is an obvious requirement. But there exists an problem in the spatial subgraph that the spatial relation overlooks the semantic meanings of regions. To address this problem, we propose the similarity subgraph as a complement to the spatial subgraph. Due to the lack of explicit labels for the local regions, we take the region representation as a substitution for the semantic label. Then, we model the similarity between these region representations to capture the semantic similarity relations between regions.

Given the node set V, we can obtain the state vector $v_i \in \mathbb{R}^h$ of each node. In similarity subgraph, we aim to obtain the strong connection between semantic similar regions. So the semantic similarity relations between regions are measured by the cosine similarity, which is defined as follows:

$$e_{ij}^{sim} = \phi(v_i)^T \phi(v_j), \quad \phi(v) = \frac{\omega^{sim} v}{\left|\left|\omega^{sim} v\right|\right|_2} \tag{7}$$

where ϕ represents the transformation of the state vector and following ℓ_2 normalization, $\omega^{sim} \in \mathbb{R}^{h/2 \times h}$ is the transformation weights. The dot product e_{ij}^{sim} of two ℓ_2 normalized vector denotes the cosine similarity between regions. To balance the impact of neighbor nodes, we perform the softmax function on each row of the cosine similarity matrix as:

$$a_{ij}^{sim} = \frac{\exp(e_{ij}^{sim})}{\sum_{j \neq i} \exp(e_{ij}^{sim})} \tag{8}$$

where A^{sim} is used as the adjacency matrix for similarity subgraph. The diagonal values in A^{sim} are zero.

3) Graph Convolution: After building the layout graph, the next step is to incorporate the spatial and semantic similarity information into the representations of regions, and generate discriminative image representations. Considering the superior performance of Graph Convolution Network (GCN) [8] on graph structured data, we adopt Graph Convolution (GC) on spatial and similarity subgraph, then combine two subgraphs. Given a graph $\mathcal{G} = \{V, A\}$, where $V \in \mathbb{R}^{N \times h}$ is the node set and $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix. One GC layer aims to combine the information of neighbor nodes and target node through relation edges to update the state vector of target node, which can be formulated as:

$$\begin{aligned} V^t &= \sigma(Z V^{t-1} \Omega_t) \\ Z &= \Lambda^{-1} \tilde{A} \\ \Lambda_{ii} &= \sum_j \tilde{A}_{ij}, \quad \tilde{A} = A + I_N \end{aligned} \tag{9}$$

where V^t is the updated state vectors of nodes in GC layer t, $\Omega_t \in \mathbb{R}^{h_{t-1} \times h_t}$ denotes weight matrix, and h_0 is the input vector dimension while h_t ($t > 0$) means hidden size of GC layer t. We utilize the non-linear function ReLU as σ.

4) Combination of Different Subgraphs: Now, we can employ graph convolution to generate the updated state vectors V_{sim}^t, V_{sp}^t for spatial, similarity subgraphs, with A^{sim}, A^{sp} obtained above, respectively. Then, we investigate how to effectively combine these two subgraphs. First, we define the combination of two subgraphs as:

$$V^t = V_{sim}^t \oplus V_{sp}^t \tag{10}$$

where $\oplus$ means the combination operator. Intuitively, we can combine the updated state vectors from two subgraphs by using element-wise *addition* or *maximum*. Beyond them, we also consider an alternative to improve the sparsity of combined representations, which is element-wise *product*. we conduct a comparison experiment in section IV-D2, which confirms that the element-wise *product* is a better choice to combine two subgraphs.

5) Global Information: PaSLs in most images cannot cover the whole areas of images, which may lose some useful information. So we decide to add global information into the layout graph. We define a global node that represents the whole image, and perform average pooling on the convolutional activation maps from the last convolutional layer to generate the initial state vector of the global node. As a result, the node set V will be $\{v_0, v_1, \ldots, v_N\}$, where v_0 denotes the global node. For spatial subgraph, We set the bounding box of global node as $d_0 = \{x_{min} = 0, y_{min} = 0, x_{max} = W_{img}, y_{max} = H_{img}\}$, where W_{img} and H_{img} denote the width and height of the whole image. The global node is connected to all local nodes, and we apply the same operations described above to obtain the new adjacency matrices A^{sp}, A^{sim}.

6) Output: To avoid overfitting, we only utilize one GC layer. We obtain the final state vectors $V^1 \in \mathbb{R}^{N \times h_1}$ from the GC layer and following ℓ_2 normalization as node representations. When only using local regions as nodes, we apply average pooling on node representations to generate the image representation as a h_1-dimensions vector. And if adding global node, we only treat the global node representation as the image representation. Besides, we have tried to averagely pool all global/local node representations to obtain the image representation, which hurts the performance. And we have also tried to concatenate the global node representation with averagely pooled local node representation to produce the image representation, while it has similar performance but needs more parameters in the later process. For scene recognition, we feed image representation into one layer fully connected network to predict the image category. And we utilize softmax function with cross entropy as the loss function to obtain the image classification loss l_g.

7) Node Weak-Supervision Mechanism: Specifically, we propose a node weak-supervision mechanism to improve the discriminative performance of each node (except global node). For the representation of each node, we force it to predict the scene category of image by using one layer fully-connected network in a weakly supervised way, which

can make the node representations more suitable for image recognition and produce the node classification loss l_n. We combine the two classification loss to form the total loss l as,

$$l = l_g + \lambda l_n \tag{11}$$

where λ is a hyperparameter. Specifically, this branch is only used in the training process.

IV. Experiments and Discussions

In this section, we evaluate our method on three widely recognized datasets, MIT67 [1], SUN397 [11], and Places365 [12].

MIT67 Dataset contains a total of 15620 images belonging to 67 indoor scene categories. Following the standard evaluation protocol, we use 80 images of each category for training and 20 images for testing. We report accuracy as evaluation metric.

SUN397 Dataset is a more challenge scene dataset, which contains 397 scene categories and 108,754 images. The dataset is divided into 10 train/test splits, each split consists of 50 training images and 50 test images per category. The average accuracy over splits is presented as evaluation metric.

Places365 Dataset is one of the largest scale scene-centric datasets, which has two training subsets, Places365-standard and Places365-challenge. In this paper, we only choose Places365-standard as training set, which consists of around 1.8 million training images and 365 scene categories. The validation set of Places365 contains 100 images per category and the testing set has 900 images per category. We report experimental results on its validation set, because its test set has no available ground truth. Both top1 and top5 accuracy are reported as evaluation metric.

A. Implementation Details

Our model can be implemented with different pretrained models as backbone CNNs. For fair comparison with other methods, we adopt three pretrained models, which are VGG-IN, VGG-PL205, ResNet-PL365. VGG-IN, VGG-PL205 are the VGG16 models pretrained on ImageNet dataset [50] and Places205 dataset [13] respectively, ResNet-P365 is the ResNet50 model pretrained on Places365 dataset [12]. To construct PaSL, we extract the convolutional activation maps from the last convolutional layer (max-pooled in VGG16). Inspired by [51], we fix the input image resolution as 448×448 for VGG16-IN and ResNet50-PL365, 352×352 for VGG16-PL205, which leads to $14 \times 14 \times 512$, $14 \times 14 \times 2048$ and $11 \times 11 \times 512$ activation maps respectively. The number of clusters N, the hidden size h_1 and the λ are set to $\{32, 8192, 4.0\}$ for LGN with backbone VGG-IN and VGG-PL205, $\{64, 8192, 1.0\}$ for ResNet-PL365.

The initial state vectors of nodes are normalized with two normalization function (Layer Normalization [52], ℓ_2 Normalization), then fed into LGN. Specifically, the Layer Normalization is not trained in our experiments. We train LGN using Adam [53] with an initial learning rate of 10^{-3} (decayed by a factor of 0.1 at 10/15/18th epoch), a batch size of 32 and weight decay of 10^{-5}. All parameters are randomly initialized following Xavier initialization method [54]. We use the model trained at 20th epoch as the final model in all experiments. Dropout is only applied on the output prediction layer with a ratio of 0.2. The ℓ_2-norm of gradients is clipped to a maximum value of 0.25. All experiments are conducted on a single NVIDIA 1080 Ti GPU by using open-sourced framework Tensorflow.

TABLE I

Comparison of LGN With Previous Methods Based on Single VGG16 Model in Single Scale. Classification Accuracy (%) Is Reported as Evaluation Metric on MIT67 and SUN397 Datasets. The Best Result of Each Column Is Marked in Bold

Method	P.S.	I.R.	MIT67	SUN397
Three [51]	IN	643	76.42	59.71
MFA-FS [55]	IN	512	79.57	61.71
HSCFVC [56]	IN	512	79.5	-
MFAFVNet [23]	IN	512	80.3	62.51
CNN-DL [6]	IN	-	78.33	60.23
LSO-VLADNet [24]	IN	448	81.7	61.6
Three [51]	PL205	256	80.90	66.23
S-HunA [57]	PL205	-	83.7	-
SpecNet [58]	PL205	256	84.3	67.6
CNN-DL [6]	PL205	-	82.86	67.90
LGN	IN	448	79.78	62.03
	PL205	352	**85.37**	**69.48**

P.S. : "Pretrain dataset", I.S. : "Input Resolution"
IN : "ImageNet", PL205 : "Places205"

B. Experimental Results

In this subsection, we first report the performances on MIT67, SUN397. These two datasets are the most popular benchmark for evaluating scene recognition methods. Thus, we can provide the comprehensive and detailed comparison with existing works about scene recognition. Meanwhile, we also conduct experiments on one of the largest scale scene dataset, Places365, to demonstrate the generalization of our model.

1) Comparison on Single Model in Single Scale (MIT67 and SUN397): Most existing scene recognition methods obtain their best performances based on multi-model or multi-scale fusion. However, to perform the fusion needs more computational time and memory usage, which will cause expensive cost. The idea of multi-scale representation is presented to alleviate the problem of the various sizes of the semantic components in scene images. Benefiting from the flexible structure of PaSL, our model can efficiently capture the different locations and sizes of semantic components, to produce the better image representations for scene recognition. To prove it, we compare the previous works with our model on single VGG16 model in single scale in Table I. The two pretrained VGG16 models, pretrained on ImageNet (VGG-IN) and Places205 (VGG-PL205) are adopted as backbone CNN models in the comparison. The backbone VGG-PL205 show impressive performance on MIT67 and SUN397, generally outperforming the VGG-IN. Compared to existing works using the same VGG-PL205 backbone, our model obtains better performance with a clear margin ($1-2\%$). While based on the VGG-IN, the LGN surpasses the most previous works, except

TABLE II
COMPARISON OF LGN WITH STATE-OF-THE-ART WORKS ON MIT67 AND SUN397. CLASSIFICATION ACCURACY (%) IS REPORTED AS EVALUATION METRIC. THE BEST RESULT OF EACH COLUMN IS MARKED IN BOLD

Method	# of scales	# of models	MIT67	SUN397
CLDL [61]	1	1	84.69	70.40
MFA-FS [55]	4	4	87.23	71.06
MFAFVNet [23]	4	4	87.97	72.01
CNN-DL [6]	2	2	86.43	70.13
SDO [35]	10	2	86.76	73.41
PowerNorm [59]	1	3	86.3	-
fgFV [62]	10	2	87.60	-
MFAFSNet [60]	4	4	**88.06**	73.35
Adi-Red [5]	3	2	-	73.59
LGN (ResNet)	1	1	**88.06**	**74.06**

MFAFVNet and LSO-VLADNet. The lower performance of VGG-IN can be concluded into two possible reasons: 1) these two previous works report better accuracy benefiting from the refinement of low level convolutional features. 2) The PaSL derived from VGG-IN has less power for capturing the spatial structure in scene images, which is verified in subsection IV-C3.

2) Comparison With the State-of-the-Art Works (MIT67 and SUN397): Table II presents the results of our best model and state-of-the-art works. Our best model is based on ResNet-PL365 pretrained model in single scale setting. Compared to the methods [5], [59], [60] based on the same pretrained model, our model achieves the best performance. Most importantly, the work [5] utilizes the similar technique to extract discriminative regions and even multi-scale regions to generate the image representations. However, it ignores the relations (either spatial or similarity relations) between local regions, leading to an inferior performance. This confirms that the relations between local regions are useful for scene recognition, and our LGN can take advantage of them. We also report state-of-the-art works that involve various combination techniques to achieve better performance. Even though these works contain multi-scale information [5], [6], [23], [35], [55], [61], [62] or multi-model combination [23], [35], [55], [60], our model still outperforms them and achieves the state-of-the-art performance for scene recognition, to the best of our knowledge.

3) Experimental Results on Places365: To make more convincing results, we report the result of our best model on Places365 in Table III. The experimental setting is same as above, except the input resolution changes to 224×224 and the number of clusters changes to 32. Compared to the baseline Places365-ResNet [12], our model can gain 1.76% improvement of Top1 accuracy, which demonstrates the effectiveness of the proposed PaSL and LGN. It is worthy to note that the proposed LGN can outperform previous works with single model in single scale, although they report better results obtained by multi-model or multi-scale combination.

C. Analysis of PaSL

We provide a deep analysis of PaSL based on MIT67, and discuss its properties.

TABLE III
CLASSIFICATION ACCURACY (%) ON PLACES365 VALIDATION SET

Method	Top1 acc.	Top5 acc.
Places365-VGG [12]	55.24	84.91
Places365-ResNet [12]	54.74	85.08
Deeper BN-Inception [63]	56.00	86.00
CNN-SMN [10]	54.3	-
LGN (ResNet)	56.50	86.24
Multi-Model CNN-SMN [10]	57.1	-
Multi-Resolution CNNs [63]	58.30	87.60

1) The Visualization of PaSL: Fig. 3 show the images with PaSL derived from the backbone VGG-PL205. All the images are plotted with 32 bounding boxes of regions in PaSL. To avoid an unclear display, we firstly sort all the local regions in PaSL, and then emphasize the top 8 regions in the yellow and thick rectangles and downplay other regions in the red and thin rectangles, when plotting PaSL on an image. Specifically, we choose the edge values of all local regions connected to the global node in two adjacency matrices for sorting these regions. In Fig. 3, the left 3 columns show the regions emphasized by similarity edges, and the right 3 columns show the regions emphasized by spatial edges in same images. It's easy to see that the regions in PaSL can vary greatly in size and location to suit the large diversity of structural layouts in scene images. Importantly, PaSL can localize some semantic regions specified for the corresponding scenes, like "liquor cabinet" in "bar," "bed" in "bed room," "meeting table" in "meeting room," and so on. When comparing the regions emphasized by spatial and similarity edges, the obvious difference is that the regions emphasized by spatial edges tend to focus on the aggregated semantic components (like a lots of chairs), and the regions emphasized by similarity edges usually concentrate on the contents similar in visual details (like texture of parts of floor or wall). This difference demonstrates that two subgraphs can explore the local information in different aspects and be complementary to each other.

2) The Difference of PaSL: Although each image has its own spatial structure, PaSLs derived from the same pretrained model will have some similar properties. From the point of view of PaSLs in the whole training data, we define a metric named Coverage Ratio, which is the ratio between the coverage area of PaSL and the area of the image, to analyze the properties of PaSLs. In Fig. 4, the boxplots show the distributions of Coverage Ratio for all training image PaSLs derived from three different pretrained models. Note that the number of regions in PaSL is fixed to 32 for a fair comparison. We find that PaSLs derived from models pretrained on scene-centric datasets (Places205 or Places365) focus on larger regions compared to them derived from the model pretrained on object-centric dataset (ImageNet). And also PaSLs derived from the model pretrained on ImageNet may focus on the regions with high objectness. So, the values of their Coverage Ratio have a larger diversity due to the wide variety of size and location of objects in scene images.

3) The Generalization of PaSL: Considering the independence between PaSL construction and LGN, we can explore the generalization of PaSL by combining PaSL with LGN when they are based on different or same pretrained models.

Fig. 3. Visualization of PaSL. We choose some examples with PaSL from five scene categories. Each example shows the image with 32 rectangles representing the local regions in PaSL. For better visualization, we sort all 32 local regions and emphasize the top 8 regions in yellow and thick rectangles and downplay other regions in red and thin rectangles. The left 3 columns show the images with the regions emphasized by similarity edges. The right 3 columns show the same images with the regions emphasized by spatial edges.

There are three kinds of PaSLs derived from different pretrained models, and three kinds of LGNs based on different pretrained models. Therefore, we conduct combination experiments on MIT67, and report nine combination results in Table IV. When the pretrained models are different for PaSL and LGN, the performance can yield a change of no more than 1.04%. Besides different combination of PaSL and LGN, we also evaluate another spatial layout formed by regions generated by Faster RCNN [28] pretrained on MSCOCO dataset. For a fair comparison, we set the number of regions in this layout to 32. Based on Table IV, we can have three observations. 1) Compared to PaSLs derived from other pretrained models, the one from VGG-PL205 has the better ability to represent the spatial structure of scene images. 2) Despite having some fluctuations in performance, PaSLs derived from different pretrained models have comparable value for scene recognition, which demonstrates their considerable generalization capability. 3) The spatial layout generated by object detection obtains the worst performances with all LGNs. One possible reason is that this layout mainly focus on some common objects, and is not suitable to capture the complex structural layouts of scene images.

D. Experimental Study of LGN

1) Configuration of Hyperparameters: Three hyperparameters are important to determine the performance of our method, the number of clusters N in constructing PaSL, the hidden size h_1 in graph convolution, and the λ in node weak-supervision mechanism. To investigate these three hyperparameters, we conduct several experiments on MIT67 dataset. Because the architectures of VGG16 and ResNet50 are different,

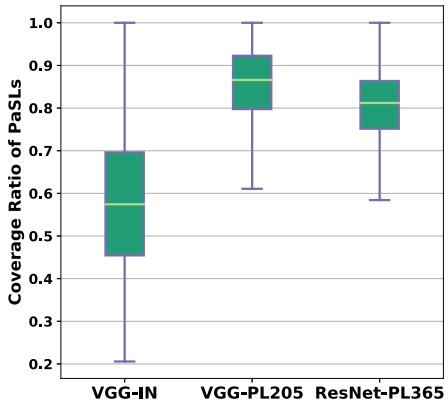
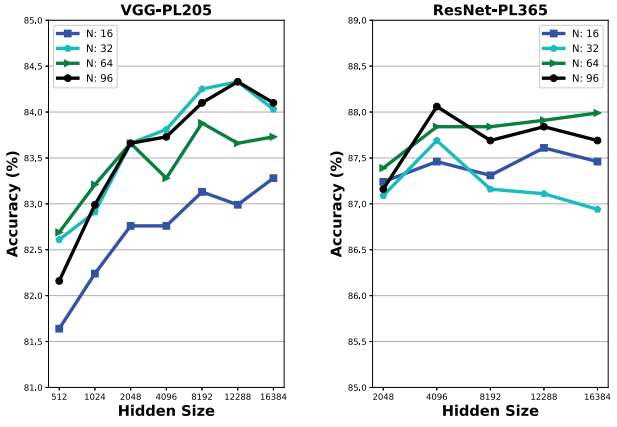


Fig. 4. The distributions of Coverage Ratio of PaSLs derived from different pretrained models (VGG-IN, VGG-PL205, and ResNet-PL365). Coverage Ratio is the ratio between the coverage area of PaSL and the area of the image.

TABLE IV

THE GENERALIZATION OF PASLS DERIVED FROM DIFFERENT PRETRAINED MODELS. WE SHOW CLASSIFICATION ACCURACY (%) OF THE COMBINATIONS OF PASL AND LGN, WHEN THEY ARE BASED ON DIFFERENT OR SAME PRETRAINED MODELS. THE BEST RESULT OF EACH ROW IS MARKED IN BOLD

LGN	PaSL			Detection
	VGG-IN	VGG-PL205	ResNet-PL365	
VGG-IN	79.78	**80.52**	80.07	68.66
VGG-PL205	84.78	**85.37**	85.29	80.52
ResNet-PL365	88.21	**88.73**	87.69	83.66



(a) (b)

Fig. 5. The effect of hidden size h_1 and the number of clusters N. We report the results on spatial subgraph without global information and node weak-supervision. (a) and (b) show the classification results of our methods based on the pretrained models, VGG-PL205 and ResNet-PL365, separately.

especially the processes from the last convolutional layer to output prediction layer, we analyze these hyperparameters on VGG-PL205 and ResNet-PL365 pretrained models, separately. We do not show the analysis on VGG-IN, since it has a similar behavior with VGG-PL205.

TABLE V

THE INFLUENCE OF λ IN NODE WEAK-SUPERVISION MECHANISM. WE SHOW THE CLASSIFICATION ACCURACY (%) BASED ON SPATIAL SUBGRAPH WITHOUT GLOBAL INFORMATION. THE BEST RESULT OF EACH ROW IS MARKED IN BOLD

λ	0.0	1.0	2.0	3.0	4.0	5.0
VGG-PL205	84.25	84.18	84.18	84.32	**84.55**	84.40
ResNet-PL365	87.84	**87.91**	87.84	87.80	87.76	87.69

TABLE VI

A COMPARISON OF DIFFERENT SUBGRAPH COMBINATION METHODS WITHOUT GLOBAL INFORMATION. CLASSIFICATION ACCURACY (%) IS REPORTED AS EVALUATION METRIC. THE BEST RESULT OF EACH COLUMN IS MARKED IN BOLD

Method	VGG-PL205	ResNet-PL365
Addition	85.07	88.28
Maximum	84.93	88.13
Product	**85.22**	**88.36**

We evaluate the effect of hidden size h_1 and the number of clusters N on spatial subgraph without global information and node weak-supervision in Fig. 5. It can be observed that the trends of accuracy caused by hidden size h_1 are different with VGG-PL205 and ResNet-PL365. In Fig. 5 (a), the accuracy has a significant increment when hidden size h_1 is lower than 8192, and then tends to be stable as hidden size h_1 increases. However, in Fig. 5 (b), we can see that the accuracy has a slight change as hidden size h_1 changes. These differences can be attributed to the aggregation techniques for generating the global image representation in different CNNs. In VGG16, the local spatial features are concatenated to produce the global representations, while they are averagely pooled in ResNet. Thus, in LGN based on VGG-PL205, aggregating the local features need to substantially enlarge the projection dimension (hidden size h_1) to prevent the information loss from averagely pooling, but not for ResNet-PL365. For instance, the ratios of hidden size h_1 to input dimension h_0 are 16 and 4 for VGG-PL205 and ResNet-PL365, respectively.

As illustrated in Fig. 5, when the number of clusters $N = 32$, we obtain better performances by using VGG-PL205, and the similar observation can be found at $N = 64$ with ResNet-PL365. Thus, we set hidden size and the number of clusters $\{h_1, N\}$ to $\{8192, 32\}$ and $\{8192, 64\}$ for VGG-PL205 and ResNet-PL365 respectively in all subsequent experiments. Besides h_1 and N, the hyperparameter λ in node weak-supervision mechanism is also important. The node weak-supervision mechanism aims to force each local node to predict the image category, which makes local representations more specific for generating discriminative image representations. We report the results on spatial subgraph without global information for different values of λ in Table V. It can be observed that, the best performances are obtained at $\lambda = 4.0$ and $\lambda = 1.0$ for VGG-PL205 and ResNet-PL365, respectively, which are set as default hyperparameters in subsequent experiments. We set the same hyperparameters $\{h_1 = 8192, N = 32, \lambda = 4.0\}$ for VGG-IN pretrained models.

TABLE VII

ABLATION STUDIES ON MIT67. CLASSIFICATION ACCURACY (%) IS REPORTED AS EVALUATION METRIC. THE BEST RESULT OF EACH COLUMN IS MARKED IN BOLD

Baseline	Spatial	Similarity	Global Information	Node Weak-supervision	Pretrained Model		
					VGG-IN	VGG-PL205	ResNet-PL365
✓	-	-	-	-	74.55	81.27	86.64
-	✓	-	-	-	77.39	84.25	87.84
-	✓	-	-	✓	78.36	84.55	87.91
-	-	✓	-	✓	78.58	84.25	87.99
-	✓	✓	-	✓	78.73	85.22	**88.36**
-	✓	✓	✓	✓	**79.78**	**85.37**	88.06
Improvement Over Baseline					5.23	4.1	1.42

2) Effect of Different Subgraph Combination Methods: We perform a comparison of three different subgraph combination methods, e.g., element-wise *addition*, *maximum* and *product*. Table VI illustrates the results of LGN without global information on MIT67 dataset. The *product* combination method outperforms other methods. Compared to *addition* and *maximum* combination methods, the *product* method will produce more zero elements in output representations when the inputs are generated from the ReLU layer. This confirms that the sparsity of representation is helpful for the improvement of recognition performance.

3) Ablation Study: We conduct detailed ablation studies of our LGN on MIT67 dataset in Table VII. We analyze the effect of four components, two subgraphs, global information, and node weak-supervision mechanism across three different pretrained models. The ℓ_2 normalized input representations of local regions are averagely pooled as the inputs to a linear SVM classifier, and then produce the classification results as baselines. In Table VII, the best results are marked in bold, which show the improvements of up to **5.23%** over baselines. When applying node weak-supervision mechanism, LGN with VGG-IN has a better improvement. It can be attributed to the worse local representations for scene recognition. Moreover, it can be observed that the global information is useful for VGG16 pretrained models, but not for ResNet50 pretrained model. This may be caused by the better representations of local regions from ResNet-PL365 for capturing the whole image information. We also validate that the spatial and similarity subgraphs are both important to boost the performances and have similar improvements over the baselines. Furthermore, when combining these two subgraphs, there still have improvements, which demonstrates that the two subgraphs exist a complementary relation.

V. CONCLUSION

We propose to construct Prototype-agnostic Scene Layout (PaSL) for each image, and introduce Layout Graph Network (LGN) to explore the spatial structure of PaSL for scene recognition. The pretrained CNN models can be used as region detectors to discover discriminative regions, then form PaSL for each image. To preserve the diverse and flexible spatial structures of PaSLs, we reform each PaSL as a layout graph where regions are defined as nodes and two kinds of independent relations between nodes are encoded as edges. Then, LGN applies graph convolution on the layout graph to integrate spatial and semantic similarity relations into image representations. The detailed ablation experiments demonstrate that LGN has a great ability to capture the spatial and similarity information in PaSL. With the qualitative and quantitative analyses, we prove that PaSLs can capture the useful and discriminative information of the images and have the considerable generalization capability. Experiments on three widely recognized datasets, MIT67, SUN397, and Places365, demonstrate that our approach can achieves superior performances in the setting of a single model in a single scale, and even obtains state-of-the-art results on MIT67 and SUN397.

In the future, we consider jointly learning scene layout and structural models, which may bring better optimization results. Another interesting direction is to explore the multi-scale information from different convolutional layers to help construct more precise and useful spatial structures of scene images.

REFERENCES

[1] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420. [Online]. Available: http://people.csail.mit.edu/torralba/publications/indoor.pdf

[2] H. Izadinia, F. Sadeghi, and A. Farhadi, "Incorporating scene context and object layout into appearance modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 232–239. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909431

[3] Z. Zuo *et al.*, "Learning contextual dependence with convolutional hierarchical recurrent neural networks," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 2983–2996, Jul. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7442840/

[4] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1287–1295.

[5] Z. Zhao and M. Larson, "From volcano to toyshop: Adaptive discriminative region discovery for scene recognition," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 1760–1768, doi: 10.1145/3240508.3240698.

[6] Y. Liu, Q. Chen, W. Chen, and I. Wassell, "Dictionary learning inspired deep network for scene recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7178–7185.

[7] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1307–1314. [Online]. Available: http://ieeexplore.ieee.org/document/6126383/

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017, pp. 1–14. [Online]. Available: https://arxiv.org/pdf/1609.02907.pdf

[9] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3319–3327. [Online]. Available: http://ieeexplore.ieee.org/document/8099837/

[10] X. Song, S. Jiang, and L. Herranz, "Multi-scale multi-feature context modeling for scene recognition in the semantic manifold," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2721–2735, Jun. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7885099/

[11] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[12] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[13] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 487–495. [Online]. Available: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf

[14] X. Song, S. Jiang, B. Wang, C. Chen, and G. Chen, "Image representations with spatial object-to-object relations for RGB-D scene recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 525–537, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8796408/

[15] Y. Huang, X. Cao, X. Zhen, and J. Han, "Attentive temporal pyramid network for dynamic scene classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8497–8504.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf

[17] Z. Ma, J. Xie, Y. Lai, J. Taghia, J.-H. Xue, and J. Guo, "Insights into multiple/single lower bound approximation for extended variational inference in non-Gaussian structured data modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 19, 2019, doi: 10.1109/TNNLS.2019.2899613. [Online]. Available: https://ieeexplore.ieee.org/document/8671492/

[18] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311. [Online]. Available: http://ieeexplore.ieee.org/document/5540039/

[19] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for large-scale image classification," in *Proc. ECCV*, 2010, pp. 143–156. [Online]. Available: http://link.springer.com/10.1007/978-3-642-15561-1_11

[20] S. Lazebnik, C. Schmid, J. Ponce, S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, Jun. 2006, pp. 2169–2178.

[21] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2775–2782.

[22] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017.

[23] Y. Li, M. Dixit, and N. Vasconcelos, "Deep scene image classification with the MFAFVNet," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5757–5765.

[24] B. Chen, J. Li, G. Wei, and B. Ma, "A novel localized and second order feature coding network for image recognition," *Pattern Recognit.*, vol. 76, pp. 339–348, Apr. 2018.

[25] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. ECCV*, 2012, pp. 73–86. [Online]. Available: http://arxiv.org/abs/1205.3137

[26] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. CVPR*, Jun. 2013, pp. 923–930.

[27] B. Zhang, W. Yang, Z. Wang, L. Zhuo, J. Han, and X. Zhen, "The structure transfer machine theory and applications," *IEEE Trans. Image Process.*, vol. 29, pp. 2889–2902, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8911371/

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7485869/

[29] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, "Symbolic graph reasoning meets convolutions," in *Proc. NeurIPS*, 2018, pp. 1853–1863. [Online]. Available: https://pdfs.semanticscholar.org/4959/7c2c8d65f4d3b817aabfa31f16f3791be974.pdf

[30] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, and Y.-Z. Song, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 21, 2019, doi: 10.1109/TCSVT.2019.2936710. [Online]. Available: https://ieeexplore.ieee.org/document/8809264/

[31] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, and Q. Ni, "Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9868–9877, Dec. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8488700/

[32] G. Wu *et al.*, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1993–2007, Apr. 2019.

[33] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Modality and component aware feature fusion for RGB-D scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5995–6004. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Wang_Modality_and_Component_CVPR_2016_paper.pdf

[34] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and Á. García-Martín, "Semantic-aware scene recognition," 2019, *arXiv:1909.02410*. [Online]. Available: http://arxiv.org/abs/1909.02410

[35] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," *Pattern Recognit.*, vol. 74, pp. 474–487, Feb. 2018, doi: 10.1016/j.patcog.2017.09.025.

[36] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5219–5227. [Online]. Available: http://ieeexplore.ieee.org/document/8237819/

[37] K. Zhang *et al.*, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 20, 2019, doi: 10.1109/TCSVT.2019.2936410. [Online]. Available: https://ieeexplore.ieee.org/document/8807207/

[38] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7887720/

[39] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[40] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in *Proc. ICLR*, 2016, pp. 1–19. [Online]. Available: http://arxiv.org/abs/1511.05493

[41] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. ICML*, 2017, pp. 1263–1272. [Online]. Available: http://arxiv.org/abs/1704.01212

[42] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *Proc. CVPR*, Jul. 2017, pp. 20–28. [Online]. Available: http://arxiv.org/abs/1612.04844

[43] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, "Situation recognition with graph neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4183–4192. [Online]. Available: http://www.cs.utoronto.ca/~rjliao/papers/iccv_2017_situation.pdf

[44] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. ECCV*, Jun. 2018, pp. 690–706. [Online]. Available: http://arxiv.org/abs/1808.00191

[45] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. CVPR*, Jun. 2018, pp. 6857–6866. [Online]. Available: http://arxiv.org/abs/1803.08035

[46] A. Torralba, B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, "Object detectors emerge in deep scene CNNs," in *Proc. ICLR*, 2015, pp. 1–12. [Online]. Available: http://arxiv.org/abs/1412.6856

[47] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Comput. Vis., Graph., Image Process.*, vol. 30, no. 1, pp. 32–46, Apr. 1985. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0734189X85900167

[48] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988. [Online]. Available: http://ieeexplore.ieee.org/document/8237584/

[49] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1068–1076. [Online]. Available: http://ieeexplore.ieee.org/document/8237383/

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5206848

[51] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: Objects, scales and dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 571–579. [Online]. Available: http://ieeexplore.ieee.org/document/7780437/

[52] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[54] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

[55] M. D. Dixit and N. Vasconcelos, "Object based scene representations using Fisher scores of local subspace projections," in *Proc. NIPS*, 2016, pp. 2811–2819.

[56] L. Liu *et al.*, "Compositional model based Fisher vector coding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2335–2348, Dec. 2017.

[57] R. Sicre, Y. Avrithis, E. Kijak, and F. Jurie, "Unsupervised part learning for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3116–3124. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/papers/Sicre_Unsupervised_Part_Learning_CVPR_2017_paper.pdf

[58] S. H. Khan, M. Hayat, and F. Porikli, "Scene categorization with spectral features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5639–5649. [Online]. Available: http://ieeexplore.ieee.org/document/8237863/

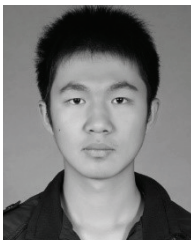
[59] P. Koniusz and H. Zhang, "A deeper look at power normalizations," in *Proc. CVPR*, Jun. 2018, pp. 5774–5783. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Koniusz_A_Deeper_Look_CVPR_2018_paper.pdf

[60] M. Dixit, Y. Li, and N. Vasconcelos, "Semantic Fisher scores for task transfer: Using objects to classify scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 10, 2019, doi: 10.1109/TPAMI.2019.2921960. [Online]. Available: https://ieeexplore.ieee.org/document/8734016/

[61] X. Jin, Y. Chen, J. Dong, J. Feng, and S. Yan, "Collaborative layer-wise discriminative learning in deep neural networks," in *Proc. ECCV*, 2016, pp. 733–749. [Online]. Available: http://link.springer.com/10.1007/978-3-319-46478-7_45

[62] Y. Pan, Y. Xia, and D. Shen, "Foreground Fisher vector: Encoding class-relevant foreground to improve image classification," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4716–4729, Oct. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8678832/

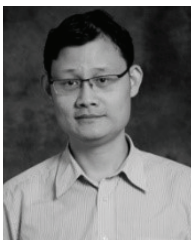
[63] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2055–2068, Apr. 2017.



Gongwei Chen received the B.S. degree from the School of Information Engineering, University of Science and Technology Beijing, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree in computer science with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer vision, machine learning, and image processing.

Xinhang Song received the B.S. degree from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2011, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2017. His research interests include image processing, large-scale image retrieval, image semantic understanding, multimedia content analysis, computer vision, and pattern recognition. He has served as a PC or TPC Member of well-known conferences, such as IJCAI, AAAI, and ACM Multimedia.

Haitao Zeng received the B.S. degree from the School of Geomatics, Shandong University of Science and Technology, Qingdao, China, in 2017. He is currently a Graduate Student in computer science with the School of Mechanical Electronic and Information Engineering, China University of Mining and Technology-Beijing, Beijing, China, and also an Intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer vision and image processing.



Shuqiang Jiang (Senior Member, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, and the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 150 articles on the related research topics. He was supported by the New-Star Program of Science and Technology of Beijing Metropolis in 2008, the NSFC Excellent Young Scientists Fund in 2013, and the Young Top-Notch Talent of Ten Thousand Talent Program in 2014. He is a Senior Member of CCF and a member of ACM. He has also served as a TPC Member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM. He won the Lu Jiaxi Young Talent Award from CAS in 2012 and the CCF Award of Science and Technology in 2012. He is the Vice Chair of the IEEE CASS Beijing Chapter and the ACM SIGMM China Chapter. He is the General Chair of ICIMCS 2015 and the Program Chair of ACM Multimedia Asia 2019 and PCM2017. He is an Associate Editor of the IEEE MULTIMEDIA and *Multimedia Tools and Applications*.