Learning Scene Attribute for Scene Recognition

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

***Abstract*—Scene recognition has been a challenging task in the field of computer vision and multimedia for a long time. The current scene recognition works often extract object features and scene features through CNN, and combine these two types of features to obtain complementary and discriminative scene representations. However, when the scene categories are visually similar, the object features might lack of discriminations. Therefore, it may be debatable to consider only object features. In contrast to the existing works, in this paper, we discuss the discrimination of scene attributes in local regions and utilize scene attributes as the complementary features of object and scene features. We extract these visual features from two individual CNN branches, one extracting the global features of the image while the other extracting the features of local regions. Through contextual modeling framework, we aggregate these features and generate more discriminative scene representations, which achieve better performance than the feature aggregation of object and scene. Moreover, we achieve the new state-of-the-art performance on both standard scene recognition benchmarks by aggregating more complementary visual features: MIT67 (88.06%) and SUN397 (74.12%).**

***Index Terms*—Scene recognition, Scene attribute**

I. Introduction

SCENE recognition has already become one of the most challenging issues in computer vision and multimedia in recent years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], which can be applied in many fields, such as Security, Autopilot, Robot and AI camera. Generally, a scene is composed of various semantic concepts, including explicit objects, backgrounds, and scene attributes. These characteristics determine that learning discriminative information of scenes requires comprehensive consideration of many semantic concepts. To better extract useful information in the scene, in this paper, we propose to explore the relationship between the three features of objects, scene attributes, and scenes, and apply them to conduct scene recognition.

In the past few years, with the impressive performance of deep learning techniques (i.e. CNN [18], [19], [20], [21], [22]) in other visual recognition issues [23], [24], [25], [26], [27], [28], many scene recognition methods [9], [10], [11], [12], [13], [14], [16], [29], [30], [31] based on CNNs were proposed. Zhou *et al* [16], [30] train a series of CNNs on a scene-centric dataset called Places, which overcomes the lack of support for large-scale scene dataset. Besides learning scene features, some works [11], [32], [33] consider improving the capability of CNN by learning more regions in the image. However, those regions are selected without explicitly semantic meanings, such as objects and scene attributes. Therefore, some methods [9], [12], [13], [14] attempt to use object features as complementary features of scene features, so as to combine these features and obtain more robust scene representations. The composition of scenes is complex and abstract, which contains not only objects but also attributes and backgrounds, etc. More importantly, the object features are not very robust when the related scene categories are visually similar. For example, in Fig. 1, we demonstrate two images from scene categories *class*room and *diner indoor*, and their similar visual contents in the local regions, such as the *chair*. Projecting the object features into scene semantic space in Fig. 1 (b), we notice that their semantic probabilities are ambiguous, and these local patches are less related to their global scene categories. This phenomenon makes the object features ambiguous and less complementary to the scenes. So it is not sufficient to consider only object features as the complementary visual features of the scene features.

In addition to objects, scene attribute is also one type of re-

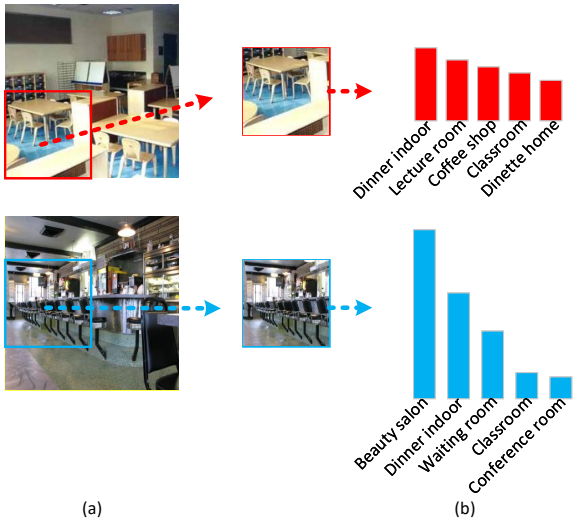


Fig. 1. The object features are not very robust when the related scene categories are visually similar: (a) images belong to classroom (top row) and diner indoor (bottom row) categories of the SUN397 dataset, (b) similar visual contents are labeled in red or blue bounding box and the corresponding semantic probability. Note that patches from different categories have similar contents, which leads the issue that the object features are less discriminative when representing these contents.

H. Zeng is with China University of Mining and Technology, Beijing, 100083, China, and also an intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China email: haitao.zeng@vipl.ict.ac.cn. X. Song is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China email: xinhang.song@vipl.ict.ac.cn. G. Chen is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China email: gongwei.chen@vipl.ict.ac.cn. S. Jiang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and also with University of Chinese Academy of Sciences, Beijing, 100049, China email: sqjiang@ict.ac.cn.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2944241, IEEE Transactions on Multimedia

gional representation. The distribution of these scene attributes is discriminative to those visually similar scenes because scene attributes not only represent the meaning of image contents from visual aspects such as objects or materials, but also from other aspects such as functions, and surface properties. For instance, in Fig. 2, two examples are from different scene categories *classroom* and *diner indoor*, which are visually similar due to the coexistence of objects such as the *chair*, and the coexistence of visual attributes *enclosed area* and *no horizon*. However, they have different functional attributes such as *reading* and *teaching-training* in the *classroom*, and *eating* in the *diner indoor*, which are helpful to distinguish those scenes. Since the scene attributes describe the similar scenes from various aspects, it makes the visually similar scenes distinguishable in the attribute aspects, which indicates that scene attributes can be modeled to represent local patches. Additionally, considering the object features are less discriminative to some scenes that often appear in some regions, the scene attribute features might be complementary to object and scene features, and combining them can obtain more appropriate features of the scene. Moreover, there are some scene attributes consistently appearing in certain categories, which are barely affected by image variance of intra-class (e.g., *enclosed area, no horizon, reading, teaching-training and wood (not part of a tree)* in the *classroom*). These highly relevant scene attributes represent the discrimination between related scene categories and share the consistent patterns of the same scene categories. Therefore, scene attributes in local regions can be used as complementary features of object and scene features in scene recognition.

Fig. 2. The images and scene attributes from athletic field/outdoor and park categories of SUN Attribute dataset. Athletic field/outdoor and park are similar scene categories due to similar objects, such as grass. But they contain discriminative scene attributes (The red words represent the scene attributes that coexist in this category).

In this paper, we develop a contextual modeling framework by taking full use of attributes for scene recognition, which exploits contexts between *object*-centric, *scene attribute*-centric and *scene*-centric features for scene recognition. First, based on the pre-trained ImageNet-CNN, we propose to train the Attribute-ImageNet model that can extract the features, which are more discriminative than original ImageNet features. Then we also extract multi-model features in the setting of multi-scale, consisting of 1) Places features, 2) ImageNet features, 3) Attribute-Places features with the (224x224, 448x448) sizes of images from the corresponding CNN models. These complementary visual features (Attribute-ImageNet, ImageNet, Attribute-Places) allow us to generate more discriminative scene representations by exploiting the relation of multiple semantic spaces. In contrast to previous scene attribute based works, we focus on the local representations of scene attributes, which can generate better complementary visual features to the scenes, benefiting from the discriminative characteristic in the local regions. To the best of our knowledge, no previous works have attempted to research scene attributes on the local regions. Finally, all features are projected into a common space and aggregated with contextual modeling approach for scene recognition. To evaluate the effectiveness of our method, we design several experiments on two standard datasets MIT67 [34] and SUN397 [35], separately, and acquire the new state-of-the-art performance which demonstrates the effectiveness of our method. The main contributions of our method are summarized as follows:

(1) We introduce the scene attributes as the complementary visual features of the scenes to describe the local patches.

(2) We incorporate both local patch features and global image features from Attribute-ImageNet and Places with a contextual modeling method and obtain better results than the combination of the visual features of ImageNet and Places.

(3) We achieve the state-of-the-art performance for both standard benchmarks (88.06% on MIT67, 74.12% on SUN397) by aggregating more complementary visual features of the scene with the multi-scale contextual modeling method.[1]

II. Related Works

In this section, we review the research works related to our paper in two aspects: scene recognition, and attributes for image recognition. We also discuss the differences or connections between these related works and our method.

A. Scene recognition

Scene recognition is an important research topic in the field of computer vision . The early methods, such as GIST [38] mainly focus on extracting the global features for scene recognition. As the local region is also useful and important for scene recognition, some local descriptors such as SIFT [1] have been employed with the Bag of Words encoding method for scene classification, such as the Fisher Vector [39]. However, these local representations could not completely describe the abstract and complex scene images, as they only describe images using low-level visual information. To deal with this problem, some methods propose to learn the mid-level concepts for scene recognition [2], [3], [4], [6]. Li *et al*. [3] propose a rich object representation Object bank for scene recognition. Zhang *et al*. [4] further enhance the description

[1]Code can be found at: https://github.com/zenght/scene_attribute_MP

of the Object bank, and propose object-to-class representations to improve the discrimination between scenes.

Recently, the deep neural networks have made significant progress in the task of image recognition. Several deep neural network structures have been exploited to facilitate the development of image recognition, such as AlexNet [18], VGGNet [19], ResNet [22] and DenseNet [21]. Therefore, some methods [9], [10], [11], [12], [14], [17], [29], [30], [33], [36] attempt to extract the visual representations for scene recognition through convolutional neural networks. However, the structure of deep neural network lacks the geometric invariance [33], which results in poor robustness to scene recognition. To deal with this problem, several works [9], [11], [12], [13], [14], [15], [33] attempt to extract the discriminative regional representations of scenes. Generally, the first step of these methods is to extract the local representations from image patches, and then aggregate these local representations by encoding methods. For example, Gong *et al*. [33] implement VLAD to aggregate the local features for scene recognition. Song *et al*. [12] use Markov Random Field to encode the multiple scale visual features. Wang *et al*. [14] propose the PatchNet that is a CNN architecture with small image region inputs and they also use the image-level label to train their networks on the large-scale dataset Places and ImageNet in weak supervision. In contrast, before weakly supervised training, we first train our models with scene attributes, and then we fine-tune these Attribute-CNNs on evaluation datasets to train the weakly supervised network.

Since the independent visual features cannot extract the abstraction of scenes well, there are several methods [9], [12], [13], [14], [17], [29], [36] attempting to combine multiple visual features, to obtain more discriminative scene representations. For instance, Cheng *et al*. [17] use object selection to generate more discriminative representations and combine them with the global scene representations. Song *et al.* [36] exploit the semantic simplex, and combine the multiple features in the same semantic space. Zhao *et al.* [9] propose an adaptive discriminative region discover (AdiRed) approach, which uses a sliding window to select the discriminative region, and extract the multi-scale features from both object-centric and scene-centric models. Li *et al.* [13] derive an embedded implementation of the mixture of factor analyzers Fisher vector (MFA-FV), and apply it to develop an end to end network, MFAFVNet, which uses the ROI pooling to accept the feature maps of multiple inputs and produces the fixed size output to the fully connected layers. The MFAFVNet is trained based on the ImageNet model, and concatenate the output with the scene-centric model.

Our contextual modeling method still follows the line of exploring the discriminative semantic concepts and meaningful regions for scene recognition. The main differences between our method and the previous methods are summarized as follows: (1) we utilize the scene attributes to describe the local regions; (2) we exploit the context relations among scene attributes, objects and scenes for scene recognition, to obtain the complementary scene representations.

B. Attributes for image recognition

The researches about attributes for image recognition are various. In the early works, the basic attributes are the central issue of researches [40], [41], [42], such as texture, shape and color. The work of [40] is the start of learning attributes, they show us that attributes can be learned for object recognition through weakly supervised learning method, and a set of classifiers is trained to predict the existence of human-labeled attributes in the data, and to discover unseen classes through intra-class transfer without the phase of training. Lampert *et al.* [42], [43] apply the attributes to zero-shot learning and create the Animals with Attributes (AwA) [43] dataset, which contains 85 attributes. Particularly, some attributes in AwA dataset are totally not conventional visual concepts, even though these attributes seem to be relatively abstract (e.g., "fast" and "weak"), their validity is demonstrated in the object classification, and these special visual attributes also appear in later works [44], [45], [46]. However, the researches about attributes for the scene are not many. Oliva *et al*. [38], [47] explore a small-scale scene attributes in their works, where eight "spatial envelope" attributes are discovered by the participant's manual division of eight scene categories. Patterson *et al.* [46] use scene attributes as intermediate representations for scene classification.

In contrast to existing methods [25], [26], [42], [48], [49] related to attributes, we do not directly apply attributes as intermediate representations to predict scene categories. We prefer to utilize scene attributes as the complementary features and combine them with object and scene features to obtain discriminative scene representations.

III. Motivating Scene Attribute for Scene Recognition

In this section, we review the scene attributes and related works. Then, we analyze the scene attributes to motivate our method.

A. Scene attribute

Patterson *et al*. [46] propose a scene attribute-centric dataset based on SUN [35], named SUN Attribute. The SUN Attribute dataset is concentrated in 102 discriminative scene attributes, and these attributes are highly related to functions/affordances, materials, surface properties, and spatial layout. Here, we demonstrate some images of scene attributes belonging to SUN Attribute and their corresponding scene categories in Fig. 3. Note that the images we demonstrate in Fig. 3 follow the rule: the selected scene attributes are often appeared in this category (probability is greater than 40%). From Fig. 3, we observe that some of the scene attributes represent the global environment and spatial layout, such as *electric–indoor lighting* in the *auditorium* and *bedroom.* Besides, there are also several scene attributes representing the functions of local regions, for example, *spectating–being in audience* and *congregating* in the *auditorium, working* and *medical activity* in the category of the *hospital room.* And a few scene attributes such as *cloth* in the *bedroom* and *hospital room,* or *paper* in the *classroom,* represent the discriminative visual contents of local

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2944241, IEEE Transactions on Multimedia

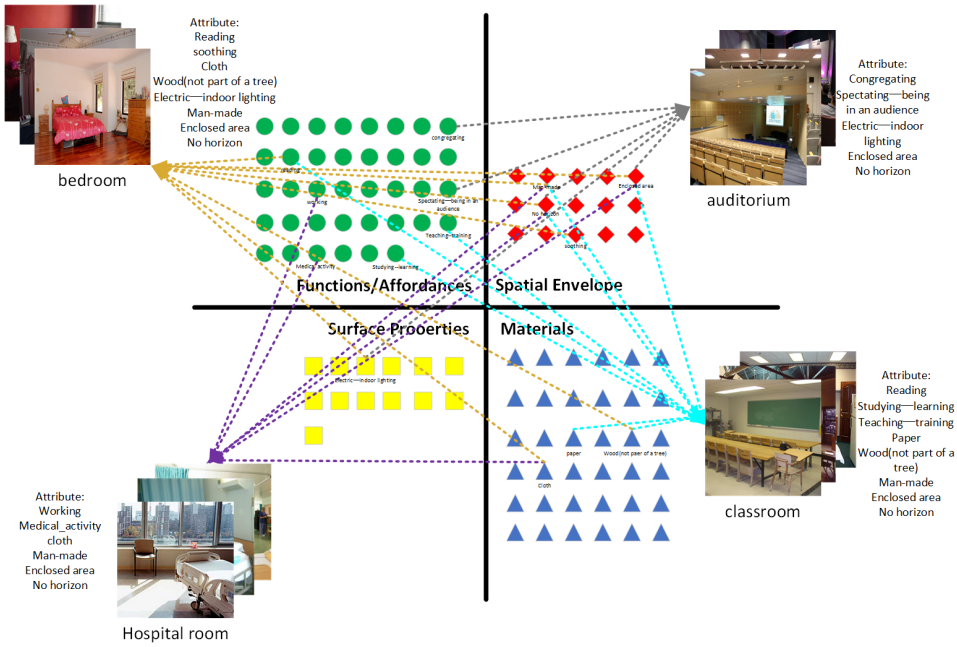


Fig. 3. Examples of scene attributes images from SUN Attribute. Red diamonds, green circles, blue triangles and yellow squares represent 15 Spatial Envelope attributes, 38 Functions/Affordances attributes, 36 materials attributes and 13 Surface properties, respectively.

regions. As we demonstrated in Fig. 3, these scene attributes are often emerged and highly related to the certain scene category, which indicates the characteristic of scenes from various aspects. Based on the above observations and analysis, we find that the scene attributes have discriminative power in the related scene categories. For example, the *hospital room* and *bedroom* are similar due to the coexistence objects (e.g., bed). But the functional scene attributes are different, such as *reading* in the *bedroom*, *working* and *medical activity* in the *hospital room,* which implies that the scene attributes are helpful to distinguish those scenes.

Patterson *et al*. [46] discuss attributes as intermediate representations for scene classification. However, their work remains some limitations: 1) they don't take into account the complementary relation among scenes, objects, and scene attributes; 2) they neither consider the fact that scene attributes are discriminative when representing local regions. Tackling the limitations of [46] motivates us to further explore the role of scene attributes in scene recognition.

B. Scene attribute analysis

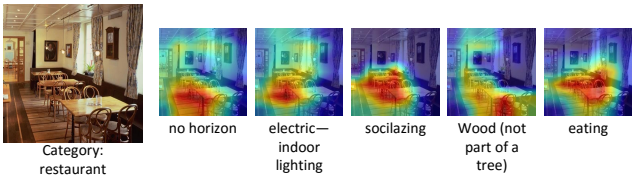


Fig. 4. The activated visualization feature maps of some scene attributes on SUN397 restaurant image based on Attribute-CNN. Grad-CAM [50] is applied to realize the visualization of the discriminative localization region of each attribute (The brighter the image, the more discriminative it is).

Generally, the scene attributes in the SUN Attribute describe the scene from various aspects, ranging from conventional visual concepts to abstract invisible concepts. The conventional visual concepts generally describe the materials in the local regions. The abstract scene attributes can also be regarded as descriptions of local regions, such as the scene regional function and surface properties. Here, inspired by Selvaraju *et al.* [50], we demonstrate the visualization of some scene attributes belonging to the *restaurant* in Fig. 4. From Fig. 4, we discover that activated regions of these visualization results are semantically significant. As shown in Fig. 4, the activated region of *wood (not part of a tree)* tends to be located in the region of wood products, indicating that local regions are composed of this material. In addition, the functional scene attributes in the local regions tend to locate where this function may occur. For instance, the activated region of *eating* in Fig. 4 is located in the dining table where *eating* might occur, overlapping with the region of *socializing*. Based on the above observations, we notice that some scene attributes are salient in describing the local regions, and these salient local scene attributes are highly relevant to the scene category. Thus, the scene attributes in the local regions can be modeled to obtain discriminative local features.

The same region might contain several scene attributes, indicating that the scene attributes in the local patches are coexisting. For instance, in Fig. 5, the scene category *diner indoor* contains several scene attributes from different aspects of the local patch, such as *enclosed area, glossy, no horizon, electric-indoor* and *congregating*. This phenomenon inspires us that scene attributes are highly relevant to the scene categories, indicating the pattern of the local regions.

The coexistence scene attributes are discriminative in similar visual contents. The examples are shown in Fig. 5 (a), three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2944241, IEEE Transactions on Multimedia

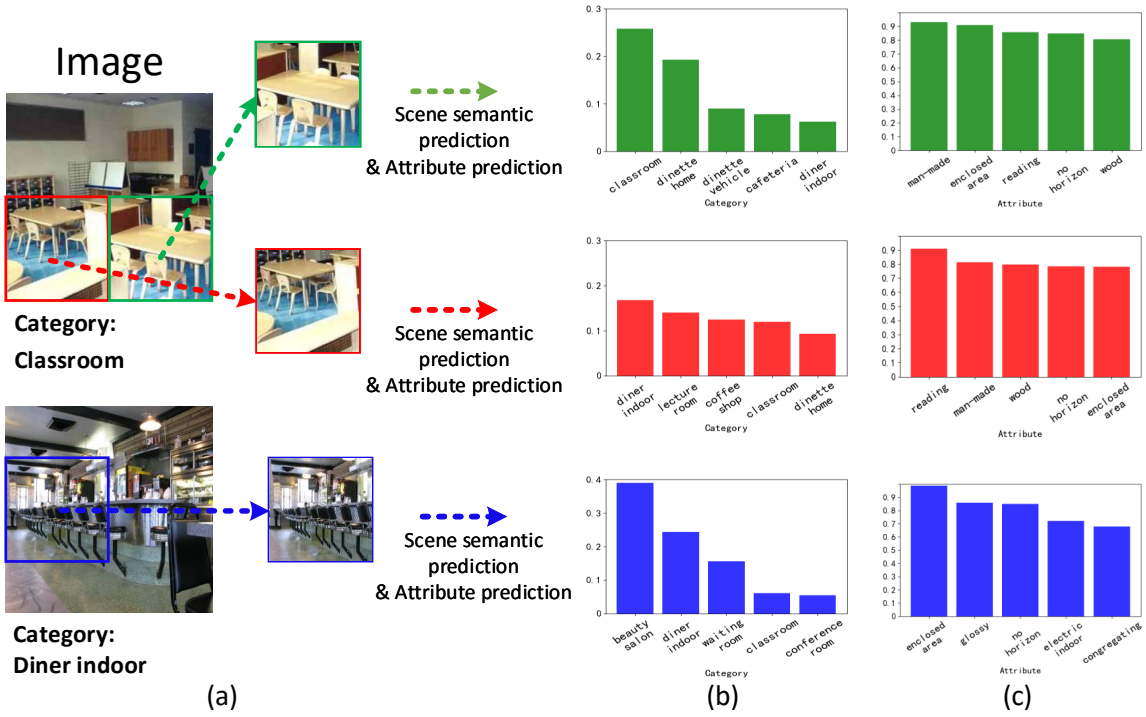


Fig. 5. The scene semantic probability and attribute prediction of the corresponding labeled patches. (a) input images of the category classroom and diner indoor from the SUN397. (b) The scene semantic probabilities of corresponding labeled patches. (c) The attribute prediction of corresponding labeled patches.

images are from two related scene categories *classroom* and *diner indoor*, which contain similar local regions due to the co-existence of objects (e.g., chair). This leads to the ambiguous representations, as shown in the scene semantic probabilities in Fig. 5 (b). However, their attribute predictions in Fig. 5 (c) demonstrate that the similar visual contents might contain different scene attributes. For example, the group of scene attribute in the *diner indoor* category is {*enclosed area, glossy, no horizon, electric-indoor* and *congregating*}, while the scene attribute group in the related *classroom* category is {*reading, no horizon, man-made, enclosed area* and *wood*}. These different scene attribute groups indicate their discriminative attributes in the related scene categories. Therefore, modeling scene attributes can effectively distinguish similar regions of different scenes.

Generally, the adjacent patches might contain coexisting visual contents, so their scene semantic probability distribution should be consistent, which we refer to as (scene) consistent patterns. However, the adjacent patches might represent inconsistent patterns due to different objects composition and spatial layout. For instance, the adjacent patches from the *classroom* category in the top two rows in Fig. 5 contain similar objects but with the different layout, which leads to inconsistent patterns, as shown in their scene semantic probabilities in Fig. 5 (b). However, we find that the adjacent patches might exhibit similar attribute prediction results, such as *reading, no horizon, man-made, enclosed area* and *wood* in Fig. 5 (c). This indicates that the adjacent patches described by scene attributes are similar. Based on the above observations, we analyze that similar scene attributes of adjacent patches are beneficial to enhance the consistent patterns of adjacent patches, thereby enhancing the discrimination of image-level representations when patch-level representations are aggregated into image-level representations.

IV. Approach

A. Scene attributes modeling

As discussed in section III, scene attributes have several advantages when describing scenes: (1) some scene attributes tend to represent the visual concept of local regions; (2) local patches have multiple coexistence scene attributes; (3) the groups of scene attribute in the similar scenes are different; (4) the scene attributes in the adjacent patches are similar. Considering the discrimination of scene attributes in representing scenes, especially in local regions, a natural question is raised: whether scene attributes can be applied to develop a more generalized scene recognition method. Thus, in our method, we consider contextual modeling method based on scene attributes. We consider the contextual relations from different aspects, including local patches, multi-scale and intra-category. The method of [12] has inspired us to model these contextual relations with Markov Random Field (MRF) model. In [12], they adopt MRF to find the relation of the feature in a contextual spatial, and to optimize each feature they employ the hierarchical message passing method that computes the

spatial distance between the current feature and others to construct the energy function, and optimizes the current feature by the gradient of the energy function.

1) Local patch scene attribute modeling: Considering the discriminative representations of scene attributes in local regions. We aim to model the scene attributes based on local patches. The adjacent image patches usually represent similar visual contents. However, because objects are composed or arranged differently, the patterns of image patches may not be consistent. Considering the adjacent patches might represent similar scene attributes, we attempt to model scene attributes to improve the consistent patterns of adjacent image patches. We define the energy as the distance between patches, and the objective is to minimize the energy function, which is equivalent to minimizing the gap between patches, the corresponding energy is:

$$E(\dot{s_1}, ...\dot{s}_{N_s}, s_1, ...s_{N_s})$$

$$= \lambda \sum_{n=1}^{N_s} G\left(\dot{s_n}, s_n\right) + \tau \sum_{n=1}^{N_s} \sum_{n' \in \left\{n, n'\right\}} G\left(\dot{s_n}, s_{n'}\right) \quad (1)$$

Where $\{s_1, ...s_{N_s}\}$ demonstrate the observed local features, and $\{\dot{s_1}, ...\dot{s_{N_s}}\}$ represents the optimized features. N_s is the number of patches, and $G\left(x, y\right)$ is the energy function that measures the distance, which is the geodesic distance [51]. $n^{'}$ in $\left\{n, n^{'}\right\}$ indicates adjacent patches of n.

2) Intra-category scene attribute modeling: Since images from the same category might exhibit the coexistence of scene attributes, modeling the coexistence of scene attributes can help us emphasize the consistent patterns of the scene categories, and increase the discrimination of the scene categories. Thus, we follow the work of [12], adding the scene category sparse regulation to penalize the flat semantic probabilities and enhance the discrimination between scene categories:

$$P\left(\mathbf{s}\right) = -\sum_{w=1}^{W} s_{(w,n)} \log\left(s_{(w,n)}\right) \quad (2)$$

In (w, n), where $w \in W$ is the scene category index, and *n* represents the number of patches.

3) Multi-scale scene attribute modeling: Multi-scale relations can be modeled to exploit the dependencies among patches from different scales, which can be applied by the semantic representations based on different scales. The semantic representations of the large-scale inputs extracted from the fixed receptive field CNN model is equivalent to the representations obtained in the image patches. Thus, the semantic representations extracted under different scale settings have certain similarities.

Since scene attributes are discriminative in representing scenes in both local and global scales. Thus, we propose to model scene attributes in both scales to enhance the consistent patterns between patches. Modeling in this way is conducive to capturing important local information in the scene and obtaining more discriminative semantic representations. Thus, we model the patches at scale $l = 1, \ldots, L$ with the patches at previous scale $l - 1$, to minimize the gap between related patches, so as to achieve the multi-scale combination. The corresponding energy with L scales is:

$$\begin{aligned} &E(\dot{s}_1^{(1)}, ...\dot{s}_{N_s}^{(1)}, \dot{s}_1^{(L)}, ...\dot{s}_{N_s}^{(L)}, s_1^{(1)}, ...s_{N_s}^{(1)}, s_1^{(L)}, ...s_{N_s}^{(L)}) \\ &= \lambda \sum_{l=1}^{L} \sum_{n=1}^{N_s} G\left(\dot{s}_1^{(l)}, s_n^{(l)}\right) \\ &+ \beta \sum_{n=1}^{N_s} \sum_{l=2}^{L} \sum_{n' \in \left\{n, n'\right\}} G\left(\dot{s}_n^{(l)}, \dot{s}_{n^l}^{(l-1)}\right) \end{aligned} \quad (3)$$

where n^l in (n, n^l) describes the adjacent patches in previous scale setting. $s^{(l)} = \{s_1^{(1)}, ..., s_n^{(1)}, ..., s_1^{(L)}, ..., s_n^{(L)}\}$ represents the set of multi-scale representations.

B. Multiple features contextual relation modeling

Some existing scene recognition methods [9], [12], [13], [14], [29], [36] aggregate object and scene features to produce the comprehensive scene representations. These works inspire us to explore the complement among different visual features. Therefore, in our method, we assume three kinds of feature sets in M, including $M_1 = \{AI, PL\}$, $M_2 = \{IM, PL\}$, and $M_3 = \{IM, PL, AI, AP\}$ that are extracted by using the fine-tuned ImageNet-CNN, Places-CNN and two kinds of Attribute-CNN, respectively. To explore the complement among different visual features, we assume the set of complementary features M_1. Since the combination of scene and object features are usually used in recent methods [9], [12], [13], [14], [29], [36], we define M_2 as a comparison set for M_1. Moreover, to fully explore the effectiveness of complementary features, we define the feature set M_3 that consists of four kinds of features.

To fully explore the effectiveness of complementary features in scene recognition, we propose a joint contextual relation model with multi-feature settings, and the corresponding energy is:

$$\begin{aligned} &E(\dot{s}_1^{(1)}, ...\dot{s}_{N_s}^{(1)}, \dot{s}_1^{(L)}, ...\dot{s}_{N_s}^{(L)}, s_1^{(1,1)}, ... \\ &s_{N_s}^{(1,1)}, s_1^{(L,M)}, ...s_{N_s}^{(L,M)}) \\ &= \lambda \sum_{l=1}^{L} \sum_{n=1}^{N_s} \sum_{m \in M} G\left(\dot{s}_n^{(l)}, s_n^{(l,m)}\right) \\ &+ \tau \sum_{n=1}^{N_s} \sum_{l \in L} \sum_{n' \in \left\{n, n'\right\}} G\left(\dot{s}_n^{(l)}, s_{n'}^{(l)}\right) \\ &+ \beta \sum_{l=2}^{L} \sum_{(n, n^l)} G\left(\dot{s}_n^{(l)}, \dot{s}_{n^l}^{(l-1)}\right) + \gamma P\left(s_n^{(l)}\right) \end{aligned} \quad (4)$$

where $\left\{s_1^{(1,1)}, ...s_{N_s}^{(1,1)}, s_1^{(L,M)}, ...s_{N_s}^{(L,M)}\right\}$ represents the observed multi-scale and multi-feature set. $\left\{\dot{s}_1^{(1)}, ...\dot{s}_{N_s}^{(1)}, \dot{s}_1^{(L)}, ...\dot{s}_{N_s}^{(L)}\right\}$ represents the optimized features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2944241, IEEE Transactions on Multimedia

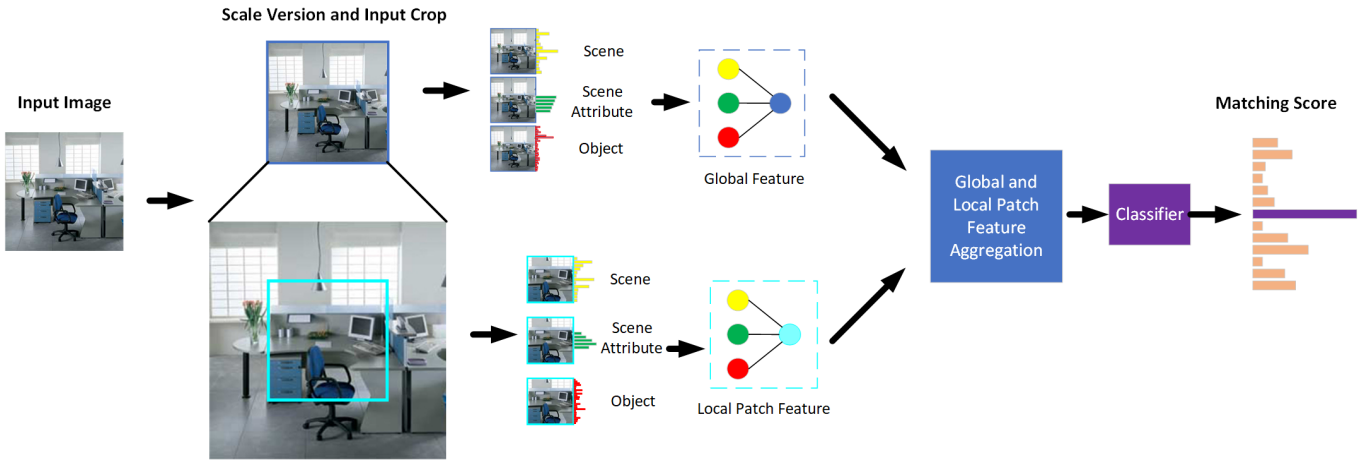


Fig. 6. The framework of our method. We first exploits multiple contexts (e.g., multi-scale and multi-feature) to optimize patch-level representations, and these local patches are aggregated into a global visual representation through MRF encoding method (red, yellow and green balls represent different semantic concepts. Blue represent different scale visual representations). Finally, scene recognition is performed on image-level representations using a linear SVM.

from different scales. In $\{n, n^l\}$, n^l demonstrates the adjacent patches in previous scale $l-1$.

However, such global energy function is difficult to be solved directly due to a large number of parameters. In this paper, to decrease the complexity of global function, we solve it with a hierarchical message passing method [12] which optimizes the energy function through local aspects instead of global. In the local function, only the current $\dot{s}_n^{(l)}$ is variable, and the others are fixed, the corresponding local function is as follows:

$$E\left(\dot{s}_n^{(l)};\varphi_n^{(l)}\right)=\lambda\sum_{m\in M}G\left(\dot{s}_n^{(l)},s_n^{(l,m)}\right)+\tau\sum_{\{n,n'\}}G\left(\dot{s}_n^{(l)},\dot{s}_{n^l}^{(l)}\right)+\beta\sum_{(n,n^l)}G\left(\dot{s}_n^{(l)},\dot{s}_{n^l}^{(l-1)}\right)+\gamma P\left(s_n^{(l)}\right) \quad (5)$$

where $\varphi_n^{(l)}$ represents the set related to $\dot{s}_n^{(l)}$, which contains the complementary features $m\in M$, and the adjacent patches in scale l, and the adjacent patches from previous scale $l-1$. The local function can be optimized by gradient descent, where gradient of $G(a,b)$ is:

$$\frac{\Delta G(a,b)}{\Delta a}=\frac{\sqrt{b}}{2\sqrt{a}\sqrt{1-\left(\sqrt{a}\sqrt{b}\right)^2}}$$

$$\Theta(a,b)=\frac{\Delta G(a,b)}{\Delta a}$$

where

$$\frac{\partial E\left(\dot{s}_n^{(l)};\varphi_n^{(l)}\right)}{\partial s_n}=\lambda\sum_{m\in M}\Theta\left(\dot{s}_n^{(l)},s_n^{(l,m)}\right)+\tau\sum_{\{n,n'\}}\Theta\left(\dot{s}_n^{(l)},\dot{s}_{n^l}^{(l)}\right)+\beta\sum_{(n,n^l)}\Theta\left(\dot{s}_n^{(l)},\dot{s}_{n^l}^{(l-1)}\right)+\gamma P\left(s_n^{(l)}\right) \quad (6)$$

Thus, the current $\dot{s}_n^{(l)}$, upgrade as $\dot{s}_n^{(l)}+k\Delta\dot{s}_n^{(l)}\rightarrow\dot{s}_n^{(l)}$, where $\Delta\dot{s}_n^{(l)}$ is the output of Eq. 6, and k is fixed update rate. To be specific, all $\dot{s}_n^{(l)}$ simultaneously calculate the gradient of their local functions and implement the update synchronously.

C. Visual representations encoding

Our task is to recognize scene labels of the images, which requires to aggregate patch-level representations (local features) into global ones to feed the scene classifier. The framework of our method is shown in Fig. 6, which is divided into two individual CNN branches: the first branch extract global features from the entire image inputs (224x224 inputs), while the other branch extracts the local patch features from the resized image inputs (448x448 inputs). The global features focus on representing the holistic visual contents of the scene images, while the local patch features aim to describe the discriminative visual contents in the local regions. The two branches represent the image from two scales, each of which is complementary. In each branch, we consider three kinds of semantic concepts. In contrast to the current methods, we consider scene attributes, objects and scenes together, each of which is complementary to the other. Then, we aggregate these global features and local patch features with the MRF contextual model [12]. Finally, the aggregated image-level representations are the input to the classifier, we adopt the linear SVM (penalty parameter C = 1) for classification, the final predicted category is determined by the maximum result of the linear SVM classifier.

V. Experiments

In this section, we demonstrate the experimental setting and illustrate the performance of our proposed method on the standard benchmark datasets.

The purpose of the experimentation is to support the following goals:

(1) Verify the performance of the Attribute-CNN on the scene recognition task and compare it with the competing CNN learned for object and scene features.

(2) Investigate the effectiveness of the Attribute-ImageNet features for scene recognition when aggregated with the Places features in the multi-scale setting.

(3) Evaluate the effect of aggregating more complementary visual features of the scene in the scene recognition.

A. Experimental dataset

SUN Attribute [46] contains 102 attributes vary from 707 categories and 14,340 images. Following the standard protocol of [46], we randomly divided the dataset into 90% training and 10% testing. The experiments run on 10 random divisions of the training set and testing set. The reported result is the mean average precision (MAP).

MIT67 [34] includes 67 indoor scene categories and 15,620 images. There are at least 100 images in each category. Following the original paper [34], we utilize 80 images for training and 20 images for testing from each class.

SUN397 [35] is a much larger dataset that varies from the *abbey* to *youth hostel,* which consists of 397 categories and 108,734 images, where each class also contains more than 100 images. We also follow the evaluation of original paper [35], where each category involves 50 images for training and 50 for testing. Since SUN397 is a large data set, it is challenging to verify the generalization capabilities of our method on it.

B. Implementation details

In our experiments, we train two kinds of Attribute-CNNs on SUN Attribute dataset, through fine-tuning the CNN models pretrained on ImageNet and Places, respectively. Only the fully connected layer is trained, which has similar performance to fine-tuning all layers. In our experiment setting, we adopt the DenseNet [21] as the basic CNN architecture. The input image size is 224x224. We set the initial learning rate to 0.08, and divide the learning rate by 10 after 10 epochs. The weight decay is 0.0001 and the momentum is 0.9, and we obtain Attribute-Places-CNN with MAP 57.62%, and Attribute-ImageNet-CNN with MAP 54.18%.

Since our experiments are conducted on the benchmark datasets MIT67 and SUN397, and their training data are not large, we intend to train the corresponding CNNs on both datasets by fine-tuning the pretrained ImageNet-CNN, Places-CNN, Attribute-ImageNet-CNN and Attribute-Places-CNN, respectively. Moreover, inspired by the PatchNet [14], we also train patch-level CNNs to extract the local features. In the patch-input setting, each image is resized into 256x256, and then cropped to 128x128 small patches as inputs. We apply the image-level label to these cropped patches, which can be regarded as weakly supervised training. Image-level annotation might be not accurate for some image patches, and the loss might be also larger than image-level CNNs. However, because the number of training samples has increased, this method is still able to learn local visual contents of scenes.

These trained CNNs allow us to extract the corresponding global and local features. Since the receptive field of the trained CNN model is fixed, the extracted features of different scales are obtained by changing the size of the input images, and the features extracted by large-scale inputs are equivalent to those obtained from the image patches. Therefore, in our experiments, we consider extracting global features from images with 224x224 inputs and extracting local features of images from 448x448 inputs. Note that we extracted two types of local features, one is obtained by CNN trained with 224x224 as inputs, and the other is obtained by CNN weakly supervised trained with 128x128 patch inputs.

C. Evaluation on Attribute-CNN

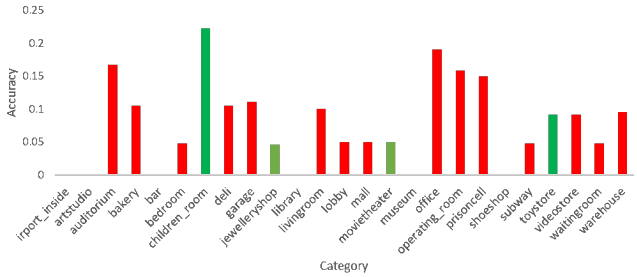


Fig. 7. Low accuracy category classification results on MIT67. (The red column represents the increase compared with the original result (ImageNet), and the green represents the decline.)

In order to evaluate our method, we begin our experiments on the standard benchmarks MIT67. Considering the limitation of object features in similar scenes, and the discrimination of attributes in the similar contents, we make a fair comparison between ImageNet-CNN and Attribute-ImageNet-CNN. To be specific, we extract the activation maps of last convolutional layer as features, and then train Softmax classifiers. We believe that the classification results lower than 80% in the baseline model (ImageNet-CNN) can be regarded as low classification categories (25 categories on MIT67). The Attribute-ImageNet-CNN might improve the classification results of these categories. We evaluate the accuracy of these categories based on Attribute-ImageNet-CNN model, the results are shown in Fig. 7. From Fig.7, we find that 15 categories have improved the classification performance, and 6 categories remain stable, and 4 categories get worse results. Moreover, there is the obvious improvement of some scenes, such as *office*, *auditorium*, *operating room* and *garage*. As a result, scene attributes have an impact on the low accuracy categories.

For further evaluation, we also demonstrate the classification results on MIT67 across all categories. From Table I, we observe that the Attribute-ImageNet-CNN obtain 82.31% of accuracy which is 1.12% higher than the ImageNet-CNN. Based on the above experiments, the improvement of classification accuracy indicates the effectiveness of Attribute-ImageNet-CNN. In addition, we believe that these results also evaluate that scene attribute and object features are complementary, and that the combination of the two features can generate more discriminative scene representations. Moreover, we also aggregate the local features through average pooling and get 75.22% accuracy of Attribute-ImageNet-CNN that is 1.04%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2944241, IEEE Transactions on Multimedia

higher than ImageNet-CNN results. The better local features aggregated results of Attribute-ImageNet also evaluate that the added scene attributes can capture more discriminative local contents, enhance the consistent patterns of local patches, and evaluate that using scene attributes to represent local regions is beneficial to scene recognition.

However, from Table I, the results of Attribute-Places CNN is slightly lower than Places-CNN. We think that the simple fusion of scene attribute and scene features is not obvious for the scene recognition, it may even have negative effects.

TABLE I
THE CLASSIFICATION RESULTS OF EACH CNN MODEL IN DIFFERENT SCALES ON MIT67

Scale	ImageNet	Attribute-ImageNet	Places	Attribute-Places
224x224	81.04%	82.31%	85.3%	85%
448x448	74.18%	75.22%	82.76%	81.42%

TABLE II
THE CLASSIFICATION RESULTS OF EACH CNN MODEL IN DIFFERENT SCALES ON SUN397

Scale	ImageNet	Attribute-ImageNet	Places	Attribute-Places
224x224	64.15%	63.82%	69.7%	69.66%
448x448	58.25%	58.08%	66.91%	66.68%

The same experiments on SUN397 are shown in Table II. However, the results of Attribute-ImageNet-CNN are slightly lower than the ImageNet-CNN. We analyze the major reason of this result is that the volume of SUN Attribute dataset is relatively small, much less than SUN397, so it performs more obvious in the relatively small evaluation dataset, while the impact on the larger dataset is slightly less.

TABLE III
RESULTS OF DIRECTLY APPLIED SCENE ATTRIBUTES FOR SCENE CLASSIFICATION

Method	MIT67	SUN397
Attribute+SVM [27]	-	27.6%
CNN(without fine-tuning)	28.36%	20.24%
CNN(ImageNet)	82.31%	63.82%
CNN(Places)	85%	69.66%

The existing methods that directly apply scene attributes to scene recognition is not many. Patterson *et al.* [46] train one-vs-all SVM using 102 scene attributes for scene classification, which obtain 27.6% accuracy on SUN397. Moreover, we also conduct further exploration, we extract features based on scene attribute CNN (without fine-tuning) to train scene classifiers, the results are shown in Table III. From Table III, we observe that CNN (without fine-tuning) get 28.36% on MIT67, and 20.24% on SUN397. Compared to other models based on fine-tuning, the results of directly applying scene attributes are lower. We believe the major reason is that the volume of SUN Attribute is small, which cannot provide the same amount of diversity data as Places, resulting in limited robustness of the trained models. Thus, in this paper, we train the Attribute-CNNs by fine-tuning the pre-trained large-scale CNN models. Moreover, attribute classification itself is also a challenging task. Once the attributes are well predicted, the performance of scene classification will be improved.

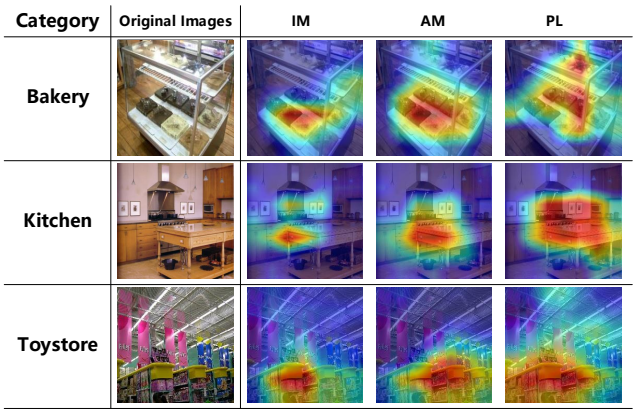


Fig. 8. Activated visualization feature maps from some images on MIT67 based on different models. Grad-CAM[50] is used to realize the visualization of the discriminative localization region of each model (The brighter the image, the more discriminative it is). From left to right: (1) Categories; (2) Original images; (3) Activated discriminative localization feature map of each model.

To evaluate the combination of scene attribute and object features, we evaluate it through a visualization method inspired by Selvaraju *et al.* [50]. And the generated visualization regions are based on the feature maps of the last convolutional layer of each model, it is relatively clear to describe the respective emphasis of each CNN model on the same image.

In Fig. 8, we describe the images from three categories of MIT67 and their visualization results under the CNN model of ImageNet, Attribute-ImageNet and Places. Specifically, we aim to explore whether Attribute-ImageNet-CNN can capture more visual contents about scenes. As shown in Fig. 8, the highlight activated visualization region of each feature map contains rich semantic information. For instance, in the category of *toystore*, the highlight region of ImageNet feature map tends to localize at the *toys'* region, while in the Attribute-ImageNet, we can observe that activated region tends to include not only *toys* but also adjacent *shelves*. This phenomenon also appears in the category of *kitchen* and *bakery*. Based on the above observations, the ImageNet-CNN mainly captures the core object concepts in the scene images, such as *toys* in *toystore*. While Attribute-ImageNet expands the captured region. In other words, the combination of scene attribute and object features focus on more visual contents. Taking the feature maps of Places as the reference, we think that Places-CNN can get better classification results probably because it focuses on more scene contents. Therefore, we think that expanding the activated regions is beneficial to scene recognition. The larger activated region of Attribute-ImageNet indicates that Attribute-ImageNet focuses more visual contents of the scene, which also evaluates that the combination of scene attribute and object features is significant to scene recognition.

D. Exploration on the combination of the attribute in multi-scale and multi-feature

In this subsection, we explore our proposed method that aggregating scene attribute, object and scene features jointly. In our experiment, we extract the features from different pre-trained models in the multi-scale setting, including ImageNet-CNN (IM), Attribute-ImageNet-CNN (AI) and Places-CNN (PL). We compare the performance of AI and PL features aggregation with the aggregation of IM and PL. Through the MRF encoding method, we aggregate multi-scale and multi-semantic visual representations into image-level representations, and the image-level representations are fed into linear SVM classifier for scene classification.

TABLE IV
PERFORMANCE COMPARISON OF MULTI-SCALE AND MULTI-FEATURE COMBINATION ON MIT67 AND SUN397.

Aggregation Strategy	Scale Size	MIT67	SUN397
AI+PL	224,448	87.24%	72.21%
IM+PL	224,448	87.01%	72.03%

We conduct the experiments on MIT67 and SUN397 to evaluate our method. The results are shown in Table IV, the aggregation of AI and PL features obtains higher results than the aggregation of IM and PL, with an accuracy of 87.24% on MIT67 and 72.21% on SUN397 through MRF encoding method. These results are reasonable as AI features not only learns the object features but also learns scene attribute features. In other words, when AI features aggregated with PL features, the aggregated image-level representations are based on three visual features, thus, it's not surprising to obtain better results.

TABLE V
PERFORMANCE COMPARISON OF MULTI-SCALE AND MULTI-FEATURE COMBINATION IN WEAKLY SUPERVISED SETTING ON MIT67 AND SUN397.

Aggregation Strategy	Scale size	MIT67	SUN397
AI+PL	224,448	87.24%	72.21%
	224,448*	**87.84%**	**73.85%**
IM+PL	224,448	87.01%	72.03%
	224,448*	87.39%	73.81%

448* means training through 128x128 inputs,
Feature aggregation method is MRF encoding

To further evaluate the discrimination of scene attributes in the local regions, we perform an additional study to explore the performance of weakly supervised local features. To keep a comparison fair, we only replace the original local features with the weakly supervised trained local features. The images in patch size contain more local and discriminative visual contents, and these local visual contents can be better preserved during the training process. The experiment results on MIT67 and SUN397 are summarized in Table V.

From Table V, the combination of AI and PL features achieves 87.84% accuracy on MIT67, which is 0.6% higher than the original. Moreover, we discover that the improvement on SUN397 is a large margin. In the weakly-supervised training setting, the performance of AI and PL features aggregation improved by 1.64% to the accuracy of 73.85%. The results on the two benchmark datasets suggest that the weakly supervised trained local features might capture more local contents of the scene, which is beneficial to scene recognition. More importantly, based on the results in Table V, we find that the results of AI and PL features combination are higher than the results of IM and PL on both benchmarks. These results indicate that using scene attributes in local patches can enhance the consistent patterns among the patches, and generate more discriminative scene representations.

The empirical experiments highlight the significance of our method in scene recognition. The combination of AI and PL features achieves better experimental results than IM and PL, which evaluates that scene attribute, object and scene features are complementary, and the combination of the three features is able to obtain more discriminative scene representations. In addition, in the weakly supervised training setting, the experiment results verify that the discrimination of scene attributes is beneficial to enhance the consistent patterns of local patches.

E. Exploration on more complementary visual features combination

TABLE VI
COMPARISON EXPERIMENT ON FOUR VISUAL FEATURES COMBINATION ON BOTH BENCHMARKS IN MULTI-SCALE SETTING.

Aggregation Strategy	Scale size	MIT67	SUN397
AI+PL	224,448	87.24%	72.21%
	224,448*	87.84%	73.85%
IM+PL	224,448	87.01%	72.03%
	224,448*	87.39%	73.81%
PL+AP+AI+IM	224,448	87.46%	72.41%
	224,448*	**88.06%**	**74.12%**

448* means training through 128x128 inputs,
Feature aggregation method by MRF encoding

So far, we have evaluated the effectiveness of our combination of AI and PL features. In this subsection, we further explore the significance of four kinds of complementary visual features in the multi-scale setting. The set is M = $\{IM, PL, AI, AP\}$. The experiments are conducted on both benchmarks, the results are shown in Table VI. The performance of four visual features achieves 88.06% accuracy on MIT67, and 74.12% on SUN397. This is reasonable as the number of visual features of the aggregation increases, more complementary features of the scene are considered. Thus, the final scene representations take more features into account and generate more discriminative scene representations.

F. Comparison with state-of-the-art works

Based on our experiments and analysis related to scene attributes, we are going to make the comparison with these state-of-the-art methods in this section. The comparative performances are shown in Table VII and VIII, our combination method obtains the best results among all methods, evaluating the better generalization of our method compared with other methods. Moreover, our four visual features aggregation effectively combines object, scene and scene attribute together, covering the widest semantic space. The 88.06% accuracy on

MIT67 and 74.12% accuracy on SUN397 are the best results on both datasets, to the best of our knowledge, state-of-the-art on scene recognition.

TABLE VII
COMPARISON WITH STATE-OF-THE-ART WORKS ON MIT67

Approaches	Accuracy
Places365+VGGNet16[30]	76.5%
MetaObject-CNN[31]	78.9%
Semantic FV[15]	79.0%
LS-DHM[11]	83.75%
VSAD+FV+ Places205-VGGNet-16[14]	86.2%
Places401-Deeper-BN-Inception (B2)[10]	86.7%
SDO[17]	86.72%
MP[12]	86.9%
MFAFVNet+Places[13]	87.97%
Our PL+AI	87.24%
Our PL+AP+AI+IM	87.46%
Our PL+AI (weakly supervised)	87.84%
Our PL+AP+AI+IM (weakly supervised)	**88.06%**

TABLE VIII
COMPARISON WITH THE STATE-OF-THE-ART WORKS ON SUN397

Approaches	Accuracy
Attribute+SVM[46]	27.6%
Xiao *et al*. [35]	38.0%
MetaObject-CNN[31]	58.11%
Semantic FV[15]	61.72%
Places365+VGGNet16[30]	63.2%
LS-DHM[11]	67.56%
Human-level performance[35]	68.5%
Places401-Deeper-BN-Inception (B2)[10]	72.0%
MFAFVNet+Places[13]	72.01%
MP[12]	72.6%
VSAD+FV+ Places205-VGGNet-16[14]	73.0%
SDO[17]	73.41%
Adi-Red [9]	73.59%
Our PL+AI	72.21%
Our PL+AP+AI+IM	72.41%
Our PL+AI (weakly supervised)	73.85%
Our PL+AP+AI+IM (weakly supervised)	**74.12%**

VI. CONCLUSIONS

Although the existing scene recognition methods mostly adopt the combination of scene features and object features to achieve the improvement of recognition accuracy, due to the limitation of object features in local regions, the discrimination of the merged features is limited. In our work, we enhance the discrimination of similar local regions by exploiting the basic visual features scene attributes. In the contextual modeling framework, we explore the visual features of scene attribute, object and scene in a common semantic space, and utilize the complement of visual features and their contextual spatial relations, and integrate them into a contextual model to obtain discriminative scene representations. The state-of-the-art performances on two challenging standard benchmarks evaluate the effectiveness of combined scene representations and verify that scene attributes are beneficial for scene recognition.

ACKNOWLEGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018 and 61902378, in part by Beijing Natural Science Foundation under Grant L182054 and Z190020, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals, in part by the National Postdoctoral Program for Innovative Talents under Grant BX201700255, and in part by China Postdoctoral Science Foundation under Grant 2018M631583.

REFERENCES

[1] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
[2] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005.
[3] L. Li, H.Su, E. Xing, and F. Li. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.
[4] L. Zhang, X. Zhen, and L. Shao. Learning object-to-class kernels for scene classification. *IEEE Trans. on Image Process.*, 23(8):3241–3253, Aug. 2014.
[5] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, 34(5):902–917, 2012.
[6] N. Rasiwasia and N. Vasconcelos. Latent dirichlet allocation models for image classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, 35(11):2665–2679, 2013.
[7] X. Li and Y. Guo. Multi-level adaptive active learning for scene classification. In *ECCV*, pages 234–249, 2014.
[8] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, pages 359–372, 2012.
[9] Z. Zhao and M. Larson. From volcano to toyshop: Adaptive discriminative region discovery for scene recognition. In *ACM MM*, pages 1760–1768, 2018.
[10] 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. on Image Process.*, 26(4):2055–2068, Apr. 2017.
[11] S. Guo, W. Huang, L. Wang, and Y. Qiao. Locally supervised deep hybrid model for scene recognition. *IEEE Trans. on Image Process.*, 26(2):808–820, Feb. 2017.
[12] X. Song, S. Jiang, and L. Herranz. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. *IEEE Trans. on Image Process.*, 26(8):2721–2735, June. 2017.
[13] Y. Li, M. Dixit, and N. Vasconcelos. Deep scene image classification with the mfafvnet. In *ICCV*, pages 5757–5765, 2017.
[14] Z. Wang, L. Wang, Y. Wang, B. Zhang, and Y. Qiao. Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *IEEE Trans. on Image Process.*, 26(4):2028–2041, Apr. 2017.
[15] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In *CVPR*, pages 2974–2983, 2015.
[16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.
[17] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou. Scene recognition with objectness. *Pattern Recognition*, 74:474–487, Feb. 2018.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–14, 2015.
[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.
[21] G. Huang, Z. Liu, L. Maaten, and K. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017.
[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
[23] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Trans. on Multimedia.*, 18(12):2528–2536, Dec. 2016.
[24] S. Xie and H. Hu. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Trans. on Multimedia.*, 21(1):211–220, Jan. 2019.

[25] J. Zhu, S. Liao, Z. Lei, and S. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image Vision Comput.*, 58:224–229, 2017.

[26] H. Li, H. Wu, S. Lin, L. Lin, X. Luo, and E. Izquierdo. Boosting zero-shot image classification via pairwise relationship learning. In *ACCV*, pages 85–99, 2016.

[27] S. Jiang, W. Min, and S. Mei. Hierarchy-dependent cross-platform multi-view feature learning for venue category prediction. *IEEE Trans. on Multimedia*, 21(6):1609–1619, 2019.

[28] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re-identification in urban surveillance videos. In *ICME*, pages 1–6, 2016.

[29] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: objects, scales and dataset bias. In *CVPR*, pages 571–579, 2016.

[30] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018.

[31] R. Wu, B. Wang, W. Wang, and Y. Yu. Harvesting discriminative meta objects with deep cnn features for scene classification. In *ICCV*, pages 1287–1295, 2015.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015.

[33] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, pages 392–407, 2014.

[34] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009.

[35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[36] X. Song, S. Jiang, L. Herranz, Y. Kong, and K. Zheng. Category co-occurrence modeling for large scale scene recognition. *Pattern Recognition*, 59:98–111, Nov. 2016.

[37] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, 2012.

[38] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[39] J. Sanchez F. Perronnin and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.

[40] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, pages 433–440, 2007.

[41] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[42] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[43] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 36(3):453–465, Mar. 2014.

[44] O. Russakovsky and F. Li. Attribute learning in large-scale datasets. In *ECCV*, pages 1–14, 2010.

[45] G. Patterson and J. Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, pages 85–100, 2016.

[46] G. Patterson, C. Xu, and H. Su. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[47] A. Oliva and A. Torralba. Scene-centered description from spatial envelope properties. In *BMCV*, pages 263–272, 2002.

[48] X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *ICCV*, pages 2120–2127, 2013.

[49] L. Liu Y.Long, L. Shao, F. Shen, G. Ding, and J. Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR*, pages 6165–6174, 2017.

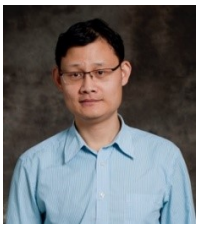
[50] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[51] D. Zhang, X. Chen, and W. S. Lee. Text classification with kernels on the multinomial manifold. In *SIGIR*, pages 266–273, 2005.

Haitao Zeng received the B.E. degree in school of geomatics, Shandong University of Science and Technology, Qingdao, China, in 2017. And he is currently a graduate student in computer science at School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, China, and he is also an intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, and image processing.

Xinhang Song received the B.S. degree in school of computer and information technology Beijing Jiaotong University, Beijing, China, in 2011, and the Ph.D. degree in computer science at Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 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 PC or TPC member for well-known conferences, such as IJCAI, AAAI and ACM Multimedia.

Gongwei Chen received the B.E. degree in School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China, in 2016. And he is currently a PhD student in computer science at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, machine learning, and image processing.



Shuqiang Jiang (SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS), Beijing and a professor in 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 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE and CCF, member of ACM, Associate Editor of IEEE Multimedia, Multimedia Tools and Applications. He is the vice chair of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He is the general chair of ICIMCS 2015, program chair of ACM Multimedia Asia2019 and PCM2017. 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.