



Scene Recognition with Comprehensive Regions Graph Modeling

Haitao Zeng^{1,2}(✉) and Gongwei Chen²

¹ China University of Mining and Technology, Beijing 100083, China

² Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
{haitao.zeng,gongwei.chen}@vip1.ict.ac.cn

Abstract. Learning the regional contents of scenes comprehensively is key to scene recognition. Due to semantic diversity and spatial complexity in scene images, modeling based on these regional contents is challenging. The current works mainly focus on some small and partial regions of the scene, while ignoring the majority region of the scene. In contrast, we propose the Semantic Regional Graph modeling framework for the comprehensive selection of discriminative semantic regions in scenes. To explore the relations of these regions, we propose to model these regions in geometric aspect based on the graph model, and generate the discriminative representations for scene recognition. Experimental results demonstrate the effectiveness of our method, which achieves state-of-the-art performances on MIT67 and SUN397 datasets.

Keywords: Scene recognition · Graph Neural Network

1 Introduction

The goal of scene recognition is to predict scene labels for images. Scene recognition is a challenging task of computer vision since the scene images are composed of various regional contents (e.g. foreground and background) with highly flexible spatial layouts. This characteristic determines that extracting the discriminative information of scenes requires the comprehensively learning of regional contents. Therefore, how to model these regional contents to obtain consistent visual representations is becoming the main challenge in the filed of scene recognition.

Some earlier methods [7, 27, 30] propose to model the local regional representations with BOW (Bag of Words) encoding for scene recognition. With the developments of Convolution Neural Networks (CNNs) [1, 9, 14], some scene recognition methods [11, 24, 26, 32, 34, 37] propose to learn regional features with the CNN models. These methods can be divided into two branches: some methods [24, 26, 32] propose to extract CNN features on local patches, which are annotated with the image-level label, and trained in weak supervision, leading to the ambiguity and noise in training. While, some other methods [34, 37] attempt to generate region proposal to locate the object regions for feature extraction,

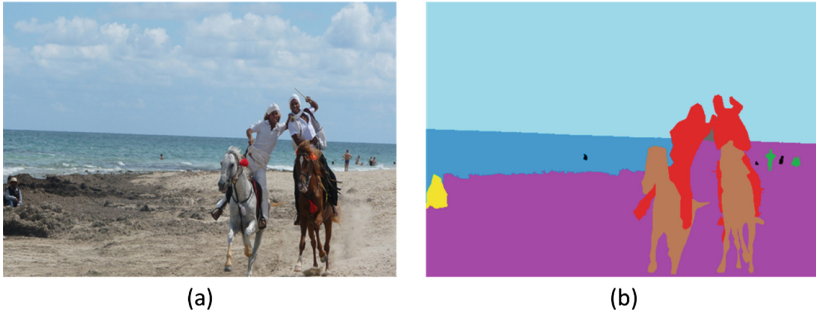


Fig. 1. (a) An example image, (b) its annotations in COCO-Stuff dataset

which are then fed to the followed networks for classification. However, considering the characteristics of scene image, the object based methods still have the limitation, since the object regions can only cover the relatively small and partial area of the scene, while the majority area of the scene is ignored, which may decrease the performance of scene recognition. In contrast, our motivation is to obtain more comprehensive information in the scenes.

Obtaining more comprehensive information in the scenes requires that the extracted region information have diversity. In addition to objects, the scenes usually consist of a much larger area of “stuff” (amorphous background regions, e.g. *sea*, *sand*, and *sky*), which also contain discrimination to different scenes. In our work, we propose to obtain discriminative regional information based on the stuff, since stuff covers a wider area, and it is essential to determine the scene category (e.g. as shown in Fig. 1 *sky*, *sand* and *sea* are the imperative elements in the beach category). Moreover, the object based works [34, 37] also inspire us that the object regions can also provide discriminative information. Therefore, in our work, we take both object and stuff into account as the discriminative semantic regions, and learn the relation of these semantic regions to generate discriminative representations.

In this paper, we propose a semantic regional graph modeling (SRG) framework for scene recognition. To perform scene recognition, we first feed an image into the pre-trained semantic segmentation network (e.g. DeeplabV2 [4] pre-trained on COCO-Stuff [3]) to generate the label map that has the same resolution as the input image. To obtain the information of scenes comprehensively, we implement three region selection methods on the label map to select the discriminative semantic regions, including the region of stuff and object. We extract these regional representations based on a pre-trained CNN through *RoIAlign* [13]. These regional features are concatenated together as the node representations of graph convolution network [15] (GCN). And we propose to learn the relations between these regions on the geometric aspects through GCN, which is used to optimize the corresponding node representations. Finally, we feed these optimized representations into classifier to predict scene labels. We conduct several experiments on MIT67 [20] and SUN397 [36], the experimental results illustrate the effectiveness of the proposed method.

2 Related Works

In this section, we briefly review the works that related to our topic in several aspects. The differences and connections of these works with ours are also being argued.

2.1 Scene Recognition

Scene recognition is an essential domain in computer vision. In some early works [23, 28], the basic visual elements (e.g. color, shape, and texture) play an important role in learning the global features of images. However, since scenes are relatively abstract, scene images are generally composed of multiple semantic regions. Thus, some works [7, 8, 12, 25, 27, 30] propose to perform scene recognition based on local region features. Lazebnik *et al.* [27] present the Spatial Pyramid Matching (SPM) which divides the image into several local sub-regions, extracts the feature on each region, and then concatenates the features of all sub-regions to predict the image label. Additionally, Perronin *et al.* [7] propose to use Fisher Vector (FV) to encode local handcrafted features (e.g. SIFT [17]) for scene recognition. Alternatively, Song *et al.* [25] propose to exploit multiple local features with context modeling, and also propose to embed multi-feature in semantic manifold.

Recently, the deep learning methods have made great impacts in some fields of computer vision, such as image recognition [1], object detection [21] and semantic segmentation [4]. Hence, some recent scene recognition works propose their methods based on the convolution neural networks (CNNs), and sharply improve the performances. Zhou *et al.* [2] present a massive scene-centric dataset Places that generate better generalization than object-centric dataset (e.g. ImageNet [22]). However, due to the structure of CNNs, some discriminative regional contents might be discarded during training. To deal with this problem, some methods propose to learn regional features. Wang *et al.* [32] propose PatchNets which is trained in weak supervision. During the training process, images are cropped into several patches and annotated with their image-level label. Song *et al.* [25] propose to embed multi-scale regional features with a hierarchical context modeling method. Wu *et al.* [34] propose to use the region proposal method to detect the discriminative object regions in the image to guide the scene recognition. In contrast to the current methods, we extract both object and stuff features as the discriminative semantic regions, and model the relations of these semantic regions in the geometric aspect through the graph network.

2.2 Graph Neural Network

Inspired by the impact of Graph Neural Network (GNN) in processing the non-Euclidean data, some recent works [5, 15, 18, 35, 39] in the computer vision have also employed the GNN to improve the performance, such as multi-label prediction [18], zero-shot recognition [35], fine-grained image recognition [5] and 3D human pose regression [16]. Yang *et al.* [39] develop an attentional graph

convolutional network to implement scene graph generation by upgrading the nodes in both visual and semantic features. While we also employ GCN [15] to upgrade the node representations, the graph we constructed is for each image, and with geometric information, thus, the relation between regions can be better captured, and the discriminative information can be preserved.

3 Semantic Regional Graph Model

The semantic regional graph modeling framework (SRG) including a semantic region selection module, a graph modeling network module, and a scene classification module. The architecture of our framework is illustrated in Fig. 2.

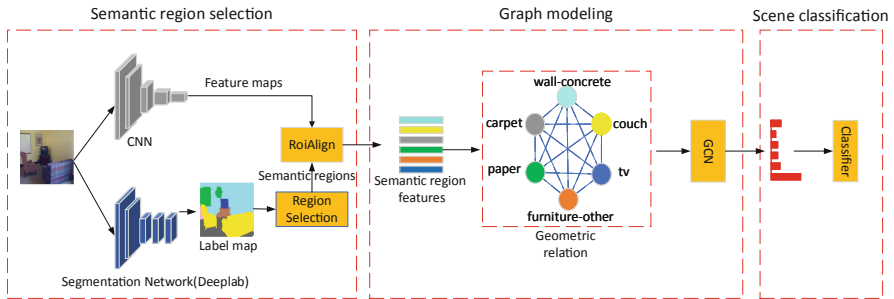


Fig. 2. The framework of SRG, which includes a semantic region selection module to determine the discriminative semantic regions, and a graph modeling module to learn the relation of semantic regions in geometric aspect, and a scene classification module to conduct classification. (*couch*, *tv* and *paper* are the object semantic regions, *wall-concrete*, *furniture-other* and *carpet* are the stuff semantic regions)

3.1 Semantic Region Selection Module

Generally, both stuff and object regions can provide discriminative information. To obtain these semantic regions comprehensively, abundant annotation is required. In our work, we adopt the COCO-Stuff [3] dataset, which contains 91 stuff categories and 80 object categories. Since the COCO-Stuff [3] dataset is a semantic segmentation dataset, we propose to implement our method based on the semantic segmentation network.

Given an image I , we feed it into a pre-trained semantic segmentation model (e.g. DeeplabV2 pre-trained on the COCO-Stuff), and obtain a label map $S \in R^{H \times W}$ as output. The label map S has the same resolution of input image I . The value S_{ij} of the pixel (i, j) in S represents the predicted category of its counterpart in I . For each category c , we can define a category binary map S^c based on S , which can be formalized as:

$$S^c_{ij} = \begin{cases} 1, & S_{ij} = c \\ 0 & S_{ij} \neq c \end{cases}$$

In practice, some category-wise binary maps will have no or few positive pixel (the value of the pixel is 1). These maps bring useless or noise information about desired semantic regions. So we set a threshold T to filter them. First, we count the number P^c of positive pixels in each category map S^c . Then, we select a new subset $\{\bar{c} \mid P^{\bar{c}} > T\}$ of categories.

Based on the binary map $S^{\bar{c}}$, we generate the connected components as semantic regions by applying the algorithm in [33]. By performing the same operation on all selected category-wise binary maps, we can obtain the set \mathbf{R} of semantic regions. Each item r in \mathbf{R} corresponds to a semantic region, and contains two elements $r = [r^1, r^2]$, where $r^1 = \{x, y, w, h\}$ contains the coordinate of central point and width and height of this region and r^2 denotes the predicted category of this region. To determine the discriminative semantic regions, we design several region selection methods:

Maximum Region (MR): The simplest selection method only consider the area of regions. Given region r , the area of region can be computed by r^1 . Then, we choose top N regions which are listed in descending order by area. We define the operator $\mathcal{S}(\cdot)$ to represent this selection process. The selected region set V is obtained by,

$$V = \mathcal{S}(\mathbf{R}, N)$$

Category guided Maximum region (CM): Considering the semantic diversity, we propose another selection method by considering the category information r^2 of region. To address the issue that many large regions in \mathbf{R} belong to a few categories, we choose the maximum region of each category in \mathbf{R} to form a new region set \mathbf{R}^{cm} . Then the operator $\mathcal{S}(\cdot)$ is performed on \mathbf{R}^{cm} to obtain the selected region set V .

Category guided Union (CU): To bring abundant and useful information, another selection operation is based on the union of regions within same predicted category. We compute the union of regions of every category \bar{c} , and use the union as the element to form a new region set \mathbf{R}^{cu} . Then the operator $\mathcal{S}(\cdot)$ is performed on \mathbf{R}^{cu} to obtain the selected region set V .

After obtaining the discriminative semantic region set V , we extract the local representations of regions through a pre-trained CNN. For each region v_i in V , we can use the coordinate of central point and width and height with *RoIAlign* [13] operation to generate the representation $x_i \in R^d$ of this region. To make use of global information, we regard the image as a global region with the geometry information $\{x = W/2, y = H/2, W, H\}$ and add it into the region set V . Finally, the region representation matrix $X \in R^{(N+1) \times d}$ is obtained.

3.2 Graph Modeling Module

In order to model these regions, we reorganize them in form of graph and perform GCN [15] to capture the discriminative relation between regions. Unlike

the conventional convolutions, the GCN is operated on the non-Euclidean data, which requires to learn a specific function $f_{gcn(\cdot)}$

$$X^{(t+1)} = f_{gcn}(X^{(t)}, A) \quad (1)$$

where $X \in R^{N \times d}$ (N indicates the number of regions, and d denotes the dimension of region representation) is the region representation matrix and $A \in R^{N \times N}$ is the corresponding adjacency matrix (we will discuss the construction process of A later). When applying the convolution operation [15], the function f_{gcn} can be formalized as:

$$\bar{X}^{(t+1)} = \eta(\tilde{\theta}^{-\frac{1}{2}} \tilde{A} \tilde{\theta}^{-\frac{1}{2}} \bar{X}^{(t)} W^{(t)}) \quad (2)$$

where $X^{(t+1)} \in R^{N \times d}$ denotes the optimized representations of regions, and $\tilde{A} = A + I_N$, $\theta_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix of \tilde{A} , $W^{(t)}$ denotes the trainable weight matrix. $\eta(\cdot)$ is the non-linear activation function ReLU.

To optimize the node representations on the graph. We need to extract the local representation set $X = \{X_1, \dots, X_i, \dots, X_N\}$, and construct the adjacency matrix A . Since we have extracted the node representations that based on the *RoIAlign* [13]. Therefore, we only discuss the way of constructing the adjacency matrix A .

Geometric Relation: To understand the connection of each node representation on the graph, we construct the adjacency matrix A . Since the impact of geometric relation in scenes is heavily. Thus, in order to model the relation of semantic regions, we define the geometric representation based on each region, and construct the corresponding geometric adjacency matrix. For a pair of regions v_i and v_j in region set V , a 4-dimensional relative geometric feature is produced, as

$$\left(\log \left(\frac{|x_i - x_j|}{w_i} \right), \log \left(\frac{|y_i - y_j|}{h_i} \right), \log \left(\frac{w_j}{w_i} \right), \log \left(\frac{h_j}{h_i} \right) \right)$$

Then, this feature is embedded into a high-dimensional (d_s -dim) representation O_{ij} by performing method in [29]. The embedded feature is projected by $W_o \in R^{d_s \times 1}$ into a scalar, which can be represented as:

$$\begin{aligned} \alpha_{ij} &= O_{ij} W_o, \\ A^{gr} &= \text{softmax}(\alpha) \end{aligned} \quad (3)$$

After constructing the adjacency matrix, we can apply the graph convolution network in Eq.(6) to update the regional representation $X^{(t)}$, and generate the updated $X^{(t+1)}$.

3.3 Scene Classification Module

To prevent over-fitting, we only adopt one-layer GCN. After the operation of graph modeling, we obtain the final region representations X^1 , then use the global region representation X_1^1 as image representation. Finally, the image representation is fed into an one-layer fully connected network for classification.

4 Experiments

In this section, we introduce the experimental details of our SRG. And we design several experiments, to evaluate the performance of SRG on two widely used scene recognition benchmarks, MIT67 [20] and SUN397 [36].

4.1 Experimental Datasets

MIT67: There are 67 indoor scene categories and 15,620 images. Each category contains at least 100 images. For evaluation experiments, each category contains 80 images for training and 20 images for test following the original protocol.

SUN397: There are 397 categories and 108,754 images in this dataset. Following the original paper, we divide 50 images for training and 50 images for test. Due to this dataset is relatively large, evaluating on this dataset is challenging.

4.2 Implementation Details

In the semantic region selection module, we adopt the DeeplabV2 [4] pre-trained on the COCO-Stuff [3] as our basic segmentation model. The resolution of the input image is fixed as 448×448 , which leads to 448×448 label map. Based on this map, we select the discriminative regions of the image by our region selection methods, in which the threshold T is 0.01, and the selected number N of selected discriminative regions is determined on the statistics of the distribution of the number of regions, which are shown in Fig. 3. The mean values of the two benchmarks are *15.61* and *10.46*, respectively. Thus, the number of regions we selected in MIT67 and SUN397 are *16* and *10* respectively (if the number of semantic regions in some images is lower than N , we fill the selected region set with fake regions, whose representations are denoted by zeros, and geometric information is $\{x = 0, y = 0, W = 1, H = 1\}$). Then, we extract these representations based on Res50-PL model (ResNet50 [14] model pre-trained on the Places365). The initial region representation matrix is $(N + 1) \times 2048$.

In the graph modeling module, we adopt one layer GCN to upgrade the node representations. The initial node representations are regularized with the L2 regularization factor, then fed into our graph model. In the training phase, we train our models for 20 epochs with the batch size of 32 and Adam optimizer, and the initial learning rate is set to 0.001, and is divided by 10 at 10/15/18th epoch. On the two benchmarks, the hidden layer units in graph convolution are 4096, and 8192, for MIT67 and SUN397 datasets respectively. We use omit regularization (dropout) in our final classifier with a rate of 0.5.

After graph modeling, we obtain the final region representations. We only adopt one layer GCN to upgrade, and use the global region representation as image representation to conduct scene classification, which can prevent the impact of the fake region representations upgrade.

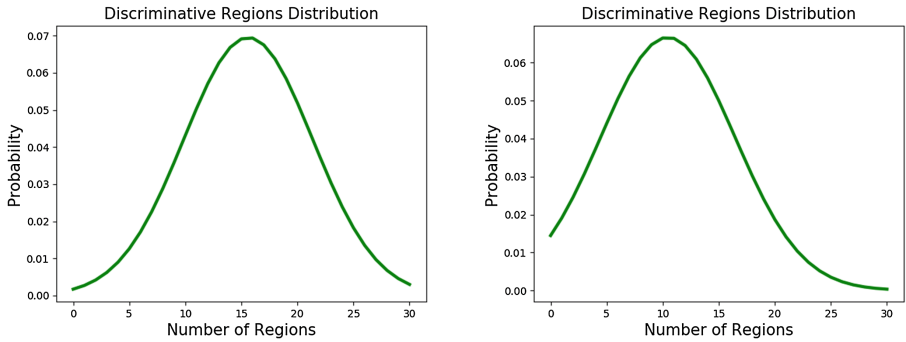


Fig. 3. The distribution of discriminative regions in MIT67 and SUN397.

4.3 Results

In this subsection, we conduct several experiments to evaluate the performance of our approach. The classification results of the linear SVM are set as the **baselines**, whose inputs are initial global region representations.

Effectiveness of Different Region Selections. In the semantic region selection module, we set three region selection methods, such as Maximum Region (MR), Category guided Maximum region (CM) and Category guided Union (CU). We conduct some detailed experiments in Table 1, and analyze the effectiveness of three selection methods. In Table 1, it can be noticed that three region selection methods have achieved higher results than the baselines, which demonstrates the effectiveness of our region selection method. In addition, we can observe that CM performs better than MR, which indicates when selecting regions based on the semantic meanings, more discriminative information of the image can be learned. Moreover, the slightly lower performance of CU demonstrates that selecting the union of regions may result in redundancy. Therefore, it's essential to ensure the diversity of semantics and avoid redundant information when selecting discriminative semantic regions.

Table 1. Comparisons of different region selection methods

Baseline	Region selection			MIT67(%)	SUN397(%)
	MR	CM	CU		
✓	–	–	–	86.87	71.53
–	✓	–	–	87.99	74.03
–	–	✓	–	88.13	74.06
–	–	–	✓	87.69	73.89

Moreover, three region selection methods are based on the same graph modeling. In Table 1, it can be noted that our best results are 1.26% and 2.53% over baselines. This confirms that the effectiveness of modeling the geometric relation between discriminative semantic regions, which can boost the performance of scene recognition.

The Effectiveness of Different Kinds of Semantic Regions. To determine the effectiveness of different kinds of semantic regions, we construct the following experiments. We divide the semantic regions into different sets, including stuff and object sets. According to the statistics, the number N of selected regions is $4/12$ (object/stuff) in MIT67, $2/8$ (object/stuff) in SUN397. We select these regions based on the CM region selection method. In Table 2, we can observe that the stuff and object are both over the baselines when the number of regions is equal, which indicates that we can obtain discriminative information from both stuff and object regions. When enlarging the number of stuff regions, there are still improvements. Furthermore, when considering both stuff and object regions, the improvement of performances are also obvious, which demonstrates that object and stuff regions can provide complementary information. Thus, obtaining comprehensive information of scene images can improve the performances of scene recognition.

Table 2. Comparisons of different semantic regions.

Regions	Object	Stuff	Object & Stuff	MIT67(%)	SUN397(%)
2	✓	–	–	–	71.92
2	–	✓	–	–	72.41
8	–	✓	–	–	73.4
10	–	–	✓	–	74.06
4	✓	–	–	87.01	–
4	–	✓	–	87.09	–
12	–	✓	–	87.46	–
16	–	–	✓	88.13	–
Baseline	–	–	–	86.87	71.53

4.4 Comparison with State-of-the-Art Methods

We compare our SRG with state-of-the-art methods. The results are shown in Table 3. It can be observed that our SRG outperforms the current state-of-the-art methods, confirming the effectiveness of our method. Compared with the region based works [24, 32, 34, 37], our SRG achieves the best performance, which demonstrates the effectiveness of our method. To the best of our knowledge, our SRG obtains state-of-the-art performance in the domain of scene recognition.

Table 3. Comparisons of our method with state-of-the-art methods

Approaches	MIT67(%)	SUN397(%)
Places365+VGGNet16 [2]	76.5	63.2
MetaObject-CNN [34]	78.9	58.11
MLR+CFV+FCR1-w [37]	82.24	64.53
LS-DHM [10]	83.75	67.56
VSAD+FV+Places205-VGGNet-16 [32]	86.2	73.0
PowerNorm [19]	86.3	–
Places401-Deeper-BN-Inception (B2)[31]	86.7	72.0
SDO [6]	86.72	73.41
MP [24]	86.9	72.6
MFAFVNet+Places [38]	87.97	72.01
Adi-Red [40]	–	73.59
Our SRG	88.13	74.06

5 Conclusion

In this paper, we propose our semantic regional graph modeling framework for scene recognition. To select the discriminative semantic regions in the scene comprehensively, we conduct several region selection methods, effectively capturing the discriminative semantic regions, ensuring the semantic diversity and avoiding redundancy. In the graph learning module, we optimize the region representations in the relation of geometric aspects, and generate the discriminative scene representations. The exploration of stuff and object regions also demonstrates the complementarity of them. Based on the comprehensive semantic regions, our method can obtain state-of-the-art performances on MIT67 and SUN397 datasets.

References

1. Sutskever, I., Krizhevsky, A., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114 (2012)
2. Bolei, Z., Aditya, K., Agata, L., Antonio, T., Aude, P.: Places: an image database for deep scene understanding. arXiv preprint [arXiv:1610.02055](https://arxiv.org/abs/1610.02055) (2016)
3. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: thing and stuff classes in context. In: CVPR (2018)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018)
5. Chen, T., Lin, L., Chen, R., Wu, Y., Luo, X.: Knowledge-embedded representation learning for fine-grained image recognition. In: IJCAI (2018)

6. Cheng, X., Lu, J., Feng, J., Yuan, B., Zhou, J.: Scene recognition with objectness. *Pattern Recogn.* **74**, 474–487 (2018)
7. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010*. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
8. Fredembach, C., Schroder, M., Susstrunk, S.: Eigenregions for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(12), 1645–1649 (2004)
9. Maaten, L., Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. In: *CVPR* (2017)
10. Guo, S., Huang, W., Wang, L., Qiao, Y.: Locally supervised deep hybrid model for scene recognition. *IEEE Trans. Image Process.* **26**(2), 808–820 (2017)
11. Heranz, L., Jiang, S., Li, X.: Scene recognition with CNNs: objects, scales and dataset bias. In: *CVPR* (2016)
12. Jiang, S., Chen, G., Song, X., Liu, L.: Deep patch representations with shared codebook for scene classification. *ACM Trans. Multimedia Comput. Commun. Appl.* **15**, 1–17 (2019)
13. Gkioxari, G., He, K., Dollár, P., Girshick, R.: Mask R-CNN. In: *ICCV*, pp. 2980–2988 (2017)
14. Ren, S., He, K., Zhang, X., Sun, J.: Deep residual learning for image recognition. In: *CVPR*, pp. 770–778 (2016)
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: *ICLR* (2017)
16. Peng, X., Zhao, L., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3D human pose regression. In: *CVPR* (2019)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* **60**, 91–110 (2004)
18. Min, Z., Wei, C.X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: *CVPR* (2019)
19. Koniusz, P., Zhang, H.: A deeper look at power normalizations. In: *CVPR* (2018)
20. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: *CVPR*, pp. 413–420 (2009)
21. Ren, S., He, K., Girshick, R., Sun, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *NIPS* (2015)
22. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252 (2015)
23. Shen, J., Shepherd, J., Ngu, A.H.H.: Semantic-sensitive classification for large image libraries. In: *International Multimedia Modelling Conference*, pp. 340–345 (2005)
24. Song, X., Jiang, S., Herranz, L.: Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. *IEEE Trans. Image Process.* **26**(8), 2721–2735 (2017)
25. Song, X., Jiang, S., Herranz, L.: Joint multi-feature spatial context for scene recognition on the semantic manifold. In: *CVPR* (2015)
26. Song, X., Jiang, S., Herranz, L., Kong, Y., Zheng, K.: Category co-occurrence modeling for large scale scene recognition. *Pattern Recogn.* **59**, 98–111 (2016)
27. Schmid, C., Lazebnik, S., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *CVPR* (2006)
28. Vailaya, A., Jain, A., Figueiredo, M., Zhang, H.: Content-based hierarchical classification of vacation images. In: *IEEE International Conference on Multimedia Computing and Systems*, pp. 518–523 (1999)

29. Vaswani, A., et al: Attention is all you need. In: NIPS (2017)
30. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
31. Wang, L., Guo, S., Huang, W., Xiong, Y., Qiao, Y.: Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs. *IEEE Trans. Image Process.* **26**(4), 2055–2068 (2017)
32. Wang, Z., Wang, L., Wang, Y., Zhang, B., Qiao, Y.: Weakly supervised patchnets: describing and aggregating local patches for scene recognition. *IEEE Trans. Image Process.* **26**(4), 2028–2041 (2017)
33. Mark, J.B., Wilhelm, B.: *Principles of Digital Image Processing: Core Algorithms*. UTICS. Springer, London (2009). <https://doi.org/10.1007/978-1-84800-195-4>
34. Wu, R., Wang, B., Wang, W., Yu, Y.: Harvesting discriminative meta objects with deep CNN features for scene classification. In: ICCV (2015)
35. Ye, Y., Wang, X., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: CVPR (2018)
36. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN Database: large-scale scene recognition from abbey to zoo. In: CVPR, pp. 3485–3492 (2010)
37. Xie, G.-S., Zhang, X.-Y., Yan, S., Liu, C.-L.: Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. *IEEE Trans. Circuits Syst. Video Technol.* **27**, 1263–1274 (2017)
38. Vasconcelos, N., Li, Y., Dixit, M.: Deep scene image classification with the MFAFVNet. In: ICCV (2017)
39. Yang, J., Lee, S., Lu, J., Batra, D., Parikh, D.: Graph R-CNN for scene graph generation. In: ECCV (2018)
40. Zhao, Z., Larson, M.: From volcano to toyshop: adaptive discriminative region discovery for scene recognition. In: ACM MM (2018)