2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems

Multi-scale Convolutional Neural Networks For Scene Recognition

1st HanLing Zhang
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China
jt_hlzhang@hnu.edu.cn

2st Yi Zheng
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China
1611742878@qq.com

***Abstract*—In recent years, convolutional neural network (CNN) has performed well in a number of image classification tasks, but it hit a bottleneck on scene recognition task, due to the multilevel semantic information in a scene. This paper is dedicated to studying the deep learning methods in scene recognition task, and making contributes to improving the classification performance in the field of scene recognition, and an effective method that captures and fuses multi-level semantic information is proposed. First of all, we compare the differences between object classification task and scene recognition task in order to apply the successful replication of CNN in object classification task to scene recognition task after resolving the differences. Then we use a multi-scale learning method to capture different scale visual features at multiple levels. In addition, on the basis of multi-scale learning, we propose a method of feature fusion at the level of category, aiming to effectively combine different scale features. The experimental results show that the success of the object classification task can be applied to the scene recognition task by our method.**

Keywords—Pattern Recognition and Intelligent System, convolutional neural network, scene recognition.

I. Introduction

The human visual system can understand the world at a glance, not only to quickly distinguish object categories, but also to accurately recognize complex scenes. Computer scene recognition is a fundamental problem in computer vision, which aims to analyze the input image and distinguish the scene category of the image by simulating human vision system. A scene can provide rich semantic information that is useful for other related visual tasks: the semantic information contains objects and backgrounds, which can be used to define the context of detection task for object and text in [1,2], even can be applied to event recognition and motion recognition [3,4] with the combination of human common habits. In addition, in fields such as intelligent autonomous driving [5], combined with a clear scene, different driving scenarios for different conditions can improve efficiency and accuracy.

In recent years, Convolutional Neural Networks(CNN) have achieved remarkable success in various image classification and understanding tasks [6-10]. These deep CNN models directly learn discriminative visual representation from the original image in an end-to-end way. Compared with traditional hand-crafted feature, they have rich modeling capabilities and powerful visual representation, with a sufficient amount of data, the CNN can learn the ability of feature extraction and classification. With the update of computer hardware, GPU acceleration calculation makes up for the computational efficiency of neural network, the performance of neural network exceeds the traditional method with the help of large-scale datasets. Recently, ImageNet Large Scale Visual Recognition Competition [11] (ILSVRC) was over, on which the classification error rate in the LOC mission has been as low as the classification error of the human eye.

Compared with the success of the object classification task, the scene recognition task progresses relatively slowly, though both have some similar problems in image classification, such as illumination variation and object occlusion. The target of the scene recognition task is more challenging due to the more complicated content. The object recognition requires to identify these objects separately while the scene recognition task requires the judge image categories represented by the environment as a whole. At the level of definition, a scene represents a place, an environment, usually composed of a series of objects, which is richer in content than object recognition tasks. And it is worth noting that, in addition to the issue of the number of objects, the correlation among objects and background environment also bring difficulties for scene recognition.

On the whole, there are many difficulties in scene recognition, such as label ambiguity, visual inconsistency, objects, etc. What's more, CNN as a black box, is difficult to accurately describe what features CNN has learned, and whether these features are helpful for scene recognition. In order to solve these problems, we have done the following work: (1) We compare the two classification tasks and find that object is the key difference between object classification and scene recognition; (2) We propose a multi-scale feature fusion method to effectively combine different object features, so as to improve the accuracy in scene recognition. First, we find that there is a remarkable difference in the distribution of object in two datasets, which are respectively used for object classification and scene recognition. Specifically, there are more objects in scene, compared with datasets used for object classification task. Then we study the effect of object on scene recognition, we find that CNN also extracts object features in scene recognition task, which is like it used to do in object classification task. So we propose a training method called multi-scale feature fusion, we found that there are many objects with different scales in each scene image, and CNN lacks the invariance of significant scaling, it is difficult to extract effective scene features. The proposed method allows our network structure to capture image features with different scales. In addition, we found that different categories of scenes have different internal structures, so we train the network to combine the different scale features to adapt to this structural difference. Moreover, the method can also alleviate the issue of visual inconsistencies leading by rich categories. We choose AlexNet [6] and Inception [9] as the basic network architecture, demonstrating the effectiveness of multi-

This work was supported by the National Natural Science Foundation of China (Grant No.61672222, 61572183, 61472131), Hunan Natural Science Foundation (Grant No.2018JJ2054), Science and Technology Planning Project of Changsha, China (Grant No. kq1706021).

978-1-7281-2058-4/19/$31.00 ©2019 IEEE
DOI 10.1109/HPCC/SmartCity/DSS.2019.00333

resolution feature fusion method on three datasets for experiments, our method has improved performance in scene recognition. Finally, we present several failure cases by our method to highlight the existing challenges scene recognition and discuss possible future research direction.

The rest of the article is organized as follows. In Sec. 1, we compare the datasets and methods used in traditional pattern recognition and deep learning. Section 2 explores the differences between object classification and scene recognition. In Sec. 3, we introduce our multi-scale feature fusion training method and compare it with the previous work. In Sec. 4 we introduce the experimental details and results. Section 5 is the conclusion.

II. Related Work

CNN has been around for a long time, but only in recent years have they been successful, partly due to lack of enough training samples and the limited computing ability of hardware. In Sec. 1 we summarize the progress in datasets and algorithmic models for scene recognition work, and introduce the differences between our work and previous work.

A. From Object Dataset to Scene Dataset

No large-scale collection is not sufficient to train a complex CNN and deep learning performed not as well as traditional machine learning methods did, due to the lack of training samples in the early time. The PASCAL Visual Object Classes Challenge [12] (VOC) used real object area annotations to provide a standard image annotation dataset and standard evaluation system for detection algorithms and learning performance. ImageNet dataset [13] had been widely applied in the field of deep learning image, including image classification, positioning and detection etc., and the visual task error rate was lower than human vision in ILSVRC2017. However, scene recognition is still rich in challenges and Places Challenge just started. Ariadna Quattoni proposed Indoor67 dataset in [14] so as to evaluate the works on the indoor scene recognition. And a wide range of scene understanding dataset SUN to define the concept of the scene was proposed in [15]. Bolei Zhou proposed the Places dataset, which became the largest set of scene data in the world [16]. In addition, Bolei Zhou also released a densely annotated dataset ADE20K dataset, which constructs a benchmark platform for scene analysis in [17]. Our work will compare the difference between the scene dataset and the object dataset in the second section, and we validated our method on the largest scene dataset Places and tested it on the indoor scene dataset Indoor67.

B. Scene Recognition Method

Besides deep learning, the hand-crafted feature was a popular method for image processing task, which is also applied to the field of scene recognition. The bag of words is the most commonly used method for image research [18], and spatial pyramid matching [19] was proposed to combine spatial layout into a word bag representation for scene recognition. Gist [20] is a well-known scene recognition feature that captures spatial layout and high efficiency in scene recognition and there are other feature representations in [21, 22].

Since AlexNet won the ILSVRC2012, more and more research focuses on the use of CNNs to deal with image processing task, including scene recognition. Bolei Zhou proposed a new scene-centric dataset Place for eliminating dataset bias, and showed the object detection effect of CNN in scene recognition task in [23]. Wang proposed the use of multi-resolution CNN for scene recognition in [24]. Luis Herranz also studied how CNN effectively combines scene-centric and object-centric knowledge in [25]. Different from previous studies, our proposed method can capture feature information with different scales in a scene and reduce the dataset bias. In addition, we don't only extract different scales of features in the feature extraction stage, but also configure optimal feature combinations for different categories of scenes in the classifier.

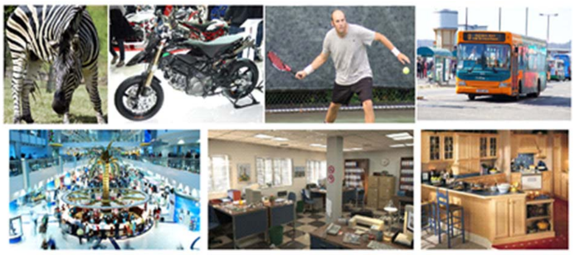

Fig. 1. Picture examples from ImageNet dataset and MIT Indoor67 dataset.

III. Object and Scene

Deep learning has achieved excellent results in object classification task, and scene recognition task is similar to the object classification task somehow, so we seek a method to improve scene recognition. In this section, we first explore the difference in datasets used for the two tasks, then introduce the impact of object in the image on scene recognition, and finally propose an improvement scheme.

A. Data Difference

Training CNN requires massive data support, and understanding the differences in the datasets involved in scene recognition task and object classification task can better explain the reason for their different performance. Datasets commonly used for object classification tasks include Pascal VOC, ImageNet, and datasets of scene recognition tasks are represented by MIT Indoor67 and Places. Our research found that the main difference between these datasets lies in the distribution of objects, which is represented by the number of objects and the scale of objects.

We randomly selected some pictures from the two datasets for comparison, as shown in Fig. 1. Since a scene usually represents an environment, it is not only composed of a series of objects, but also includes different background. Therefore, the content of scene image is richer and more complex than that of object image. There is usually only a single significant target in the object image, while there are many different types of objects in the scene image. We further study the object in the distribution of datasets, by parsing the annotation file of dataset, Table. I shows the average number of each object per image. The number of object is usually less than 3 while there are more than 7 objects in scene image, the number can even reach 19 in indoor scene. It is quite obvious that the distribution of object on the two datasets varies greatly in number.

TABLE I. Distribution of object in various datasets

Dataset	Object categories	Number of objects on average
ImageNet	1000	1.7
COCO	91	3.5
Indoor	8267	19
ADE20K	2693	7.9
SUN	4479	9.8

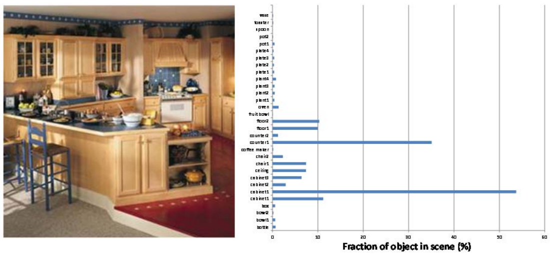

Fig. 2. Distribution of objects in kitchen scene.

As shown in Fig. 2, our further study found that the scale of object varies in the scene dataset. The picture in object dataset is usually shot from the perspective of close-up, and object lies in the center of the image with a large size, occupying more than three quarters of the whole size. On the other hand, the picture in scene dataset is shot from a long shot, and the distribution of objects is relatively scattered, some of them lie in the edge of the image and some lie in the middle position. There are a lot of objects in the scene, each object occupies a small proportion in the whole image on average, and different objects have different scales, ranging from large size to small size. We further study this scale feature of object, which is also related to the category of the scene.

B. The Impact of Object on Scene Recognition

The characteristics of object in quantity and scale are the main differences between the two datasets, so we further study the impact of object on scene recognition task. As an important part of scene, object can provide very discriminating information for scene recognition. For example, if there are seats in the scene image, such a scene is most likely the category of indoor, and if there are vehicles in the scene, it is most likely one of outdoor scene. Luis Herranz proposed in that object is the intermediate representation of CNN in Scene recognition task [23], which indicates that CNN still detects object information and extract the feature of objects in scene recognition task, just like it used to do in object classificationtask. Therefore, it can be inferred that object information is essential for scene recognition.

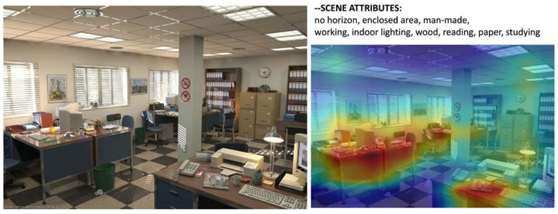
--SCENE ATTRIBUTES:
no horizon, enclosed area, man-made,
working, indoor lighting, wood, reading, paper, studying

Fig. 3. Class activation map and scene attributes.

Although the object is critical in the scene, it is not the only discriminative information. We cannot judge the scene category only according to the object category in the image, for example, the scene which contains desk, chairs and computers may be an office room while it can be a study room, too. The definition of a scene includes not only the various objects, but also the correlation between the objects and background, which can affect the category of the scene. Fig. 3 shows the class activation map and scene attributes, demonstrating global layout and correlation among objects are also important. We can distinguish the general category of the scene by the object information that usually appears in the scene, but to further improve the prediction, we need to combine the internal correlation of these objects.

CNN can autonomously learn the categories of objects in the scene and even the correlation between the objects through the massive dataset, but the CNN is also flawed—the CNN lacks the invariance of significant scaling. Compared with other visual classification tasks, the scale of object in the scene varies more, and the scene recognition task needs to process more scale information. For this reason, we propose multi-scale feature fusion of our method in the fourth part.

IV. Multi-scale Feature Fusion

In this section, we propose our method in order to reduce the impact from objects cause of its diversity appeared in scene image. Specifically, we use the multi-scale feature to represent multiple objects and fuse the features to match the correlation among objects, and we compare our method with the former multi-scale training network.

A. Multi-scale Feature

Generally, there are a variety of objects in a scene, which have different size and vary greatly. However, the inherent structure of CNN is difficult to effectively express the visual feature when the scale of object varies from a size to another size in an image. Therefore, in the process of feature extraction, we use a multi-scale network to capture the features of multiple objects. We define the feature as small-scale feature and large-scale feature that represent small object and large object, respectively.

Small-scale feature is the fine information of image, representing local details of a scene, which the network gets by using a relatively small receptive field in convolution calculation. Oppositely, large-scale feature represents coarse information in the scene, specifically the large-scale object or the overall structure of the scene. The size of receptive field

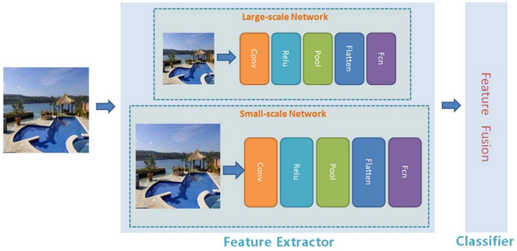

Fig. 4. We train CNN to extract different scale features by changing the resolution of training set as input in feature extractor, and fuse the features in the classifier.

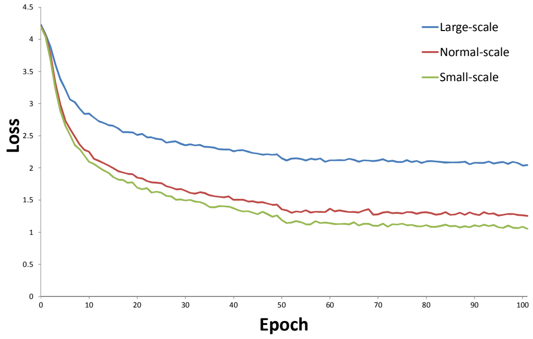

Fig. 5. Loss of CNN with different scales in training process.

can be controlled by changing the resolution of the input image or the size of convolution kernel, we choose to vary the resolution of the input image without modifying the original structure of the network in order to prove the effectiveness of our method, as shown in Fig. 4.

In Fig. 5, the performance of network with small-scale is better than others, which is due to that small-scale features can be used to describe meaningful local details. On the other hand, the training time of network with large-scale is shorter while small-scale increases the cost of computation relatively. Each of them has its own advantage and disadvantage, a powerful scene deep network should be able to capture multi-scale features so as to describe multi-level visual concepts. Our method trains CNN to learn different scale features by taking different resolution images as input, so that the network can describe various objects at different scales for scene recognition.

The idea of multi-scale training strategy has been proposed for a long time, VGGNet adopted multi-scale cropping for network training [7]. Different from multi-scale cropping, our method captures multi-level information from different resolutions while previous works all rely on a single resolution, and these multi-scale features will be applied in classifier together.

B. Feature Fusion

The similarity between different categories is one of existing challenges for scene recognition, because these scenes may contain the same objects, the correlation of which can determine the category of scene. We consider an effective way to combining the scale feature of various objects in the classifier so that CNN can learn the correlation by fusing the feature at the level of category, as shown in Fig. 6.

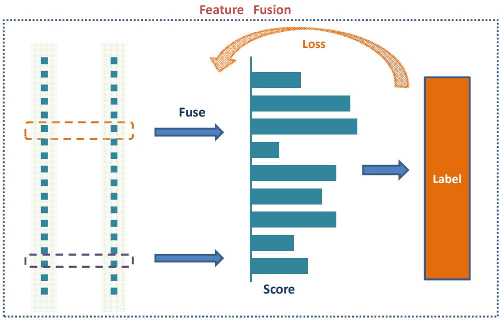

Fig. 6. Fuse different scale features in the classifier

We describe the recognition effect of CNN on scene category, as shown in Fig. 7. Although small-scale feature network performs better than large-scale in most categories, there are some exceptions. Our method considers the advantages of each scale so that the scale feature with a high score can complement the one with a low score in some categories. Specifically, we consider both precision and recall as benchmark, so we use F-measure:

$$F_{\beta} = (1+\beta^2)\frac{precision \cdot recall}{\beta^2 \cdot precison + recall}, \tag{1}$$

Where we take $\beta = 1$ and we get F_1 score as follows:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}, \tag{2}$$

The classifier is designed to fuse features with different scales at the level of category, we define it as follows:

$$\begin{cases} S_c = \dfrac{1}{A}\sum_{i=1}^{A} K_i(p_i \cdot w_i + b_i) \\ S = \{S_1, S_2, \cdots, S_c\}_{c=1,2,\ldots,N} \end{cases} \tag{3}$$

Where A represents the number of scales and N is the number of categories, p is the output of network with different scales, c is the category. We train the classifier to update k, w, b from the loss function where $y_n = S$:

$$-\frac{1}{N}\sum_{n=1}^{N}[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)] \tag{4}$$

The Multi-Resolution CNN has been proposed in [24], different from the work, we allow the network to learn the combination of different features at the level of category while the features are roughly computed as an arithmetic average in previous work.

V. Experiment

In this section, we introduce our experimental dataset and training method, and describe the performance of our method in the test, including MIT Indoor, SUN and Places. Finally, we present several failure examples from our method and discuss the possible reasons.

A. Datasets

We evaluate the proposed method with three widely used scene benchmarks. MIT Indoor contains 67 categories, each of which has 80 indoor images for training. Indoor scenes tend to be rich in objects, which are more difficult for scene recognition task. SUN is widely accepted as benchmark for

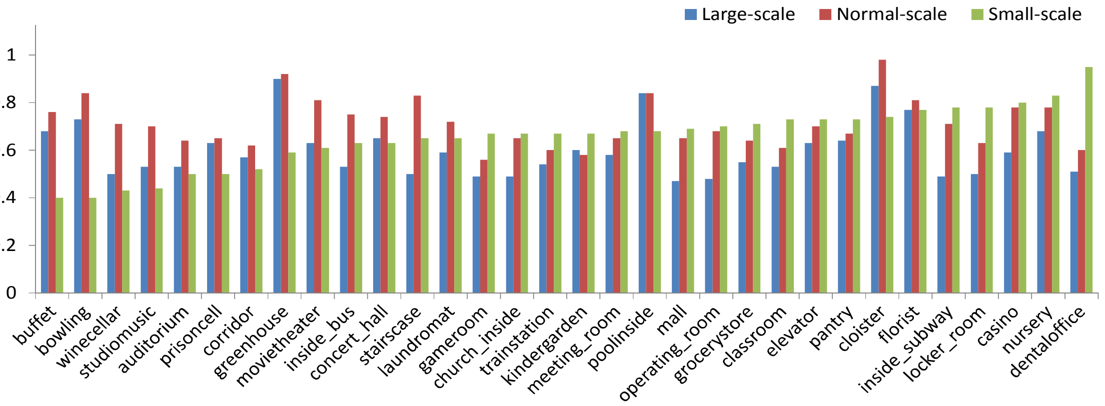

Fig. 7. Recognition effect of CNN with different scales on scene category

scene recognition, containing 397 categories. Places is the largest scene recognition dataset currently, we perform experiments on Places365, which contains 365 categories and more than 5000 images each category for training.

B. *Training*

Considering the training cost, we adopted two architectures in the experiment: the classical network AlexNet and the network Inception using the small convolution kernel. The Training set is resized from 172*172 to 428*428 as input and we adopted a few layers of model slightly to the classifier, for example, the size of pooling layer is 6*6 when the resolution is 172*172 in Inception network. We train the network with stochastic gradient utilizing the Pytorch[26] on two Nvidia Titan GPUs with batch size 32 for 100 epochs and we use a learning rate of 0.001, decayed every 20 epoch using an exponential rate of 0.5.

C. *Performance*

First, we randomly use 60% data in dataset, on which we train CNN with different scales, and we choose the other 20% data to be validation set. After the training stage, we choose the model that performed best in validation to test, the test set is the rest of unused data, the results are summarized in Table. II.

TABLE II. EVALUATION OF CNN WITH DIFFERENT SCALES

	Indoor	**SUN**	**Places365**
Places365-AlexNet [16]	**70.7%**	**56.1%**	**53.17%**
AlexNet-smallscale	71.5%	56.3%	53.3%
AlexNet- largescale	60.7%	48.6%	48.3%
Places365-Inception [16]	**73.3%**	**58.3%**	**53.6%**
Inception-smallscale	78.6%	71.5%	56.4%
Inception- largescale	68.5%	67.1%	53.0%

Then, we carried out experiments on combinations of different network characteristics including arithmetic mean method, and Table. III shows several results. As the result of multi-resolution CNN, this simple fusion scheme slightly improved the recognition performance of the three datasets, but our method is more accurate. These improvements show that the multi-level information captured by the two CNN images trained by different resolutions are strongly complementary, and the complementary correlation can be traced. Notably, our method is a modular learning framework that can be easily applied to any existing network architecture to enhance its capacity.

TABLE III. EVALUATION OF CNN WITH DIFFERENT FUSION

	Indoor	**SUN**	**Places365**
AlexNet-fusion1	72.3%	57.2%	53.5%
AlexNet-fusion2	**72.6%**	**57.5%**	**54.1%**
Inception-fusion1	78.2%	71.7%	56.7%
Inception-fusion2	**81.6%**	**73.5%**	**58.1%**

D. *Failure case*

Finally, we present a number of failure examples by our method, as shown in Fig. 8. Notice that part of failure cases are very relevant to the ground truth, they are still correct in top-5 prediction and the reason is mainly due to the label ambiguity, such as heliport and airfield. On the other hand, the less-typical activities caused by human action happen in a scene, leading to the confused semantic information CNN extracted. These results suggest the need to multi-ground truth labels and representation of correlation among objects for scene recognition.

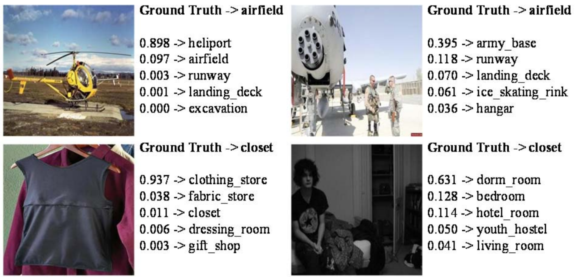

Fig. 8. Failure cases of CNN for scene recognition task.

VI. Conclusion

This paper proposes the main problems of scene recognition by comparing object classification and scene recognition task: the quantity of object and the variation of scale. We propose that the multi-scale CNN can capture the different object features in the scene, and learn the appropriate weights in combination with the scene categories, effectively combining the object features of different scales. Compared with the previous work, the multi-scale method proposed in this paper combines multi-scale features in a new way by adopting the appropriate hybrid mode for different scales to different scene category.

Actually, current researches tend to improve the structure of CNN, aiming to reduce gradient disappearance to achieve deeper network and extract more effective and comprehensive visual features. However, there are few types of research on the details of CNN in various classification tasks. The performance of Scene Recognition can't reach the optimal state by the CNN which is trained the same as the way used for object classification. This paper studies the characteristic of scene recognition and proposes the multi-scale feature fusion method, so as to reduce the interference from scene diversity. In the future, we will study confusable scene categories, explore the distribution impact of objects in these categories, and propose a deeper mode of fusing features.

References

[1] Wang. LiMin, Yirui Wu, Tong Lu, and Kang Chen, “Multiclass object detection by combining local appearances and context,” ACM international conference on Multimedia, 2011, pp. 1161-1164.

[2] Zhu, Anna, Guoyou Wang, Yangbo Dong, and Brian Kenji Iwana, “Detecting text in natural scene images with conditional clustering and convolution neural network,” Journal of Electronic Imaging, vol. 24, no. 5, 2015.

[3] Wang, Limin, Zhe Wang, Wenbin Du, and Yu Qiao, “Object-scene convolutional neural networks for event recognition in images,” IEEE conference on computer vision and pattern recognition workshops, pp. 30-35, 2015.

[4] Wang. L, Qiao.Y, and Tang.X, “MoFAP: A multi-level representation for action recognition,” International Journal of Computer Vision, vol. 119, no. 3, pp. 254-271, 2016.

[5] Sun Zehang, George Bebis, and Ronald Miller, “On-road vehicle detection: A review,” IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 5, pp. 694-711 , 2006.

[6] Krizhevsky Alex, Ilya Sutskever, and Geoffrey E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, pp. 1097-1105 , 2012.

[7] Simonyan Karen and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv, pp. 1409-1556 , 2014.

[8] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 , 2016.

[9] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, “Rethinking the inception architecture for computer vision,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826 , 2016.

[10] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, “Densely connected convolutional networks,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708 , 2017

[11] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252 , 2015.

[12] Everingham, Mark, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, “The pascal visual object classes challenge: A retrospective,” International journal of computer vision, vol.111, no.1, pp. 98-135 , 2015.

[13] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” Computer Vision and Pattern Recognition, pp. 248-255, 2009.

[14] Quattoni Ariadna and Antonio Torralba, “Recognizing indoor scenes,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 413-420 , 2009.

[15] Xiao, Jianxiong, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 3485-3492 , 2010.

[16] Zhou, Bolei, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, “Learning deep features for scene recognition using places database,” Advances in neural information processing systems, pp. 487-495, 2014.

[17] Zhou, Bolei, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, “Scene parsing through ade20k dataset,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 2, 2017.

[18] Yang, Jun, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo, “Evaluating bag-of-visual-words representations in scene classification,” ACM Proceedings of the international workshop on Workshop on multimedia information retrieval, pp. 197-206 , 2007.

[19] Parizi S N, John G. Oberlin, and Pedro F. Felzenszwalb, “Reconfigurable models for scene recognition,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 2775-2782 , 2012.

[20] Oliva Aude and Antonio Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope.” International journal of computer vision, vol. 42, no. 3, pp.145-175 , 2001.

[21] Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, “Speeded-up robust features (SURF),” Computer vision and image understanding, vol. 110, no.3, pp. 346-359 , 2008.

[22] Wu Jianxin and Jim M. Rehg, “CENTRIST: A visual descriptor for scene categorization,” IEEE transactions on pattern analysis and machine intelligence, vol. 33, no.8, pp. 1489-1501 , 2011.

[23] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Object detectors emerge in deep scene cnns,” arXiv, pp. 1412-6856 , 2014.

[24] Wang, Limin, Sheng Guo, Weilin Huang, Yuanjun Xiong, and Yu Qiao, “Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs,” IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 2055-2068 , 2017.

[25] Herranz Luis, Shuqiang Jiang and Xiangyang Li, “Scene recognition with CNNs: objects, scales and dataset bias,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 571-579 , 2016.

[26] Ketkar Nikhil, “Introduction to pytorch,” Deep Learning with Python, Apress, Berkeley, CA, pp. 195-208 , 2017.