**Learning Deep Features for Scene Recognition using Deep Learning**

Prof. Aakanksha Jain

Department of Computer Science & Engineering

Shri Ram Institute of Science and Technology

Jabalpur (MP), India

aakankshajain836@gmail.com

Prof. Pinkal Jain

Department of Computer Science & Engineering

Gyan Ganga College of Technology

Jabalpur (MP), India

pinkaljain@ggct.co.in

**Abstract**

Recognizing scenes is a fundamental yet complex task in computer vision, as it provides essential context for accurately identifying objects within images. While recent advances in object recognition—driven by large-scale datasets like ImageNet and the development of Convolutional Neural Networks (CNNs)—have shown impressive results, progress in scene recognition has lagged behind. One reason for this may be that features learned from object-focused datasets are not well-suited for capturing the broader contextual information required for scene understanding.

To bridge this gap, we introduce *Places*, a large-scale, scene-oriented dataset containing roughly 7 million labeled images spanning a diverse range of environments. This dataset offers a new framework for analyzing the diversity and density of visual data and matches or exceeds other scene datasets in richness and coverage. We leverage CNNs trained on Places to extract scene-specific deep features, leading to new state-of-the-art results on several scene recognition benchmarks. Additionally, by examining the internal activations of these networks, we demonstrate clear differences in feature representation between models trained on scenes versus those trained on objects.

1. **Introduction**

One of the most remarkable abilities of the human brain is its capacity to interpret and understand visual environments almost instantly. Within just a few milliseconds, we can categorize a scene or identify an object, underlining the significance of rapid, feedforward visual processing. This efficiency is largely driven by our brain’s ability to learn and remember a vast number of environments and visual examples throughout our lives. The human visual system constantly samples its surroundings, taking in numerous images every second—even if only for a moment—which results in exposure to millions of natural scenes each year. This raises an important question: how much experience and training would an artificial system need to achieve a level of scene recognition comparable to that of a human?

Inspired by the brain’s layered structure, which processes visual information through increasingly complex stages, Convolutional Neural Networks (CNNs) have been developed and have shown exceptional results in object classification tasks, especially when paired with large-scale datasets. However, CNNs trained on object-centric datasets often underperform on scene recognition tasks, producing results similar to those achieved using traditional handcrafted features and complex classifiers. We argue that this performance gap stems from the fact that object-focused images lack the visual complexity and contextual richness necessary for effective scene understanding. Consequently, CNNs trained on such images develop different high-level representations than those trained on diverse scene-centric data.

To address this, we introduce *Places*, a large-scale dataset of scene-oriented images that is significantly more extensive—about 60 times larger—than the SUN database. Using this dataset and a standard CNN framework, we establish new performance records on several prominent benchmarks, including MIT Indoor67, SUN, SUN Attribute, and Scene15. Interestingly, while the same network architecture trained on ImageNet achieved strong object classification results, training it with Places led to notably superior performance on scene classification tasks.

This paper is structured as follows: Section 2 explains the data collection process and introduces the Places dataset. Section 3 presents a comparative analysis of Places, ImageNet, and SUN, including evaluations of dataset density and diversity using Amazon Mechanical Turk (AMT). Section 4 demonstrates the improvements in scene recognition accuracy using deep features learned from the Places dataset. Finally, we explore the activations across different CNN layers to reveal the internal differences between object- and scene-centric models.

2. **Literature Review**

The Scene15 dataset served as one of the earliest benchmarks for scene classification, featuring only 15 categories with a few hundred images per class. Today, many classifiers have nearly saturated this dataset, achieving performance levels close to human accuracy at around 95%. To offer a broader and more complex challenge, the MIT Indoor67 dataset was introduced, covering 67 distinct indoor environments. Expanding the scope further, the SUN dataset was created to represent a comprehensive range of scene categories, encompassing 397 types, each with over 100 annotated images.

Although ImageNet includes a limited number of scene categories, their representation is minimal (as illustrated in Fig. 2), making existing scene-centric datasets significantly smaller in comparison to contemporary object-centric datasets. To address this imbalance, we introduce the *Places* dataset—a large-scale, scene-focused collection designed to complement the object-centric nature of ImageNet. Comprising over 7 million images across 476 scene categories, Places stands as the most extensive scene and environment image archive to date. It is also the first scene-centric dataset capable of supporting the training needs of CNNs and other deep learning models that require large volumes of data.

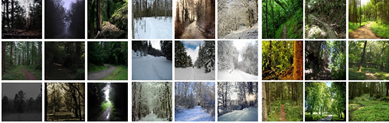
The hierarchical structure of the SUN taxonomy was retained and extended in the Places dataset. To enhance dataset diversity, we selected 666 commonly used English adjectives (e.g., "bright," "messy," "spare," "gloomy") and combined them with each scene category to formulate search queries. These queries were submitted to major image search engines—Google Images, Bing Images, and Flickr. This strategy enabled the retrieval of a broader variety of scene appearances, resulting in the download of over 40 million candidate images, which surpasses the volume of images found in ImageNet. Duplicate URLs were removed, and only distinct color images with a minimum resolution of 200x200 pixels were retained.

To prevent overlap between the Places and SUN datasets, we employed a Principal Component Analysis (PCA)-based duplicate filtering method within each category of the Places dataset and between corresponding categories in SUN. This ensured both datasets remained unique and suitable for integration if required.

After the initial filtering, we conducted two rounds of manual image annotation using Amazon Mechanical Turk (AMT). Each image was presented with a specific category prompt (e.g., "Is this a living room scene?") along with its textual definition. In the first round, the default selection was "No," prompting workers to actively select images that matched the category. In the second round, only the positively labeled images were re-evaluated, this time with the default set to "Yes," to better identify any false positives or remaining negatives.

Each Human Intelligence Task (HIT) presented 750 images for review, including 30 confirmed positive and 30 negative samples from the SUN database as quality control checks. Workers were required to achieve at least 90% accuracy on these control samples for their submissions to be accepted. After completing both annotation phases, the final Places dataset contained 7,076,580 validated images spanning 476 distinct scene categories.Fig. 1 displays some images created using a few of the adjectives found in the questions.







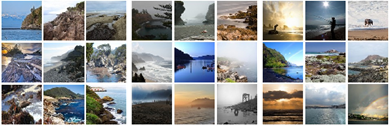


Figure 1: Image samples from the scene categories grouped by their queried adjectives.

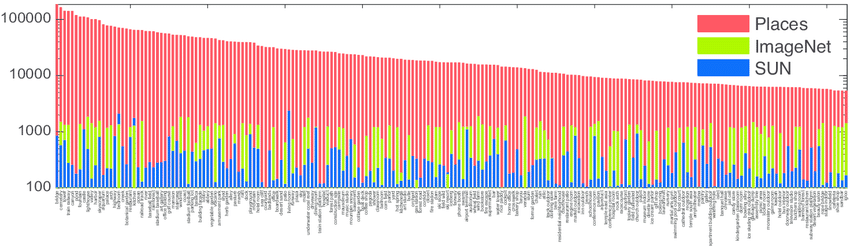


Figure 2: Comparison of the number of images per scene category in three databases.

To facilitate consistent evaluation, two specific subsets of the Places dataset were created and are used as benchmarks throughout this study. The first subset, known as *Places205*, comprises approximately 5,000 images across 205 scene categories. Figure 2 provides a comparison of the number of images in Places205 with those in ImageNet and SUN for the same categories. Notably, the SUN dataset includes all 205 categories (referred to here as *SUN205*), with a minimum of 50 images per category, whereas ImageNet contains only 128 of these categories.

The second subset used in this paper is *Places88*, which includes the 88 scene categories that are common across all three datasets and for which ImageNet has at least 1,000 images per category. For consistency, the corresponding groups in the ImageNet and SUN datasets are referred to as *ImageNet88* and *SUN88*, respectively.

**3.Comparing Scene-centric Databases**

Despite the critical role that training datasets and benchmarks play in computer vision, comparing them remains a complex task. Even datasets that cover the same visual categories can differ greatly in how well they support model generalization when used to train classifiers [7]. These differences often stem from subtle factors such as variations in camera angles, scene layouts, object arrangements, or interior styles—elements that are important but not easily quantified beyond just counting images or categories.

While the quality of a dataset often depends on its intended use, two key characteristics are generally desirable: diversity, meaning the dataset captures a wide range of appearances and viewpoints; and density, indicating that the data within each category is rich and representative. However, measuring diversity and density in image datasets is inherently difficult because they rely on notions of visual similarity that are subjective and loosely defined. For instance, two scene images may appear similar if they share objects, spatial configurations, and stylistic elements—but whether they are truly similar is often a matter of interpretation.

To address this challenge, we introduce relative metrics that focus on ranking similarities rather than assigning absolute values. These metrics enable a more practical comparison of datasets in terms of visual density and diversity. In this section, we apply these relative measures to analyze and compare the SUN, ImageNet, and Places datasets.

**Relative Density and Diversity**

Data concentration is measured by density. We make the assumption that a high density in a picture set corresponds to the fact that the images' neighbours are generally similar. The goal of relative density is to determine which of two databases, A and B, has the most similar nearest neighbours. We choose the closest neighbors of every set, a2 from A and b2 from B, and suppose that b1 from set B and a1 from set A are random images from respective sets. Should A be denser than B, a1 and a2 are more likely to be closer to one another than b1 and b2. Based on this idea, DenB(A) = p (d(a1, a2) < d(b1, b2)) where d(a1, a2) is a distance measure between two pictures (short distance denotes great similarity).

With this definition of relative density we have that A is

denser than B if, and only if, DenB(A) >DenA(B). This definition can be extended to an arbitrary number of datasets, A1, ..., AN:

DenA2,...,AN (A1) = p(d(a11, a12) <min d(ai1, ai2))

i=2:N

where ai are randomly selected and ai2 are near neighbours of their respective ai1

A dataset's quality cannot be determined only by its density. Consider, for example, a collection of 100,000 photos collected in the same bedroom. Since every image in this dataset would have a fairly identical appearance, it would have a very high density but a very little diversity. A dataset with significant diversity is also predicted to generalise well.

To describe the richness of an ecosystem, a variety of diversity metrics are commonly employed in biology (for a summary, see [9]). This section will employ a metric that draws inspiration from the Simpson Index of Diversity [10]. The likelihood that two randomly selected individuals from an ecosystem are members of the same species is measured by the Simpson index. It is correlated with the distribution's entropy and indicates how evenly dispersed individuals are among various species in an ecosystem. If subcategories are not annotated, it is difficult to extend this metric for assessing the diversity of photos inside a category. Therefore, we suggest using this to gauge the relative diversity of picture collections A and B.concept: two random samples from set A are less likely to be visually similar than two random photos from set B if set A is more diverse than set B. Then, DivB(A) = 1 − p(d(a1, a2) < d(b1, b2)), where a1, a2 ∈ A and b1, b2 ∈ B are randomly chosen, can be used to determine the diversity of A with regard to B. If and only if DivB(A) >DivA(B), then A is more diverse than B according to this definition of relative diversity. A1,..., AN: DivA2,...,AN (A1) = 1 − p(d(a11, a12) < min d(ai1, ai2)) (2) i=2:N, where ai1, ai2 ∈ Ai are chosen at random for an arbitrary number of datasets.

**4. Experimental Results**

Using Amazon Mechanical Turk (AMT), we assessed the relative *density* and *diversity* of the SUN, ImageNet, and Places datasets. The same experimental interface was employed for both evaluations: workers were shown sets of image pairs and asked to identify which pair appeared more visually similar. Our results indicated a high level of agreement among annotators, suggesting consistent judgments when comparing image similarity.

The primary difference between the two experiments—measuring density versus diversity—lay in how the image pairs were selected. For the *diversity* analysis, image pairs from each dataset were chosen randomly. Each trial presented participants with 12 image pairs—four from each dataset—to increase the chances of encountering a similar-looking pair and reduce the likelihood of skipped questions. In each trial, AMT workers were asked to select the most visually similar pair.

We focused on the 88 scene categories common to the ImageNet, SUN, and Places datasets. For each category, 40 trials were conducted, with two separate annotators reviewing each trial. Figure 3a shows examples of image pairs from one of the density tests, with the pair selected as most similar by AMT workers highlighted.

For the *density* analysis, image pairs more likely to resemble each other visually were selected. Rather than manually identifying the true nearest neighbors—which would be computationally expensive—we estimated visual similarity using the Euclidean distance between Gist descriptors of the images. For each pair, we randomly selected one image and paired it with its fifth closest match based on Gist similarity, intentionally skipping the top four closest matches to reduce potential bias.

****

Figure 3: a) Examples of pairs for the diversity experiment.

****

b) Examples of pairs for the density experiment.

close to duplicates, which would create an incorrect impression of high density). In this instance, we conduct 25 trials per category rather than 40 to prevent repetitive searches, but we still display 12 pairs of images at each trial. Examples of pairs from each density experiment are displayed in Fig. 3b, along with a highlight for the chosen pair. You'll notice that the pairs appear more alike overall in the density experiment (where we calculated neighbours) than in the diversity trial.

A scatter plot of relative diversity vs relative density for each of the three databases and all 88 categories is displayed in Fig. 3c. If the diversity and density of each dataset were the same, the spot where the two black lines intersect is where all of the findings should fall. The average of each dataset's density and diversity across all categories is also displayed in the figure.

The three datasets are generally rather similar in terms of density. On the other hand, diversity varies more, indicating that Places is the most diverse of the three datasets. For each dataset, the average relative diversity is 0.50 for SUN, 0.67 for ImageNet, and 0.83 for Places. Users rated pairs from the Places database to be the most comparable just 17% of the time in the experiment, but they chose pairs from the SUN database to be the closest 50% of the time. Playground, veranda, and waiting room are the categories with the biggest diversity variance among the three datasets.

**4.1Cross Dataset Generalization**

Due to the dataset bias issue, training and testing on various datasets typically leads in a performance decline. In this instance, the disparities in density and variety among the three datasets are one of the main causes of the bias between them. The classification results from training and testing on various permutations of the three datasets are displayed in Fig. 4. We use a linear SVM and features taken from a pre-trained ImageNet-CNN to achieve these results. For a specific number of training examples, the best results are obtained while training and testing on the same dataset in all three scenarios. Due to its size, the Places database performs best on two test sets when all of the training data is used. In contrast to a network trained using ImageNet, we will demonstrate in the following section that a CNN network trained with the Places database achieves a notable improvement over scene-centred benchmarks.

|  |
| --- |
| Table 1: Classification accuracy on the test set of Places 205 and the test set of SUN 205. |
| Places 205 SUN 205 |
| Places-CNN 50.0% 66.2% |
| ImageNet CNN feature+SVM 40.8% 49.6% |

**4.2 Training Neural Network for Scene Recognition and Deep Features**

On the ImageNet benchmark, deep convolutional neural networks have demonstrated remarkable classification performance [12]. We choose 2,448,873 photos at random from 205 Places categories (called Places 205) as the train set for Places-CNN training, with a minimum of 5,000 and a maximum of 15,000 images per category. There are 200 photographs each category in the test set and 100 images per category in the validation set, for a total of 41,000 images. Places-CNN is trained on an NVIDIA Tesla K40 GPU using the Caffe package. About six days were needed to complete 300,000 training iterations.Places-CNN shares the same network architecture as the Caffe reference network [10]. The architecture of the network suggested by [12] is similar to that of the Caffe reference network, which is trained on 1.2 million images of ImageNet (ILSVRC 2012). Throughout the subsequent comparative tests, we refer to the Caffe reference network as ImageNet-CNN.

**4.3 Visualization of the Deep Features**

Given that ImageNet-CNN and Places-CNN have the same architecture, we may better comprehend their differences by visualising the responses of the units for different network layer levels. The learnt representation of the units at the two networks' Conv 1, Pool 2, Pool 5, and FC 7 levels is shown in Fig. 5. We use the mean image method to visualise the units of the higher layers, while Conv 1 units can be visualised directly (they capture the orientated edges and opponent colours from both networks). To do this, we first combine the test set of ImageNet LSVRC2012 (100,000 images) and SUN397 (108,754 images) as input for both networks, and then we sort all of these images according to the activation response.Considering each unit at each layer; as a sort of receptive field (RF) visualisation of each unit, we then average the top 100 images with the biggest reactions for each unit. Fig. 5 shows mean pictures sorted by their first principal component so that the units from the two networks may be compared. Even though the technique is straightforward, there are significant changes between the units in the two networks, beginning with Pool 2. The units in ImageNet-CNN gradually have RFs that resemble object-blobs from Pool 2 to Pool 5 and FC 7, but the units in Places-CNN have more RFs that resemble landscapes with richer spatial organisation. The variations in the training data have a direct bearing on these learnt unit architectures.

The similarities and contrasts between the RF at various object-centric network and scene-centric network layers and the known object-centred and scene-centred neural cortical pathways found in the human brain would be intriguing to relate in future work (for a review). We will demonstrate in the following part that these two networks produce radically different results on a range of recognition benchmarks, with the sole difference being in the training sets.

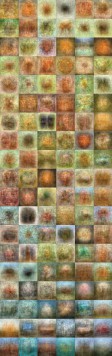
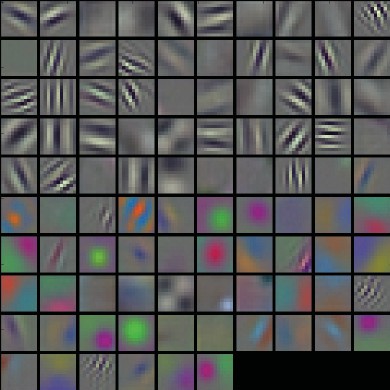
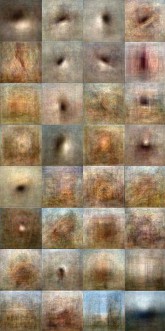
**4.4 Results on Places 205 and SUN 205**

Following training of the Places-CNN, we classify photos in the Places 205 and SUN 205 test set using the network's final layer output (Soft-max). Table 1 lists the categorisation result. The outcomes of a linear SVM trained on ImageNet-CNN features of 5000 photos per category in Places 205 and 50 images per category in SUN 205, respectively, are displayed as a baseline comparison. Places-CNN is far more effective. We also calculate the Places-CNN's performance in terms of the top-5 error rate (if the ground-truth label is not one of the model's top 5 predicted labels, one test sample is considered misclassified). For the Places 205 test set, the top-5 error rate is 18.9%.whereas 8.1% is the top-5 error rate for the SUN 205 test set.

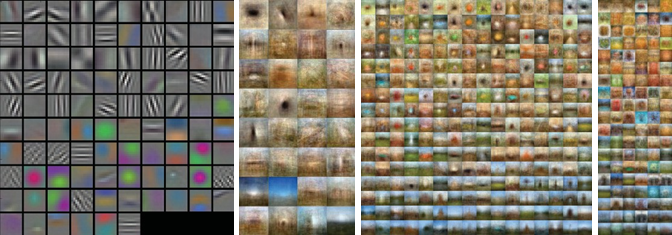
**4.5 Generic Deep Features for Visual Recognition**

For visual recognition challenges, we utilise the trained CNN's replies as generic features. On a variety of picture datasets, responses from CNN's higher-level layers havedemonstrated state-of-the-art performance as efficient generic features.

**Conv 1 Pool 2 Pool 5 FC 7**



ImageNet-CNN

****

Places-CNN

Figure 5: Visualization of the units’ receptive fields at different layers for the ImageNet-CNN and Places-CNN. Conv 1 units contains 96 filters. The Pool 2 feature map is 13×13×256; The Pool 5 feature map is 6×6×256; The FC 7 feature map is 4096×1. Subset of units at each layer are shown.

Places-CNN

Table 2:Classification accuracy/precision on scene-centric databases and object-centric databases for the Places-CNN feature and ImageNet-CNN feature.The classifier in all the experiments is a linear SVM with the same parameters for the two features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SUN397 | MITIndoor67 | Scene15 | SUNAttribute |
| Places-CNNfeature | 54.32±0.14 | 68.24 | 90.19±0.34 | 91.29 |
| ImageNet-CNNfeature | 42.61±0.16 | 56.79 | 84.23±0.37 | 89.85 |
|  | Caltech101 | Caltech256 | Action40 | Event8 |
| Places-CNNfeature | 65.18±0.88 | 45.59±0.31 | 42.86±0.25 | 94.12±0.99 |
| ImageNet-CNNfeature | 87.22±0.92 | 67.23±0.27 | 54.92±0.33 | 94.42±0.76 |

SUN397, MIT Indoor67, Scene15, SUN Attribute, Caltech101, Caltech256, Stanford Action40, and UIUC Event8 are the scene and object benchmarks on which the Places-CNN provided deep features. Every experiment complies with the guidelines in those publications. 2. On the same benchmarks, we compare the performance of the deep feature from the ImageNet-CNN. Though they are trained on scene-centric and object-centric data, respectively, Places-CNN and ImageNet-CNN share the same network architecture. The Fully Connected Layer (FC) 7 of the CNNs is the last fully connected layer before generating the class predictions, and we use the deep features from its response. Only a slight distinction exists between FC 7's feature and that ofFC 6 layer. Each image's deep feature is a 4096-dimensional vector.

The classification accuracy for the ImageNet-CNN and Places-CNN features across multiple datasets is compiled in Table 2. The classification accuracy for several visual aspects on the SUN397 database and SUN Attribute dataset is presented in Fig. 6. For the two deep features (C=1), the classifier is a linear SVM with the same default parameters. On scene classification benchmarks, the Places-CNN feature performs admirably, surpassing the state-of-the-art techniques for MIT Indoor67 (66.87% ) and SUN397 (47.20%). However, when it comes to object-related databases, the ImageNet-CNN function performs better.

Table 3: Classification accuracy/precision on various databases for Hybrid-CNN feature. The numbers in bold indicate the results outperform the ImageNet-CNN feature or Places-CNN feature.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUN397 | MIT Indoor67 | Scene15 | SUN Attribute | Caltech101 | Caltech256 | Action40 | Event8 |
| 53.86±0.21 | 70.80 | 91.59±0.48 | 91.56 | 84.79±0.66 | 65.06±0.25 | 55.28±0.64 | 94.22±0.78 |

This demonstrates that, as anticipated from the benchmark datasets used to train both networks, Places-CNN and ImageNet-CNN have complimentary strengths on scene-centric and object-centric tasks. Additionally, we fine-tune Places-CNN on SUN397 using the same experimental setup of train and test split in [1]; the accuracy of the fine-tuned Places-CNN is 56.2%, while the accuracy of the fine-tuned ImageNet-CNN in [1] is 52.2%. Take note that scene category is predicted straight from the fine-tuned CNN's final output.

Furthermore, we train a Hybrid-CNN by merging the Places-CNN and ImageNet-CNN training sets. The training set of Hybrid-CNN contains 3.5 million images from 1183 categories after we eliminate the overlapping scene categories from the ImageNet training set. Hybrid CNN uses the same network architecture as Places-CNN and ImageNet-CNN and is trained over 700,000 iterations. 52.3% is the accuracy on the validation set. We use the standards listed in Table 3 to assess the deep feature (FC 7) from Hybrid-CNN. For a few benchmarks, combining the two datasets results in an extra speed boost.

**5.Conclusion**

Deep convolutional neural networks are designed to benefit and learn from massive amounts of data. We introduce a new benchmark with millions of labeled images, the Places database, designed to represent places and scenes found in the real world. We introduce a novel measure of density and diversity, and show the usefulness of these quantitative measures for estimating dataset biases and comparing different datasets. We demonstrate that object-centric and scene-centric neural networks differ in their internal representations, by introducing a simple visualization of the receptive fields of CNN units. Finally, we provide the state-of-the-art performance using our deep features on all the current scene benchmarks.

**References**

[1] J. Malik, R. Girshick, and P. Agrawal. evaluating multilayer neural networks' object identification ability. 2014. In Proc. ECCV.

[2] Bengio, Y. AI deep architecture learning. R in Machine Learning: Foundations and Trends, 2009.

[3] L.-J. Li, K. Li, L. Fei-Fei, R. Socher, W. Dong, and J. Deng. An extensive hierarchical image database is called Imagenet. 2009, in Proc. CVPR.   
[4] A. Gupta, A. A. Efros, and C. Doersch. Discovery of mid-level visual elements as a discriminative mode searching technique. Neural Information Processing Systems Advances, 2013.   
[5]J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Y. Jia, O. Vinyals, and J. Donahue. For general visual recognition, Decaf is a deep convolutional activation feature. 2014.

X.-R. Wang, C.-J. Lin, C.-J. Hsieh, K.-W. Chang, and R.-E. Fan [6]. A library for massive linear classification is called LIBLINEAR. 2008.   
[7] P. Perona, R. Fergus, and L. Fei-Fei. An incremental bayesian method was tried on 101 object categories to learn generative visual models from a small number of training instances. Image Understanding and Computer Vision, 2007.   
[8]A. Holub, P. Perona, and G. Griffin. Caltech-256 dataset for object categories. 2007.   
[9] K. Soetaert, P. Herman, and C. Heip. signs of equality and diversity. Oceanis (1998).   
[10]Caffe, an open-source convolutional architecture for quick feature embedding, was developed by Y. Jia. Berkeley Vision, 2013. http://caffe.berkeleyvision.org/.   
[11] A. Oliva, G. A. Alvarez, T. F. Brady, and T. Konkle. The function of categories in visual long-term memory: scene memory is more intricate than you may imagine. PsychologicalPsych Science, 2010.   
[12] G. E. Hinton, I. Sutskever, and A. Krizhevsky. Using deep convolutional neural networks for imagenet categorisation. Neural Information Processing Systems Advances, 2012.   
[13] J. Ponce, C. Schmid, and S. Lazebnik. Beyond feature-rich bags: Recognising categories in natural scenes via spatial pyramid matching. 2006, in Proc. CVPR.   
[14] LeCun, Y., Hubbard, B., Denker, J. S., Henderson, D., Howard, R. E., and Jackel, L. D. Handwritten zip code recognition using backpropagation. 1989; Neural Computation.   
[15]L. Fei-Fei and L.-J. Li. Who, what, and where? utilising object and scene recognition to categorise events. Proceedings of the ICCV, 2007.   
[16]Oliva, A. Perception of the scene (chapter 51). 2013's The New Visual Neurosciences

[17] A. Torralba and A. Oliva. A comprehensive depiction of the spatial envelope is used to model the scene's shape. International Computer Vision Journal, 2001.   
[18] J. Hays and G. Patterson. Finding, labelling, and identifying scene tributes in the Sun attribute database. 2012, in Proc. CVPR.   
[19] A. Torralba and A. Quattoni. identifying scenes that are indoors. 2009, in Proc. CVPR.   
[20] S. Carlsson, J. Sullivan, H. Azizpour, and A. S. Razavian. Off-the-shelf CNN features provide an amazing starting point for recognition. The 2014 preprint arXiv is arXiv:1403.6382.