**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**

One of the main challenges in computer vision is scene recognition, which enables the definition of a context for object recognition. The availability of massive datasets like ImageNet and the development of Convolutional Neural Networks (CNNs) for learning high-level features are responsible for the remarkable recent advancements in object recognition tasks; nevertheless, scene recognition performance has not reached the same level of success. This might be as a result of the inadequacy of the deep features currently trained from ImageNet for these kinds of tasks. We present Places, a brand-new scene-centric database that contains more than 7 million labelled images of scenes. We demonstrate that Places is just as dense as other scene datasets and have new techniques for comparing the diversity and density of image datasets greater variety. We achieve new state-of-the-art performances on multiple scene-centric datasets by using CNN to learn deep features for scene recognition tasks. We may demonstrate the distinctions between object-centric and scene-centric networks' internal representations by visualising the CNN layers' responses.

1. **Introduction**

One of the greatest achievements of the human brain is the ability to comprehend the world in a single glance; it only takes a few tens of milliseconds to identify an object or environment's category, highlighting the crucial role feedforward processing plays in visual identification. Our ability to learn and retain a wide variety of locations and exemplars is one of the mechanisms underlying effective human visual recognition [1]. By sampling the world multiple times per second, our neural architecture continuously registers new inputs, even for brief periods of time, reaching exposure to millions of natural images in a single year. How much learning would it take for an artificial system to match a human's scene recognition skills?

A key characteristic of the primate brain is its hierarchical organisation in layers of increasing processing complexity, which has inspired Convolutional Neural Networks, or CNNs[2], in addition to being exposed to a wide range of natural images. These designs have achieved remarkable results on object classification tasks especially when combined with recent huge databases[3]. Nonetheless, these networks' baseline performance on scene categorisation tasks falls within the range of performance based on complex classifiers and hand-designed features. Here, we demonstrate that one of the causes of this disparity is that the richness and diversity of visual information that photos of scenes and surroundings offer for learning to recognise them is not present in iconic images of items, which is why object-centric CNNs learn different higher-level features than scene-centric CNNs.

We provide Places, a scene-focused image collection that is 60 times bigger than SUN. Using this database and a typical CNN architecture, we set new accuracy benchmarks on MIT Indoor67, SUN database, SUN Attribute Database, and Scene15 were among the scene datasets that the deep features from thesame network architecture trained with ImageNet1 produced results that were noticeably better.

The structure of the paper is as follows: We present the Places database and outline the gathering process in Section 2. We contrast Places with the other two sizable image datasets, ImageNet and SUN, in Section 3. To compare the density and variety of these three datasets, we conduct experiments on Amazon Mechanical Turk (AMT)[4]. When training deep features from millions of labelled scene photos, we demonstrate improved scene classification performance in Section 4. Lastly, we illustrate the responses of the units at various CNN layers.

2. **Literature Review**

The Scene15 database, which was based on, served as the initial standard for scene classification. Present classifiers are saturating this dataset, approaching human performance at 95%, with only 15 scene categories and a few hundred photos per class. There are 67 categories for indoor locations in the MIT Indoor67 database. To give a broad coverage of scene categories, the SUN database was created. It is divided into 397 categories, each of which has over 100 photographs.

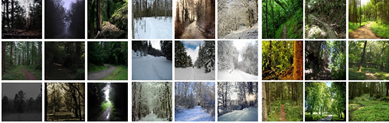
All of these scene-centric datasets, however, pale in comparison to existing object datasets like ImageNet (keep in mind that ImageNet also includes scene categories, but in a very minor percentage, as illustrated in Fig. 2). We offer here a scene-centric database, which we call the Places database, as a complement to ImageNet, which is mostly object-centric. locations is currently the largest image library of scenes and locations to date, with over 7 million images from 476 place categories[5]. It was also the first scene-centric database that was competitive enough to train algorithms like CNNs that require massive amounts of data.

The Places database inherited the same list of scene categories from the rich scene taxonomy of the SUN database. To generate the image URL query, 696 common adjectives (messy, spare, sunny, desolate, etc.) are manually selected from a list of popular English adjectives and combined with each scene category name. The queries are then sent to three image search engines (Google Images, Bing Images, and Flickr). By adding adjectives to the queries, we are able to download more images than ImageNet can offer and increase the variety of visual appearances. After that, we eliminate duplicate URLs and download the raw images with unique URLs. Over 40 million photos have been downloaded thus far. Only 200x200 pixel colour photos are retained. In order to guarantee that the Places and SUN databases do not contain identical photos, PCA-based duplicate removal is carried out inside each scene category in the Places database and across the same scene category in the SUN database. This enables us to merge the two datasets.

Following this preliminary screening, the images are submitted to Amazon Mechanical Turk for two iterations of individual image annotation. The query, "Is this a living room scene?" appears at the top of a screen for a particular category name, along with its definition as in [6]. Employees are requested to touch a Yes or No key as one image at a time is displayed in the centre of a sizable window. The worker must actively pick up the positive photos in the first labelling round because the default response is set to No.

After the first round of annotation, the positive images are forwarded for a second round of annotation, where the default response is Yes (to pick up the remaining negative images). 750 downloaded photos are included for annotation in each HIT (one assignment per worker), and as a control, 30 positive and 30 negative samples with ground truth from the SUN database are also randomly injected. An accuracy of 90% or more on these control images is necessary for valid HITs retained for additional analysis. The Places database now contains 7,076,580 photos from 476 scene categories following the two annotation rounds and the publication of this paper. Fig. 1 displays some images created using a few of the adjectives found in the questions.







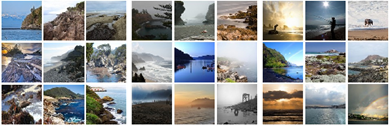


Figure 1: Image samples from the scene categories grouped by their queried adjectives.

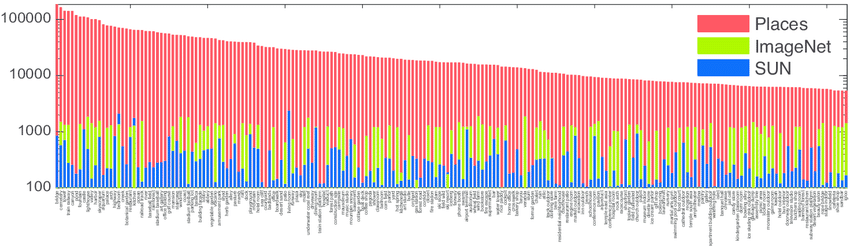


Figure 2: Comparison of the number of images per scene category in three databases.

Two subsets of Places were created and will serve as benchmarks throughout the article. The first is Places 205, which has almost 5000 photos in 205 categories. The quantity of photos in Places 205 using ImageNet and SUN is contrasted in Fig. 2. Keep in mind that SUN includes all 205 categories (we'll refer to this collection as SUN 205, and it has at least 50 images per category), but ImageNet only has 128 of them. Places 88 is the second subset of Places that is used in this paper. It includes all 88 of ImageNet's common categories, which means that ImageNet has at least 1000 photos. The relevant subsets are referred to as ImageNet 88 and SUN 88.

**3.Comparing Scene-centric Databases**

Comparing datasets remains a challenge in computer vision, despite the significance of training datasets and benchmarks. When used to train a classifier, even datasets spanning the same visual classes exhibit significant variances, resulting in varying generalisation performance [7]. Some factors, such as the variation in camera poses, décor styles, or objects in the scene, are significant but challenging to measure beyond the quantity of photos and classifications.

Although a database's quality will vary depending on the purpose, it is safe to say that a good database should be diverse (including a high degree of heterogeneity in appearances and viewpoints) and dense (including a high degree of data concentration). Because they rely on an idea of picture similarity that is generally ill-defined, density and diversity are both difficult to measure in image sets. If two pictures of scenes feature comparable objects, are in similar poses and spatial arrangements, and have similar decorative techniques, they might be regarded as similar. It is challenging to determine whether these two photos are similar, though, because this idea is nebulous and subjective. Therefore, we define relative measures that simply need ranking similarities in order to compare datasets in terms of density and diversity. We will use these relative measures to compare the densities and diversities of SUN, ImageNet, and Places in this section.

**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 nearest neighbours of each set, a2 from A and b2 from B, and assume that a1 is a random image from set A and b1 from set B. It is more likely that a1 and a2 are closer to one another than b1 and b2 if A is denser than B. The relative density is defined as DenB(A) = p (d(a1, a2) < d(b1, b2)) based on this concept, where d(a1, a2) is a distance measure between two images (small distance implies high 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**

With AMT, we calculated the relative densities and diversities of SUN, ImageNet, and Places. The same experimental interface was used for both measures: employees were shown various image pairs and asked to choose the one with the most similar photos. We found that while determining whether two photos are more similar than one another, different annotators consistently make the same decision.

The generation of the pairings is the only distinction between density and diversity estimation in these trials. Each database's couples are chosen at random for the diversity experiment. There are 12 pairings available for selection, with each trial consisting of 4 pairs from each database. To improve the likelihood of discovering a similar pair and prevent users from having to skip trials, we employed four pairs per database. On each trial, AMT employees had to choose the pair that was the most comparable. For the 88 categories shared by the ImageNet, SUN, and Places databases, we conducted 40 trials per category with two observers each trial. Examples of pairs from one of the density tests are displayed in Fig. 3a.Highlighted is the pair that AMT employees determined to be more similar.

Pairs that were more likely to be visually similar were chosen for the density experiments. This would necessitate first determining each image's genuine nearest neighbour, which would be expensive to do experimentally. Rather, we employed visual similarity as determined by the Euclidean distance between two images' Gist descriptor. Each image pair was made up of one randomly chosen image and its fifth closest neighbour using Gist; the first four neighbours were disregarded to prevent.

****

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.

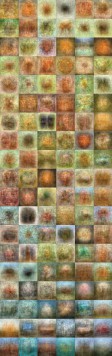
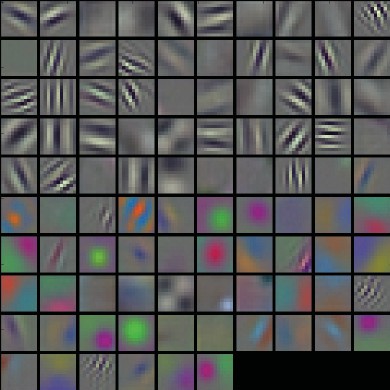
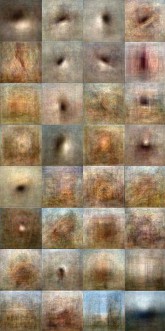
**4.4 Results on Places 205 and SUN 205**

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.

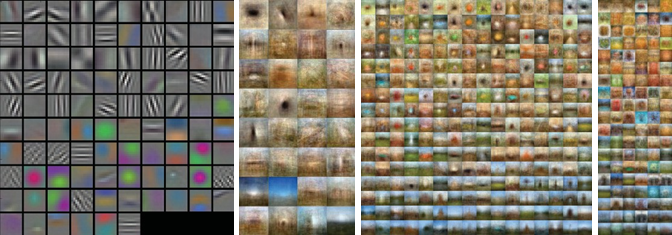
**4.5 Generic Deep Features for Visual Recognition**

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 |

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.