This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

Places: A 10 million Image Database for Scene Recognition

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba

Abstract—The rise of multi-million-item dataset initiatives has enabled data-hungry machine learning algorithms to reach near-human semantic classification performance at tasks such as visual object and scene recognition. Here we describe the Places Database, a repository of 10 million scene photographs, labeled with scene semantic categories, comprising a large and diverse list of the types of environments encountered in the world. Using the state-of-the-art Convolutional Neural Networks (CNNs), we provide scene classification CNNs (Places-CNNs) as baselines, that significantly outperform the previous approaches. Visualization of the CNNs trained on Places shows that object detectors emerge as an intermediate representation of scene classification. With its high-coverage and high-diversity of exemplars, the Places Database along with the Places-CNNs offer a novel resource to guide future progress on scene recognition problems.

Index Terms—Scene classification, visual recognition, deep learning, deep feature, image dataset.

1 Introduction

If a current state-of-the-art visual recognition system would send you a text to describe what it sees, the text might read something like: "There is a sofa facing a TV set. A person is sitting on the sofa holding a remote control. The TV is on and a talk show is playing". Reading this, you would likely imagine a living-room. However, that scenery can very well happen in a resort by the beach.

For an agent acting into the world, there is no doubt that object and event recognition should be a primary goal of its visual system. But knowing the place or context in which the objects appear is as equally important for an intelligent system to understand what might have happened in the past and what may happen in the future. For instance, a table inside a kitchen can be used to eat or prepare a meal, while a table inside a classroom is intended to support a notebook or a laptop to take notes.

A key aspect of scene recognition is to identify the place in which the objects seat (*e.g.*, beach, forest, corridor, office, street, ...). Although one can avoid using the place category by providing a more exhaustive list of the objects in the picture and a description of their spatial relationships, a place category provides the appropriate level of abstraction to avoid such a long and complex description. Note that one could avoid using object categories in a description by only listing parts (i.e. two eyes on top of a mouth for a face). Like objects, places have functions and attributes. They are composed of parts and some of those parts can be named and correspond to objects, just like objects are composed of parts, some of which are nameable as well (*e.g.*, legs, eyes).

- *B. Zhou, A. Khosla, A. Oliva, A. Torralba are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA.*
- *A. Lapedriza is with Universitat Oberta de Catalunya, Spain.*

Whereas most datasets have focused on object categories (providing labels, bounding boxes or segmentations), here we describe the Places database, a quasi-exhaustive repository of 10 million scene photographs, labeled with 434 scene semantic categories, comprising about 98 percent of the type of places a human can encounter in the world. Image samples are shown in Fig. 1 while Fig. 2 shows the number of images per category, sorted in decreasing order.

Departing from Zhou *et al.* [1], we describe in depth the construction of the Places Database, and evaluate the performance of several state-of-the-art Convolutional Neural Networks (CNNs) for place recognition. We compare how the features learned in a CNN for scene classification behave when used as generic features in other visual recognition tasks. Finally, we visualize the internal representations of the CNNs and discuss one major consequence of training a deep learning model to perform scene recognition: object detectors emerge as an intermediate representation of the network [2]. Therefore, while the Places database does not contain any object labels or segmentations, it can be used to train new object classifiers.

1.1 The Rise of Multi-million Datasets

What does it take to reach human-level performance with a machine-learning algorithm? In the case of supervised learning, the problem is two-fold. First, the algorithm must be suitable for the task, such as Convolutional Neural Networks in the large scale visual recognition [1], [3] and Recursive Neural Networks for natural language processing [4], [5]. Second, it must have access to a training dataset of appropriate coverage (quasi-exhaustive representation of classes and variety of exemplars) and density (enough samples to cover the diversity of each class). The optimal space for these datasets is often task-dependent, but the rise of multi-million-item sets has enabled unprecedented performance in many domains of artificial intelligence.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 1. Image samples from various categories of the Places Database (two samples per category). The dataset contains three macro-classes: Indoor, Nature, and Urban.

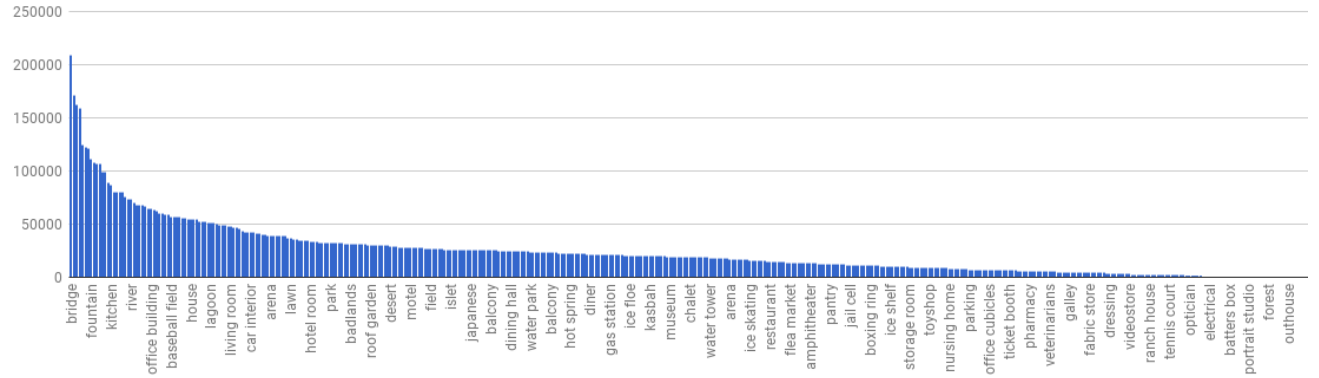


Fig. 2. Sorted distribution of image number per category in the Places Database. Places contains 10,624,928 images from 434 categories. Category names are shown for every 6 intervals.

The successes of Deep Blue in chess, Watson in "Jeopardy!", and AlphaGo in Go against their expert human opponents may thus be seen as not just advances in algorithms, but the increasing availability of very large datasets: 700,000, 8.6 million, and 30 million items, respectively [6]–[8]. Convolutional Neural Networks [3], [9] have likewise achieved near human-level visual recognition, trained on 1.2 million object [10]–[12] and 2.5 million scene images [1]. Expansive coverage of the space of classes and samples allows getting closer to the right ecosystem of data that a natural system, like a human, would experience. The history of image datasets for scene recognition also sees the rapid growing in the image samples as follows.

1.2 Scene-centric Datasets

The first benchmark for scene recognition was the Scene15 database [13], extended from the initial 8 scene dataset in [14]. This dataset contains only 15 scene categories with a few hundred images per class, and current classifiers are saturated, reaching near human performance with 95%. The MIT Indoor67 database [15] with 67 indoor categories and the SUN (Scene Understanding, with 397 categories and 130,519 images) database [16] provided a larger coverage of place categories, but failed short in term of quantity of data needed to feed deep learning algorithms. To complement large object-centric datasets such as ImageNet [11], we build the Places dataset described here.

Meanwhile, the Pascal VOC dataset [17] is one of the earliest image dataset with diverse object annotations in scene context. The Pascal VOC challenge has greatly advanced the development of models for object detection and segmentation tasks. Nowadays, COCO dataset [18] focuses on collecting object instances both in polygon and bounding box annotations for images depicting everyday scenes of common objects. The recent Visual Genome dataset [19] aims at collecting dense annotations of objects, attributes, and their relationships. ADE20K [20] collects precise dense annotation of scenes, objects, parts of objects with a large and open vocabulary. Altogether, annotated datasets further enable artificial systems to learn visual knowledge linking

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

parts, objects and scene context.

2 Places Database

2.1 Coverage of the categorical space

The first asset of a high-quality dataset is an expansive coverage of the categorical space to be learned. The strategy of Places is to provide an exhaustive list of the categories of environments encountered in the world, bounded by spaces where a human body would fit (e.g. closet, shower). The SUN (Scene UNderstanding) dataset [16] provided that initial list of semantic categories. The SUN dataset was built around a quasi-exhaustive list of scene categories with different functionalities, namely categories with unique identities in discourse. Through the use of WordNet [21], the SUN database team selected 70,000 words and concrete terms that described scenes, places and environments that can be used to complete the phrase "I am in a *place*", or "let's go to the/a *place*". Most of the words referred to basic and entry-level names ([22]), resulting in a corpus of 900 different scene categories after bundling together synonyms, and separating classes described by the same word but referring to different environments (e.g. inside and outside views of churches). Details about the building of that initial corpus can be found in [16]. Places Database has inherited the same list of scene categories from the SUN dataset, with a few changes that are described in section 2.2.4.

2.2 Construction of the database

The construction of the Places Database is composed of four steps, from querying and downloading images, labeling images with ground truth category, to scaling up the dataset using a classifier, and further improving the separation of similar classes. The detail of each step is introduced in the following sections.

The data collection process of the Place Database is similar to the image collection in other common datasets, like ImageNet and COCO. The definition of categories for the ImageNet dataset [11] is based on the synset of WordNet [21]. Candidate images are queried from several Image search engines using the set of WordNet synonyms. Images are cleaned up through AMT in the format of the binary task similar to the ours. Quality control is done by multiple users annotating the same image. There are about 500-1200 ground-truth images per synset. On the other hand, COCO dataset [18] focuses on annotating the object instances inside the images with more scene context. The candidate images are mainly collected from Flickr, in order to include less iconic images commonly returned by image search engines. The image annotation process of COCO is split into category labeling, instance spotting, and instance segmentation, with all the tasks done by AMT workers. COCO has 80 object categories with more than 2 million object instances.

2.2.1 Step 1: Downloading images using scene category and attributes

From online image search engines (Google Images, Bing Images, and Flickr), candidate images were downloaded using a query word from the list of scene classes provided by the SUN database [16]. In order to increase the diversity of visual appearances in the Places dataset, each scene class query was combined with 696 common English adjectives [1] (e.g., messy, spare, sunny, desolate, etc.). In Fig. 3) we show some examples of images in Places grouped by queries. About 60 million images (color images of at least 200×200 pixels size) with unique URLs were identified. Importantly, the Places and SUN datasets are complementary: PCA-based duplicate removal was conducted within each scene category in both databases, so that they do not contain the same images.

2.2.2 Step 2: Labeling images with ground truth category

Image ground truth label verification was done by crowdsourcing the task to Amazon Mechanical Turk (AMT). Fig.4 illustrates the experimental paradigm used. First, AMT workers were given instructions relating to a particular category at a time (e.g. cliff), with a definition, sample images belonging to the category (true images), and sample images not belonging to the category (false images). As an example, Fig.4.a shows the instructions for the category *cliff*. Workers then performed a verification task for the corresponding category. Fig.4.b shows the AMT interface for the verification task. The experimental interface displayed a central image, flanked by smaller version of images the worker had just responded (on the left), and the images the worker will respond to next (on the right). Information gleaned from the construction of the SUN dataset suggests that in the first iteration of labeling more than 50% of the the downloaded images are not true exemplars of the category. For this reason the default answer in the interface the default answer was set to *NO* (notice that all the smaller versions of the images in the left are marked with a bold red contour, which denotes that the image do not belong to the category). Thus, if the worker just presses the space bar to move, images will keep the default *NO* label. Whenever a true exemplar appears in the center, the worker can press a specific key to mark it as a positive exemplar (responding *YES*). As the response is set to *YES* the bold contour of the image turns to green. The interface also allows moving backwards to revise previous annotations. Each AMT HIT (Human Intelligence Task, one assignment for one worker), consisted of 750 images for manual annotation. A control set of 30 positive samples and 30 negative samples with ground-truth category labels from the SUN database were intermixed in the HIT as well. As a quality control measure, only worker HITs with an accuracy of 90% or higher on these control images were kept.

1. The list of adjectives used in querying can be found in https://github.com/CSAILVision/places365/blob/master/adjectives_download.csv

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 3. Image samples from four scene categories grouped by queries to illustrate the diversity of the dataset. For each query we show 9 annotated images.

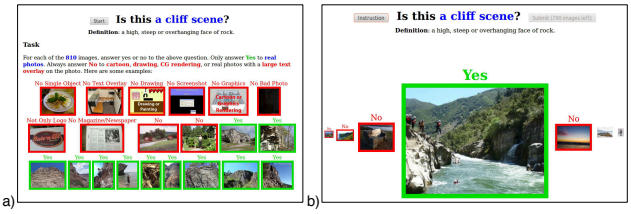


Fig. 4. Annotation interface in the Amazon Mechanical Turk for selecting the correct exemplars of the scene from the downloaded images. a) instruction given to the workers in which we define positive and negative examples. b) binary selection interface.

The positive images resulting from the first cleaning iteration were sent for a second iteration of cleaning. We used the same task interface but with the default answer was set to *YES*. In this second iteration, 25.4% of the images were relabeled as *NO*. We tested a third cleaning iteration on a few exemplars but did not pursue it further as the percentage of images relabeled as *NO* was not significant.

After the two iterations of annotation, we collected one scene label for 7,076,580 images pertaining to 476 scene categories. As expected, the number of images per scene category vary greatly (i.e. there are many more images of bedroom than cave on the web). There were 413 scene categories that ended up with at least 1000 exemplars, and 98 scene categories with more than 20,000 exemplars.

2.2.3 Step 3: Scaling up the dataset using a classifier

As a result of the previous round of image annotation, there were 53 million remaining downloaded images not assigned to any of the 476 scene categories (e.g. a *bedroom* picture could have been downloaded when querying images for *living-room* category, but marked as negative by the AMT worker). Therefore, a third annotation task was designed to re-classify then re-annotate those images, using a semi-automatic bootstrapping approach.

A deep learning-based scene classifier, AlexNet [3], was trained to classify the remaining 53 million images: We first randomly selected 1,000 images per scene category as training set and 50 images as validation set (for the 413 categories which had more than 1000 samples). AlexNet achieved 32% scene classification accuracy on the validation set after training. The trained AlexNet was then used to classify the 53 million images. We used the predicted class score by the AlexNet to rank the images within one scene category as follow: for a given category with too few exemplars, the top ranked images with predicted class confidence higher than 0.8 were sent to AMT for a third round of manual annotation using the same interface shown in Fig.4. The default answer was set to *NO*.

After completing this third round of AMT annotation, the distribution of the number of images per category flattened out: 401 scene categories had more than 5,000 images per category and 240 scene categories had more than 20,000 images. In total, about 3 million images were added into the dataset.

2.2.4 Step 4: Improving the separation of similar classes

Despite the initial effort to bundle synonyms from WordNet, the scene list from the SUN database still contained a few categories with very close synonyms (e.g. 'ski lodge' and 'ski resort', or 'garbage dump' and 'landfill'). We manually identified 46 synonym pairs like these and merged their images into a single category.

Additionally, we observed that some scene categories could be easily confused with blurry categorical boundaries, as illustrated in Fig. 5. This means that, for images in these blurry boundaries, answering the question "Does image I belong to class A?" might be difficult. However, it can be easier to answer the question "Does image I belong to class A or B?". With this question, the decision boundary becomes clearer for a human observer and it also gets closer to the final task that a computer system will be trained to solve, which is actually separating classes even when the boundaries are blurry.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 5. Boundaries between place categories can be blurry, as some images can be made of a mixture of different components. The images shown in this figure show a soft transition between a field and a forest. Although the extreme images can be easily classified as field and forest scenes, the middle images can be ambiguous.

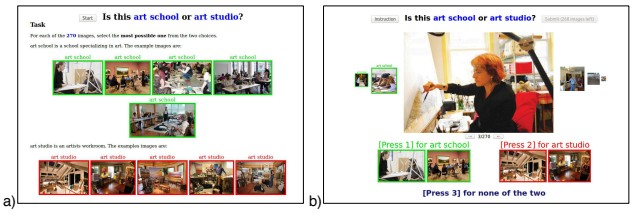


Fig. 6. Annotation interface in Amazon Mechanical Turk for differentiating images from two similar categories. a) instruction in which we give several typical examples in each category. b) the binary selection interface, in which the worker has to classify the shown image into one of the classes or none of them.

After checking the annotations, we confirmed that in the previous three steps of the AMT annotation, workers were confused with some pairs (or groups) of scene categories. For instance, there was an overlap between 'canyon' and 'mountain' or 'butte' and 'mountain'. There were also images mixed in the following category pairs: 'jacuzzi' and 'swimming pool indoor'; 'pond' and 'lake'; 'volcano' and 'mountain'; 'runway' and 'highway and road'; 'operating room' and 'hospital room'; among others. In the whole set of categories, we identified 53 different ambiguous pairs.

To further differentiate the images from the categories with shared content, we designed a new interface for a fourth annotation step. The instructions for the task are shown Fig. 6.a, while Fig. 6.b shows the annotation interface. The interface combines exemplar images from the two categories with shared content (such as 'art school' and 'art studio'), and AMT workers were asked to classify images into one of the categories or neither of them.

After this fourth annotation step, the Places database was finalized with over 10 millions labeled exemplars (10,624,928 images) from 434 place categories.

3 Places Benchmarks

Here we describe four subsets of Places database as benchmarks. Places205 and Places88 are from [1]. Two new benchmarks have been added: from the 434 categories, we selected 365 categories with more than 4000 images each to create *Places365-Standard* and *Places365-Challenge*. The details of each benchmark are the following:

- **Places365-Standard** has 1,803,460 training images with the image number per class varying from 3,068 to 5,000. The validation set has 50 images per class and the test set has 900 images per class. Note that the experiments in this paper are reported on Places365-Standard.
- **Places365-Challenge** contains the same categories as *Places365-Standard*, but the training set is significantly larger with a total of 8 million training images. The validation set and testing set are the same as the Places365-Standard. This subset was released for the Places Challenge 2016[2] held in conjunction with the European Conference on Computer Vision (ECCV) 2016, as part of the ILSVRC Challenge.
- **Places205**. Places205, described in [1], has 2.5 million images from 205 scene categories. The image number per class varies from 5,000 to 15,000. The training set has 2,448,873 total images, with 100 images per category for the validation set and 200 images per category for the test set.
- **Places88**. Places88 contains the 88 common scene categories among the ImageNet [12], SUN [16] and Places205 databases. Places88 contains only the images obtained in round 2 of annotations, from the first version of Places used in [1]. We call the other two corresponding subsets ImageNet88 and SUN88 respectively. These subsets are used to compare performances across different scene-centric databases, as the three datasets contain different exemplars per category (i.e. none of these three datasets contain common images). Note that finding correspondences between the classes defined in ImageNet and Places brings some challenges. ImageNet follows the WordNet definitions, but some WordNet definitions are not always appropriate for describing places. For instance, the class 'elevator' in ImageNet refers to an object. In Places, 'elevator' takes different meanings depending on the location of the observer: elevator door, elevator interior, or elevator lobby. Many categories in ImageNet do not differentiate between indoor and outdoor (e.g., ice-skating rink) while in Places, indoor and outdoor versions are separated as they do not necessarily afford the same function.

4 Comparing Scene-centric Datasets

Scene-centric datasets correspond to images labeled with a scene, or place name, as opposed to object-centric datasets, where images are labeled with object names. In this section we use the Places88 benchmark to compare Places dataset with the tow other biggest scene datasets: ImageNet88 and SUN88. Fig. 7 illustrates the differences among the number of images found in the different categories for Places88, ImageNet88 and SUN88. Notice that Places Database is

2. http://places2.csail.mit.edu/challenge.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

the largest scene-centric image dataset so far. The next subsection presents a comparison of these three datasets in terms of image diversity.

4.1 Dataset Diversity

Given the types of images found on the internet, some categories will be more biased than others in terms of viewpoints, types of objects, or even image style [23]. However, bias can be compensated with a high diversity of images, with many appearances represented in the dataset. In this section we describe a measure of dataset diversity to compare how diverse images from three scene-centric datasets (Places88, SUN88 and ImageNet88) are.

Comparing datasets is an open problem. Even datasets covering the same visual classes have notable differences providing different generalization performances when used to train a classifier [23]. Beyond the number of images and categories, there are aspects that are important but difficult to quantify, like the variability in camera poses, in decoration styles or in the type of objects that appear in the scene.

Although the quality of a database is often task dependent, it is reasonable to assume that a good database should be **dense** (with a high degree of data concentration), and **diverse** (it should include a high variability of appearances and viewpoints). Imagine, for instance, a dataset composed of 100,000 images all taken within the same bedroom. This dataset would have a very high density but a very low diversity as all the images will look very similar. An ideal dataset, expected to generalize well, should have high *diversity* as well. While one can achieve high density by collecting a large number of images, diversity is not an obvious quantity to estimate in image sets, as it assumes some notion of similarity between images. One way to estimate similarity is to ask the question *are these two images similar?* However, similarity in the wild is a subjective and loose concept, as two images can be viewed as similar if they contain similar objects, and/or have similar spatial configurations, and/or have similar decoration styles and so on. A way to circumvent this problem is to define *relative measures* of similarity for comparing datasets.

Several measures of diversity have been proposed, particularly in biology for characterizing the richness of an ecosystem (see [24] for a review). Here, we propose to use a measure inspired by the *Simpson index of diversity* [25]. The Simpson index measures the probability that two random individuals from an ecosystem belong to the same species. It is a measure of how well distributed the individuals across different species are in an ecosystem, and it is related to the entropy of the distribution. Extending this measure for evaluating the diversity of images within a category is non-trivial if there are no annotations of subcategories. For this reason, we propose to measure the relative diversity of image datasets A and B based on the following idea: if set A is more diverse than set B, then two random images from set B are more likely to be visually similar than two random samples from A. Then, the diversity of A with respect to B can be defined as

$$\mathrm{Div}_B(A) = 1 - p(d(a_1, a_2) < d(b_1, b_2)) \quad (1)$$

where $a_1, a_2 \in A$ and $b_1, b_2 \in B$ are randomly selected. With this definition of relative diversity we have that A is more diverse than B if, and only if, $\mathrm{Div}_B(A) > \mathrm{Div}_A(B)$.

For an arbitrary number of datasets, $A_1, ..., A_N$, the diversity of A_1 with respect to $A_2, ..., A_N$ can be defined as

$$\mathrm{Div}_{A_2,...,A_N}(A_1) = 1 - p(d(a_{11}, a_{12}) < \min_{i=2:N} d(a_{i1}, a_{i2})) \quad (2)$$

where $a_{i1}, a_{i2} \in A_i$ are randomly selected, $i = 2 : N$.

We measured the relative diversities between SUN88, ImageNet88 and Places88 using AMT. Workers were presented with different pairs of images and they had to select the pair that contained the most similar images. The pairs were randomly sampled from each database. Each trial was composed of 4 pairs from each database, giving a total of 12 pairs to choose from. We used 4 pairs per database to increase the chances of finding a similar pair and avoiding users having to skip trials. AMT workers had to select the most similar pair on each trial. We ran 40 trials per category and two observers per trial, for the 88 categories in common between ImageNet88, SUN88 and Places88 databases. Fig. 8.a-b shows some examples of pairs from the diversity experiments for the scene categories playground (a) and bedroom (b). In the figure only one pair from each database is shown. We observed that different annotators were consistent in deciding whether a pair of images was more similar than another pair of images.

Fig. 8.c shows the histograms of relative diversity for all the 88 scene categories common to the three databases. If the three datasets were identical in terms of diversity, the average diversity should be 2/3 for the three datasets. Note that this measure of diversity is a relative measure between the three datasets. In the experiment, users selected pairs from the SUN database to be the closest to each other 50% of the time, while the pairs from the Places database were judged to be the most similar only on 17% of the trials. ImageNet88 pairs were selected 33% of the time.

The results show that there is a large variation in terms of diversity among the three datasets, showing Places to be the most diverse of the three datasets. The average relative diversity on each dataset is 0.83 for Places, 0.67 for ImageNet88 and 0.50 for SUN. The categories with the largest variation in diversity across the three datasets were *playground*, *veranda* and *waiting room*.

4.2 Cross Dataset Generalization

As discussed in [23], training and testing across different datasets generally results in a drop of performance due to the dataset bias problem. In this case, the bias between datasets is due, among other factors, to the differences in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

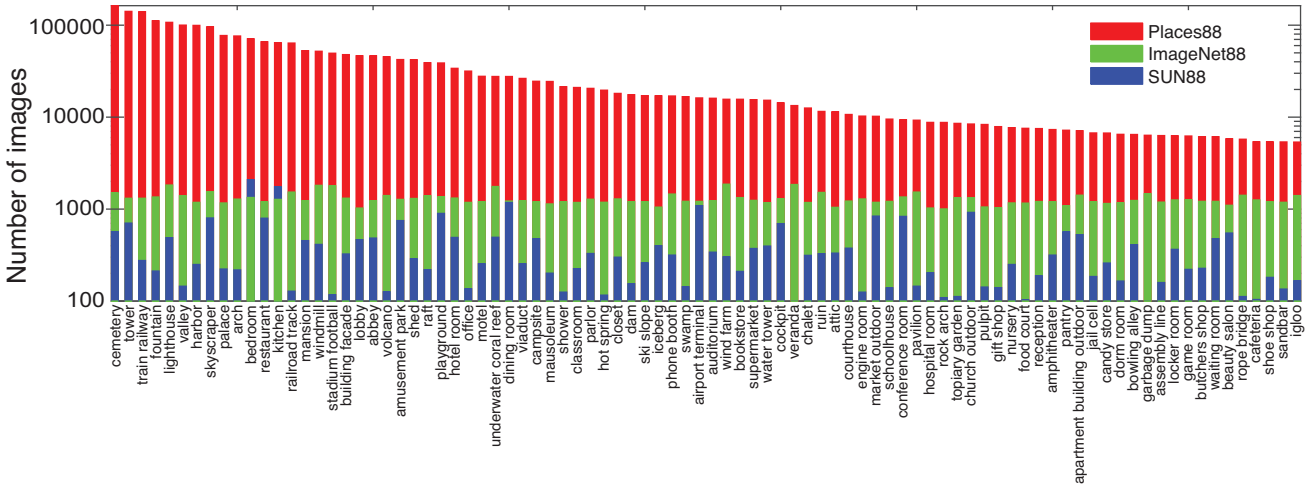
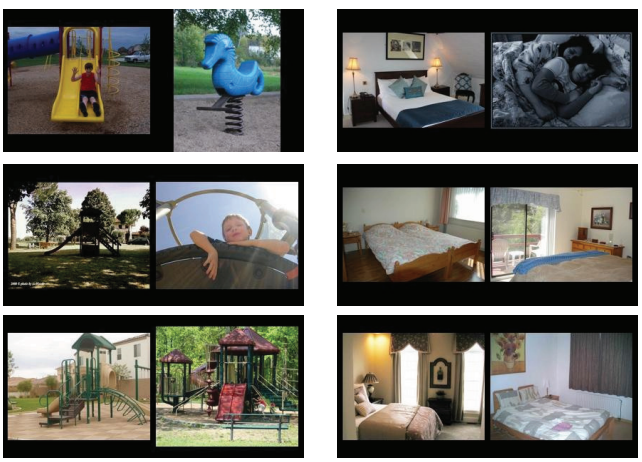


Fig. 7. Comparison of the number of images per scene category for the common 88 scene categories in Places, ImageNet, and SUN datasets.



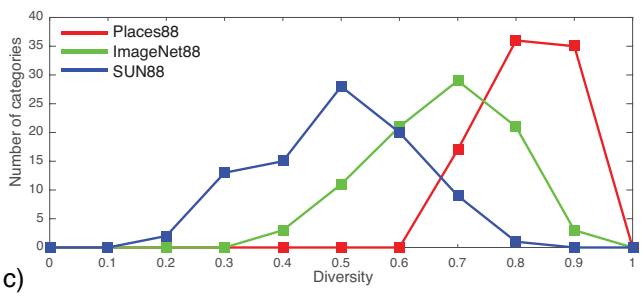


Fig. 8. Examples of pairs for the diversity experiment for a) playground and b) bedroom. Which pair shows the most similar images? The bottom pairs were chosen in these examples. c) Histogram of relative diversity per each category (88 categories) and dataset. Places88 (in blue line) contains the most diverse set of images, then ImageNet88 (in red line) and the lowest diversity is in the SUN88 database (in yellow line) as most images are prototypical of their class.

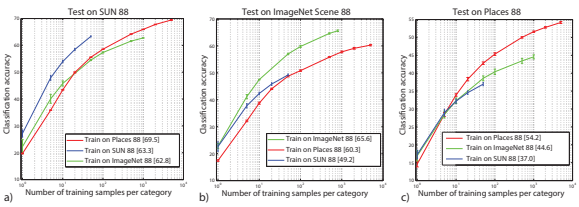


Fig. 9. Cross dataset generalization of training on the 88 common scenes between Places, SUN and ImageNet then testing on the 88 common scenes from: a) SUN, b) ImageNet and c) Places database.

the diversity between the three datasets. Fig. 9 shows the classification results obtained from the training and testing on different permutations of the 3 datasets. For these results we use the features extracted from a pre-trained ImageNet-CNN and a linear SVM. In all three cases training and testing on the same dataset provides the best performance for a fixed number of training examples. As the Places database is very large, it achieves the best performance on two of the test sets when all the training data is used.

5 Convolutional Neural Networks for Scene Classification

Given the impressive performance of the deep Convolutional Neural Networks (CNNs), particularly on the ImageNet benchmark [3], [12], we choose three popular CNN architectures, **AlexNet** [3], **GoogLeNet** [26], and **VGG** 16 convolutional-layer CNN [27], then train them on **Places205** and **Places365-Standard** respectively to create baseline CNN models. The trained CNNs are named as Places*Subset-CNN*, i.e., Places205-AlexNet or Places365-VGG.

All the Places-CNNs presented here were trained using the Caffe package [28] on Nvidia GPUs Tesla K40 and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

Titan X[3]. Additionally, given the recent breakthrough performances of the Residual Network (ResNet) on ImageNet classification [29], we further fine-tuned **ResNet152** on the Places365-Standard (termed as Places365-ResNet) and compared it with the other trained-from-scratch Places-CNNs for scene classification.

5.1 Results on Places205 and Places365

After training the various Places-CNNs, we used the final output layer of each network to classify the test set images of Places205 and SUN205 (see [1]). The classification results for Top-1 accuracy and Top-5 accuracy are listed in Table 1. The Top-1 accuracy is the percentage of the testing images where the top predicted label exactly match the ground-truth label. The Top-5 accuracy is that the percentage of testing images where the ground-truth label is among the top ranked 5 predicted labels given by an algorithm. Since there are some ambiguity between some scene categories, the Top-5 accuracy is a more suitable criteria of measuring scene classification performance.

As a baseline comparison, we show the results of a linear SVM trained on ImageNet-CNN features of 5000 images per category in Places205 and 50 images per category in SUN205 respectively. We observe that Places-CNNs perform much better than the ImageNet feature+SVM baseline while, as expected, Places205-GoogLeNet and Places205-VGG outperformed Places205-AlexNet with a large margin, due to their deeper structures. To date (Oct 2, 2016) the top ranked results on the test set of Places205 leaderboard[4] is 64.10% on Top-1 accuracy and 90.65% on Top-5 accuracy. Note that for the test set of SUN205, we did not fine-tune the Places-CNNs on the training set of SUN205, as we directly evaluated them on the test set of SUN.

We further evaluated the baseline Places365-CNNs on the validation set and test set of Places365. The results are shown in Table 2. We can see that Places365-VGG and Places365-ResNet have similar top performances compared with the other two CNNs[5]. Even though Places365 has 160 more categories than Places205, the Top-5 accuracy of the Places205-CNNs (trained on the previous version of Places [1]) on the test set only drops by 2.5%.

To evaluate how extra categories bring improvements, we compute the accuracy for the 182 common categories between Places205 and Places365 (we merge some categories in Places205 when building Places365 thus there are less common categories) for Places205-CNN and Places365-CNN. On the validation set of Places365, we select the images of the 182 common categories, then use the aligned 182 outputs of the Places205-AlexNet and Places365-AlexNet to predict the labels respectively. The Top1 accuracy for Places205-AlexNet is 0.572, the one for Places365-AlexNet is 0.577. Thus Places365-AlexNet not only predicts more categories, but also has better accuracy on the previous existing categories.

Fig.10 shows the responses to examples correctly predicted by the Places365-VGG. Notice that most of the Top-5 responses are very relevant to the scene description. Some failure or ambiguous cases are shown in Fig.11. Broadly, we can identify two kinds of mis-classification given the current label attribution of Places: 1) less-typical activities happening in a scene, such as taking group photo in a construction site and camping in a junkyard; 2) images composed of multiple scene parts, which make one ground-truth scene label not sufficient to describe the whole environment. These results suggest the need to have multi-ground truth labels for describing environments.

It is important to emphasize that for many scene categories the Top-1 accuracy might be an ill-defined measure: environments are inherently multi-labels in terms of their semantic description. Different observers will use different terms to refer to the same place, or different parts of the same environment, and all the labels might fit well the description of the scene, as we observe in the examples of Fig.11.

5.2 Web-demo for Scene Recognition

Based on our trained Places-CNN, we created a web-demo for scene recognition[6], accessible through a computer browser or mobile phone. People can upload photos to the web-demo to predict the type of environment, with the 5 most likely semantic categories, and relevant scene attributes. Two screenshots of the prediction result on the mobile phone are shown in Fig.12. Note that people can submit feedback about the result. The top-5 recognition accuracy of our recognition web-demo in the wild is about 72% (from the 9,925 anonymous feedbacks dated from Oct.19, 2014 to May 5, 2016), which is impressive given that people uploaded all kinds of photos from real-life and not necessarily places-like photos. Places205-AlexNet is the back-end prediction model in the demo.

5.3 Places365 Challenge Result

The subset **Places365-Challenge**, which contains more than 8 million images from 365 scene categories, was used in the Places Challenge 2016 held as part of the ILSVRC Challenge in the European Conference on Computer Vision (ECCV) 2016.

The rule of the challenge is that each team can only use the provided data in the Places365-Challenge to train their networks. Standard CNN models trained on Imagenet-1.2million and previous Places are allowed to use. Each team can submit at most 5 prediction results. Ranks are based on the top-5 classification error of each submission. Winners teams are then invited to give talks at the ILSVRC-COCO Joint Workshop at ECCV'16.

3. All the Places-CNNs are available at https://github.com/CSAILvision/places365

4. http://places.csail.mit.edu/user/leaderboard.php

5. The performance of the ResNet might result from fine-tuning or under-training, as the ResNet is not trained from scratch.

6. http://places.csail.mit.edu/demo.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Classification accuracy on the test set of Places205 and the test set of SUN205. We use the class score averaged over 10-crops of each test image to classify the image. $*$ shows the top 2 ranked results from the Places205 leaderboard.

	Test set of Places205		Test set of SUN205	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
ImageNet-AlexNet feature+SVM	40.80%	70.20%	49.60%	80.10%
Places205-AlexNet	50.04%	81.10%	67.52%	92.61%
Places205-GoogLeNet	55.50%	85.66%	71.6%	95.01%
Places205-VGG	**58.90%**	**87.70%**	**74.6%**	**95.92%**
SamExynos$*$	64.10%	90.65%	-	-
SIAT MMLAB$*$	62.34%	89.66%	-	-

TABLE 2
Classification accuracy on the validation set and test set of Places365. We use the class score averaged over 10-crops of each testing image to classify the image.

	Validation Set of Places365		Test Set of Places365	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
Places365-AlexNet	53.17%	82.89%	53.31%	82.75%
Places365-GoogLeNet	53.63%	83.88%	53.59%	84.01%
Places365-VGG	**55.24%**	84.91%	**55.19%**	85.01%
Places365-ResNet	54.74%	**85.08%**	54.65%	**85.07%**

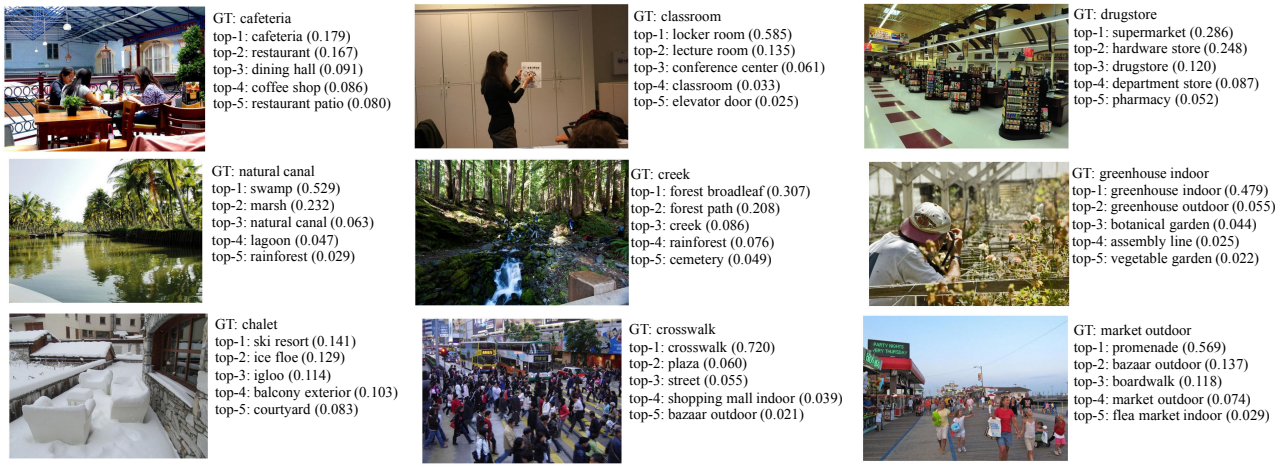


Fig. 10. The predictions given by the Places365-VGG for the images from the validation set. The ground-truth label (GT) and the top 5 predictions are shown. The number beside each label indicates the prediction confidence.

There were totally 92 valid submissions from 27 teams. Finally team *Hikvision* [30] won the 1st place with 9.01% top-5 error, team *MW* [31] won the 2nd place with 10.19% top-5 error, and team *Trimps-Soushen* [32] won the 3rd place with 10.30% top-5 error. The leaderboard and the team information are available at the challenge result page[7]. The ranked results of all the submissions are plotted in Fig.13. The entry from the winner team outperforms our best baseline with a large margin ($\sim 6\%$ in top-5 accuracy). Note that our baselines are trained with the Places365-Standard while those challenge entries are trained on the Places365-Challenge which has 5.5 million more training images.

5.4 Generic Visual Features from Places-CNNs and ImageNet-CNNs

We further used the activation from the trained Places-CNNs as generic features for visual recognition tasks using different image classification benchmarks. Activations from the higher-level layers of a CNN, also termed *deep features*, have proven to be effective generic features with state-of-the-art performance on various image datasets [33], [34]. But most of the deep features are from the CNNs trained on ImageNet, which is mostly an object-centric dataset.

Here we evaluated the classification performances of the deep features from scene-centric CNNs and object-

7. http://places2.csail.mit.edu/results2016.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

GT: construction site
top-1: martial arts gym (0.157)
top-2: stable (0.156)
top-3: boxing ring (0.091)
top-4: locker room (0.090)
top-5: basketball court (0.056)

GT: junkyard
top-1: campsite (0.306)
top-2: sandbox (0.276)
top-3: beer garden (0.052)
top-4: market outdoor (0.035)
top-5: flea market indoor (0.033)

GT: aquarium
top-1: restaurant (0.213)
top-2: ice cream parlor (0.139)
top-3: coffee shop (0.138)
top-4: pizzeria (0.085)
top-5: cafeteria (0.078)

GT: lagoon
top-1: balcony interior (0.136)
top-2: beach house (0.134)
top-3: boardwalk (0.123)
top-4: roof garden (0.103)
top-5: restaurant patio (0.068)

Fig. 11. Examples of predictions rated as incorrect in the validation set by the Places365-VGG. GT states for ground truth label. Note that some of the top-5 responses are often not wrong per se, predicting semantic categories near by the GT category. See the text for details.

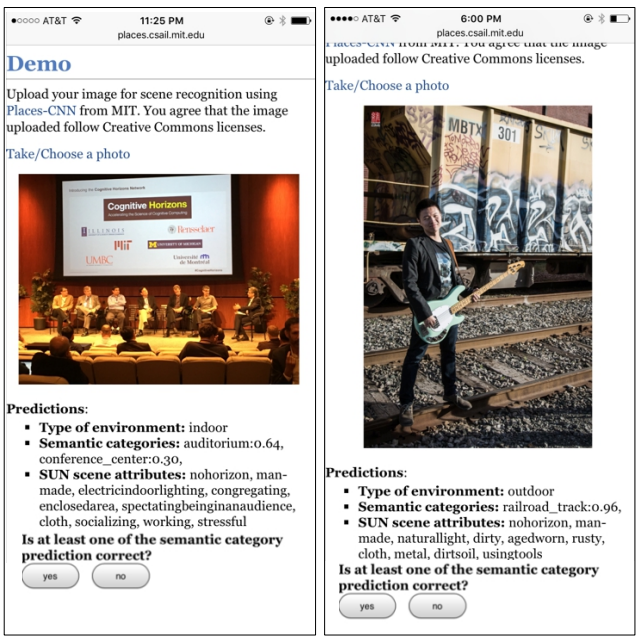


Fig. 12. Two screenshots of the scene recognition demo based on the Places-CNN. The web-demo predicts the type of environment, the semantic categories, and associated scene attributes for uploaded photos.

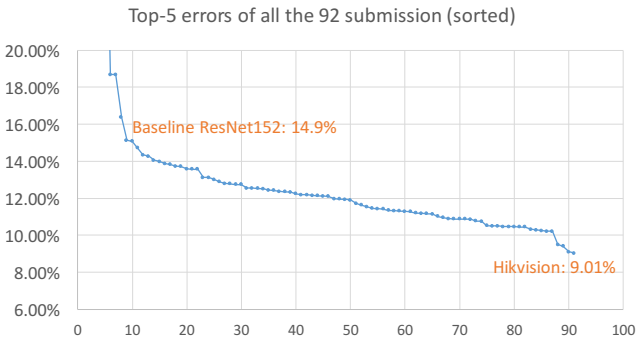


Fig. 13. The ranked results of all the 92 valid submissions. The best baseline trained on Places365-standard is the Resnet152 which has the top5-error as 14.9%, while the winner network from HikVision gets 9.01% top5-error which outperform the baseline with large margin.

centric CNNs in a systematic way. The deep features from several Places-CNNs and ImageNet-CNNs on the following scene and object benchmarks are tested: SUN397 [16], MIT Indoor67 [15], Scene15 [13], SUN Attribute [35], Caltech101 [36], Caltech256 [37], Stanford Action40 [38], and UIUC Event8 [39].

All of the presented experiments follow the standards in the mentioned papers. In the SUN397 experiment [16], the training set size is 50 images per category. Experiments were run on 5 splits of the training set and test set given in the dataset. In the MIT Indoor67 experiment [15], the training set size is 100 images per category. The experiment is run on the split of the training set and test set given in the dataset. In the Scene15 experiment [13], the training set size is 50 images per category. Experiments are run on 10 random splits of the training set and test set. In the SUN Attribute experiment [35], the training set size is 150 images per attribute. The reported result is the average precision. The splits of the training set and test set are given in the paper. In Caltech101 and Caltech256 experiment [36], [37], the training set size is 30 images per category. The experiments are run on 10 random splits of the training set and test set. In the Stanford Action40 experiment [38], the training set size is 100 images per category. Experiments are run on 10 random splits of the training set and test set. The reported result is the classification accuracy. In the UIUC Event8 experiment [39], the training set size is 70 images per category and the test set size is 60 images per category. The experiments are run on 10 random splits of the training set and test set.

Places-CNNs and ImageNet-CNNs have the same network architectures for AlexNet, GoogLeNet, and VGG, but they are trained on scene-centric data (Places) and object-centric data (ImageNet) respectively. For AlexNet and VGG, we used the 4096-dimensional feature vector from the activation of the Fully Connected Layer (`fc7`) of the CNN. For GoogLeNet, we used the 1024-dimensional feature vector from the response of the global average pool-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

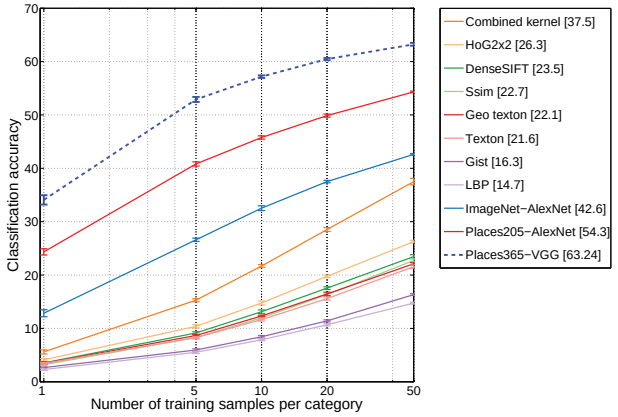


Fig. 14. Classification accuracy on the SUN397 Dataset. We compare the deep features of Places365-VGG, Places205-AlexNet (result reported in [1]), and ImageNet-AlexNet, to hand-designed features (HOG, gist, etc). The deep features of Places365-VGG outperforms other deep features and hand-designed features by a large margin. Results of other hand-designed features/kernels are fetched from [16].

ing layer before softmax producing the class predictions. The classifier in all of the experiments is a linear SVM with the default parameter for all of the features.

Table 3 summarizes the classification accuracy on various datasets for the deep features of Places-CNNs and the deep features of the ImageNet-CNNs. Fig.14 plots the classification accuracy for different visual features on the SUN397 database over different numbers of training samples per category. The classifier is a linear SVM with the same default parameters for the two deep feature layers (C=1) [40]. The Places-CNN features show impressive performance on scene-related datasets, outperforming the ImageNet-CNN features. On the other hand, the ImageNet-CNN features show better performance on object-related image datasets. Importantly, our comparison shows that Places-CNN and ImageNet-CNN have complementary strengths on scene-centric tasks and object-centric tasks, as expected from the type of the datasets used to train these networks. On the other hand, the deep features from the Places365-VGG achieve the best performance (63.24%) on the most challenging scene classification dataset SUN397, while the deep features of Places205-VGG performs the best on the MIT Indoor67 dataset. As far as we know, they are the state-of-the-art scores achieved by a single feature + linear SVM on those two datasets. Furthermore, we merge the 1000 classes from the ImageNet and the 365 classes from the Places365-Standard to train a VGG (Hybrid1365-VGG). The deep feature from the Hybrid1365-VGG achieves the best score averaged over all the eight image datasets.

5.5 Visualization of the Internal Units

Through the visualization of the unit responses for various levels of network layers, we can have a better understanding of what has been learned inside CNNs and what are the differences between the object-centric CNN trained on ImageNet and the scene-centric CNN trained on Places given that they share the same architecture AlexNet. Following the methodology in [2] we feed 100,000 held-out testing images into the two networks and record the max activation of each unit pooled over the whole spatial feature map for each of the images respectively. For each unit, we get the top three ranked images by ranking their max activations, then we segment the images by bilinear upsampling the binarized spatial feature map mask.

The image segmentation results of the units from different layers are shown in Fig.15. We can see that from `conv1` to `conv5`, the units detect visual concepts from low-level edge/texture to high-level object/scene parts. Furthermore, in the object-centric ImageNet-CNN there are more units detecting object parts such as dog and people's heads in the `conv5` layer, while in the scene centric Places-CNN there are more units detecting scene parts such as bed, chair, or buildings in the `conv5` layer.

Thus the specialty of the units in the object-centric CNN and scene-centric CNN yield very different performances of generic visual features on a variety of recognition benchmarks (object-centric datasets vs scene-centric datasets) in Table 3.

We further synthesized preferred input images for the Places-CNN by using the image synthesis technique proposed in [41]. This method uses a learned prior deep generator network to generate images which maximize the final class activation or the intermediate unit activation of the Places-CNN. The synthetic images for 50 scene categories are shown in Fig.16. These abstract image contents reveal the knowledge of the specific scene learned and memorized by the Places-CNN: examples include the buses within a road environment in the bus station, and the tents surrounded by forest-types of features for the campsite. Here we used Places365-AlexNet (other Places365-CNNs generated similar results). We further used the synthesis technique to generate the images preferred by the units in the `conv5` layer of Places365-AlexNet. As shown in Fig.17, the synthesized images are very similar to the segmented image regions by the estimated receptive field of the units.

6 Conclusion

From the Tiny Image dataset [42], to ImageNet [11] and Places [1], the rise of multi-million-item dataset initiatives and other densely labeled datasets [18], [20], [43], [44] have enabled data-hungry machine learning algorithms to reach near-human semantic classification of visual patterns, like objects and scenes. With its category coverage and high-diversity of exemplars, Places offers an ecosystem of visual context to guide progress on scene understanding problems. Such problems could include determining the actions happening in a given environment, spotting inconsistent objects or human behaviors for a particular place, and predicting future events or the cause of events given a scene.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Classification accuracy/precision on scene-centric databases (the first four datasets) and object-centric databases (the last four datasets) for the deep features of various Places-CNNs and ImageNet-CNNs. All the accuracy/precision is the top-1 accuracy/precision.

Deep Feature	SUN397	MIT Indoor67	Scene15	SUN Attribute	Caltech101	Caltech256	Action40	Event8	Average
Places365-AlexNet	**56.12**	**70.72**	89.25	92.98	66.40	46.45	46.82	90.63	69.92
Places205-AlexNet	54.32	68.24	**89.87**	**92.71**	65.34	45.30	43.26	94.17	69.15
ImageNet-AlexNet	42.61	56.79	84.05	91.27	**87.73**	**66.95**	**55.00**	**93.71**	72.26
Places365-GoogLeNet	**58.37**	73.30	**91.25**	**92.64**	61.85	44.52	47.52	91.00	70.06
Places205-GoogLeNet	57.00	**75.14**	90.92	92.09	54.41	39.27	45.17	92.75	68.34
ImageNet-GoogLeNet	43.88	59.48	84.95	90.70	**89.96**	**75.20**	**65.39**	**96.13**	75.71
Places365-VGG	**63.24**	76.53	**91.97**	**92.99**	67.63	49.20	52.90	90.96	73.18
Places205-VGG	61.99	**79.76**	91.61	92.07	67.58	49.28	53.33	93.33	73.62
ImageNet-VGG	48.29	64.87	86.28	91.78	**88.42**	**74.96**	**66.63**	**95.17**	77.05
Hybrid1365-VGG	61.77	79.49	92.15	92.93	88.22	76.04	68.11	93.13	**81.48**

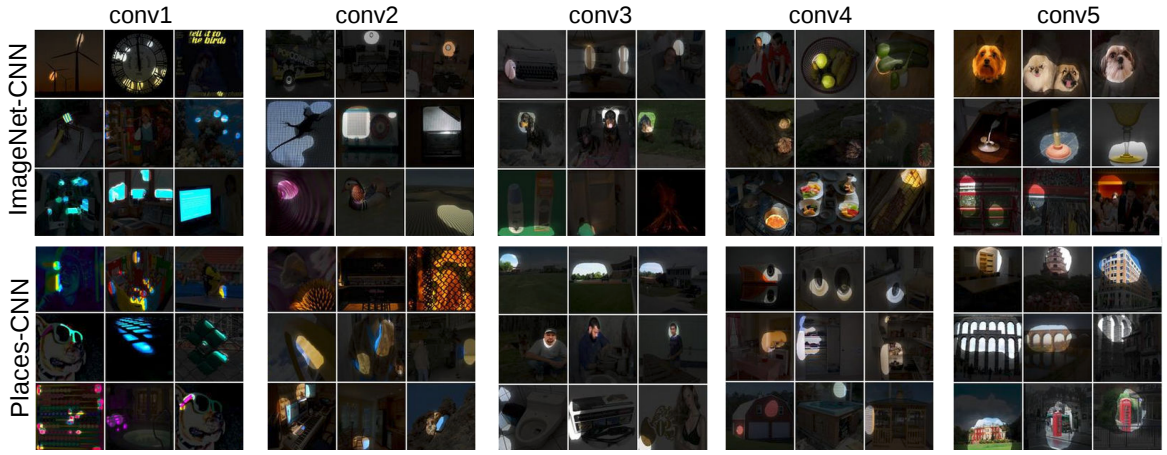


Fig. 15. a) Visualization of the units' receptive fields at different layers for the ImageNet-CNN and Places-CNN. Subsets of units at each layer are shown. In each row we show the top 3 most activated images. Images are segmented based on the binarized spatial feature map of the units at different layers of ImageNet-CNN and Places-CNN. Here we take ImageNet-AlexNet and Places205-AlexNet as the comparison examples. See the detailed visualization methodology in [2].

ACKNOWLEDGMENTS

The authors would like to thank Santani Teng, Zoya Bylinskii, Mathew Monfort and Caitlin Mullin for comments on the paper. Over the years, the Places project was supported by the National Science Foundation under Grants No. 1016862 to A.O and No. 1524817 to A.T; the Vannevar Bush Faculty Fellowship program sponsored by the Basic Research Office of the Assistant Secretary of Defense for Research and Engineering and funded by the Office of Naval Research through grant N00014-16-1-3116 to A.O.; the MIT Big Data Initiative at CSAIL, the Toyota Research Institute / MIT CSAIL Joint Research Center, Google, Xerox and Amazon Awards, and a hardware donation from NVIDIA Corporation, to A.O and A.T. B.Z is supported by a Facebook Fellowship.

REFERENCES

[1] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *In Advances in Neural Information Processing Systems*, 2014.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *International Conference on Learning Representations*, 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." in *In Advances in Neural Information Processing Systems*, 2012.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[5] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 1999.

[6] M. Campbell, A. J. Hoane, and F.-h. Hsu, "Deep blue," *Artificial intelligence*, 2002.

[7] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: beyond jeopardy!" *Artificial Intelligence*, 2013.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, 2016.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. CVPR*, 2015.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int'l Journal of Computer Vision*, 2015.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

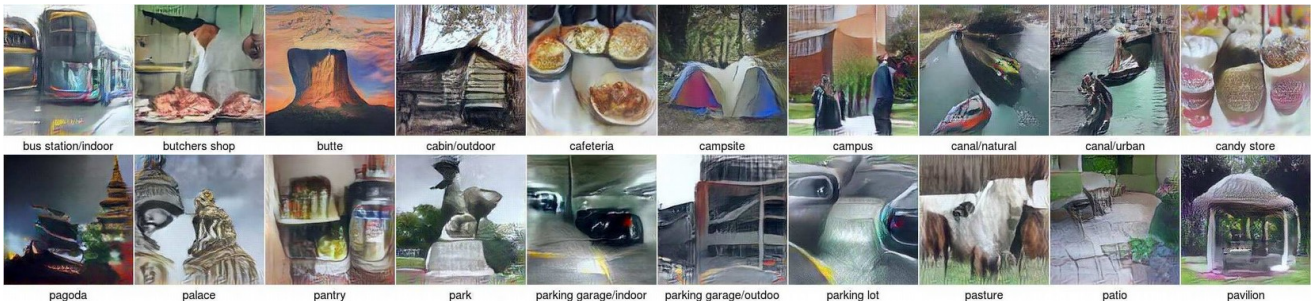


Fig. 16. The synthesized images preferred by the final output of Places365-AlexNet for 20 scene categories.

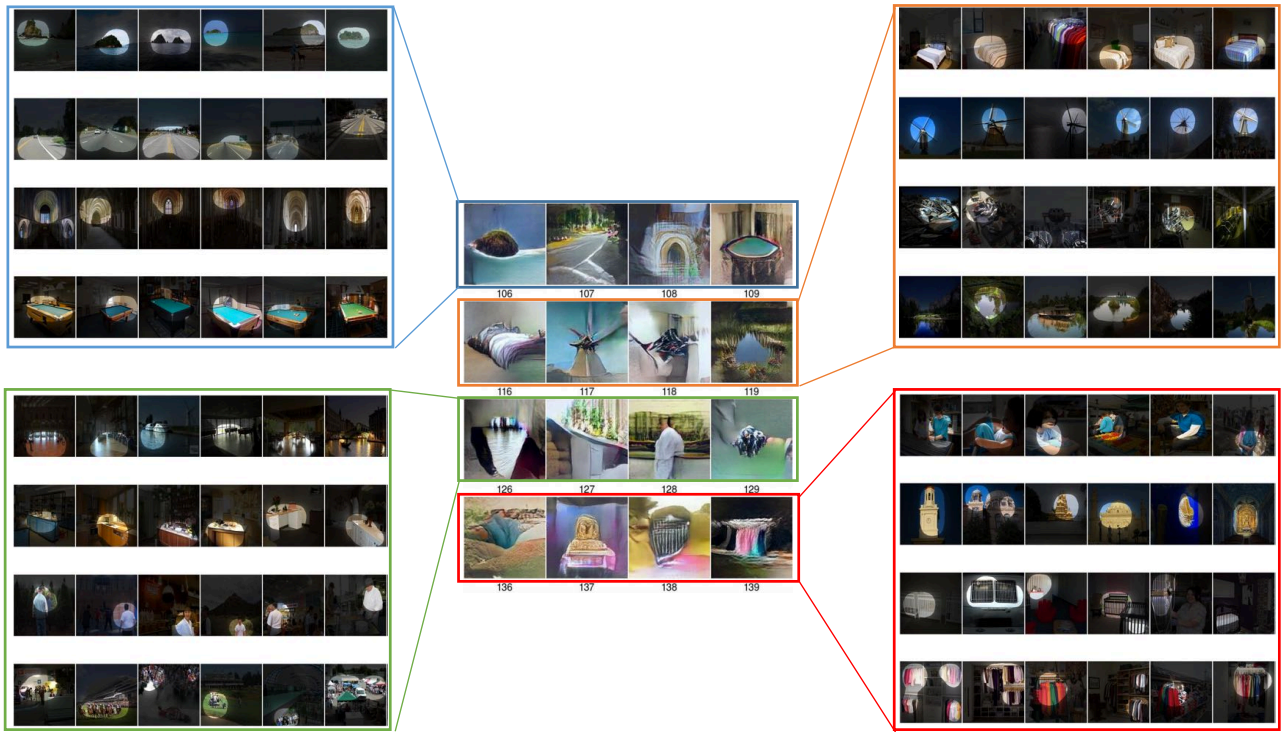


Fig. 17. The synthesized images preferred by the `conv5` units of the Places365-AlexNet corresponds to the segmented images by the receptive fields of those units. The synthetic images are very similar to the segmented image regions of the units. Each row of the segmented images correspond to one unit.

Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006.

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int'l Journal of Computer Vision*, 2001.

[15] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. CVPR*, 2009.

[16] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. CVPR*, 2010.

[17] M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó *et al.*, "The pascal visual object classes challenge 2007 (voc2007) results," 2007.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *ijcv*, 2016.

[20] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene paring through ade20k dataset," *Proc. CVPR*, 2017.

[21] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[22] P. Jolicoeur, M. A. Gluck, and S. M. Kosslyn, "Pictures and names: Making the connection," *Cognitive psychology*, 1984.

[23] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, 2011.

[24] C. Heip, P. Herman, and K. Soetaert, "Indices of diversity and evenness," *Oceanis*, 1998.

[25] E. H. Simpson, "Measurement of diversity." *Nature*, 1949.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proc. CVPR*, 2015.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2014.

[28] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," http://caffe.berkeleyvision.org/, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2723009, IEEE Transactions on Pattern Analysis and Machine Intelligence

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. CVPR*, 2016.
[30] Q. Zhong, C. Li, Y. Zhang, H. Sun, S. Yang, D. Xie, and S. Pu, "Towards good practices for recognition and detection," http://image-net.org/challenges/talks/2016/Hikvision_at_ImageNet_2016.pdf, 2016.
[31] L. Shen, Z. Lin, G. Sun, and J. Hu, "Places401 and places365 models," https://github.com/lishen-shirley/Places2-CNNs, 2016.
[32] J. Shao, X. Zhang, Z. Ding, Y. Zhao, Y. Chen, J. Zhou, W. Wang, L. Mei, and C. Hu, "Good pratices for deep feature fusion," http://image-net.org/challenges/talks/2016/Trimps-Soushen@ILSVRC2016.pdf, 2016.
[33] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014.
[34] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," *CVPR workshop*, 2014.
[35] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. CVPR*, 2012.
[36] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, 2007.
[37] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.
[38] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. ICCV*, 2011.
[39] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Proc. ICCV*, 2007.
[40] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," 2008.
[41] A. Nguyen, A. Dosovitskiy, T. Yosinski, Jason band Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," *In Advances in Neural Information Processing Systems*, 2016.
[42] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2008.
[43] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int'l Journal of Computer Vision*, 2010.
[44] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *Proc. CVPR*, 2016.

Bolei Zhou is a Ph.D. Candidate in Computer Science and Artificial Intelligence Lab (CSAIL) at the Massachusetts Institute of Technology. He received M.Phil. degree in Information Engineering from the Chinese University of Hong Kong and B.Eng. degree in Biomedical Engineering from Shanghai Jiao Tong University in 2010. His research interests are computer vision and machine learning. He is an award recipient of the Facebook Fellowship, the Microsoft Research Asia Fellowship, and the MIT Greater China Fellowship.

Agata Lapedriza is an Associate Professor at the Universitat Oberta de Catalunya. She received her MS deegree in Mathematics at the Universitat de Barcelona in 2003, and her Ph.D. degree in Computer Science at the Computer Vision Center in 2009, at the Universitat Autonoma Barcelona. She was working as a visiting researcher in the Computer Science and Artificial Intelligence Lab, at the Massachusetts Institute of Technology, from 2012 until 2015. Her research interests are related to high-level image understanding, scene and object recognition, and affective computing.

Aditya Khosla received the BS degree in computer science, electrical engineering and economics from the California Institute of Technology, and the MS degree in computer science from Stanford University, in 2009 and 2011 respectively. He completed his PhD in computer science from the Massachusetts Institute of Technology in 2016 with a focus on computer vision and machine learning. In his thesis, he developed machine learning techniques that predict human behavior and the impact of visual media on people.

Aude Oliva is a Principal Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). After a French baccalaureate in Physics and Mathematics, she received two M.Sc. degrees and a Ph.D in Cognitive Science from the Institut National Polytechnique of Grenoble, France. She joined the MIT faculty in the Department of Brain and Cognitive Sciences in 2004 and CSAIL in 2012. Her research on vision and memory is cross-disciplinary, spanning human perception and cognition, computer vision, and human neuroscience. She received the 2006 National Science Foundation (NSF) Career award, the 2014 Guggenheim and the 2016 Vannevar Bush fellowships.

Antonio Torralba received the degree in telecommunications engineering from Telecom BCN, Spain, in 1994 and the Ph.D. degree in signal, image, and speech processing from the Institut National Polytechnique de Grenoble, France, in 2000. From 2000 to 2005, he spent postdoctoral training at the Brain and Cognitive Science Department and the Computer Science and Artificial Intelligence Laboratory, MIT. He is now a Professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT). Prof. Torralba is an Associate Editor of the International Journal in Computer Vision, and has served as program chair for the Computer Vision and Pattern Recognition conference in 2015. He received the 2008 National Science Foundation (NSF) Career award, the best student paper award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2009, and the 2010 J. K. Aggarwal Prize from the International Association for Pattern Recognition (IAPR).