


1 xNet+SC: Classifying Places Based on Images by 2 Incorporating Spatial Contexts

3 **Bo Yan**

4 STKO Lab, University of California, Santa Barbara, USA

5 boyan@geog.ucsb.edu

6  <https://orcid.org/0000-0002-4248-7203>

7 **Krzysztof Janowicz**

8 STKO Lab, University of California, Santa Barbara, USA

9 jano@geog.ucsb.edu

10 **Gengchen Mai**

11 STKO Lab, University of California, Santa Barbara, USA

12 gengchen@geog.ucsb.edu

13 **Rui Zhu**

14 STKO Lab, University of California, Santa Barbara, USA

15 ruizhu@geog.ucsb.edu

16 — Abstract —

17 With recent advancements in deep convolutional neural networks, researchers in geographic in-
18 formation science gained access to powerful models to address challenging problems such as
19 extracting objects from satellite imagery. However, as the underlying techniques are essentially
20 borrowed from other research fields, e.g., computer vision or machine translation, they are often
21 not spatially explicit. In this paper, we demonstrate how utilizing the rich information embedded
22 in spatial contexts (SC) can substantially improve the classification of place types from images
23 of their facades and interiors. By experimenting with different types of spatial contexts, namely
24 spatial relatedness, spatial co-location, and spatial sequence pattern, we improve the accuracy
25 of state-of-the-art models such as ResNet – which are known to outperform humans on the Im-
26 ageNet dataset – by over 40%. Our study raises awareness for leveraging spatial contexts and
27 domain knowledge in general in advancing deep learning models, thereby also demonstrating that
28 theory-driven and data-driven approaches are mutually beneficial.

29 **2012 ACM Subject Classification** Computing methodologies → Computer vision tasks; Com-
30 puting methodologies → Neural networks; Theory of computation → Bayesian analysis

31 **Keywords and phrases** Spatial context, Image classification, Place types, Convolutional neural
32 network, Recurrent neural network

33 **Digital Object Identifier** 10.4230/LIPIcs.GIScience.2018.18

34 **1** Introduction

35 Recent advancements in computer vision models and algorithms have quickly permeated
36 many research domains including GIScience. In remote sensing, computer vision methods
37 facilitate researchers to utilize satellite images to detect geographic features and classify
38 land use [5, 26]. In urban planning, researchers collect Google Street View images and
39 apply computer vision algorithms to study urban change [22]. In cartography, pixel-wise
40 segmentation has been adopted to extract lane boundary from satellite imagery [32] and deep
41 convolutional neural network (CNN) has been utilized to recognize multi-digit house numbers
42 from Google Street View images [10]. These recent breakthroughs in computer vision are



© Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 18; pp. 18:1–18:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

43 achieved, in equal parts, due to advances in deep neural networks as well as the ever-increasing
44 availability of extensive training datasets. For example, the classification error in the latest
45 image classification challenge using the ImageNet dataset is down to about 0.023.¹

46 However, such impressive results do not imply that these models have reached a level
47 in which no further improvement is necessary or meaningful. On the contrary, such deep
48 learning models which primarily depend on visual signals are susceptible to error. In fact,
49 studies have shown that deep (convolutional) neural networks suffer from a lack of robustness
50 to adversarial examples and a tendency towards biases [25]. Researchers have discovered that,
51 by incorporating adversarial perturbations of inputs that are indistinguishable by humans,
52 the most advanced deep learning models which have achieved high accuracy on test sets can
53 be easily fooled [6, 11, 28]. In addition, deep learning models are also vulnerable to biased
54 patterns learned from the available data and these biases usually resemble many unpleasant
55 human behaviors in our society. For instance, modern neural information processing systems
56 such as neural network language models and deep convolutional neural networks have been
57 criticized for amplifying racial and gender biases [3, 4, 25, 33]. Such biases, which can
58 be attributed to a discrepancy between the distribution of prototypical examples and the
59 distribution of more complex real world systems [16], have already caused some public debates.
60 To give a provocative example, almost three years after users revealed that Google erroneously
61 labeled photos of black people as “gorillas”, no robust solutions have been established besides
62 simply removing such labels for now.²

63 The above-mentioned drawbacks are being addressed by improvements to the available
64 training data as well as the used methods [23, 3]. In our work, we follow this line of thought to
65 help improve image classification. In our case, these images depict the facades or interiors of
66 different types of places, such as restaurants, hotels, and libraries. Classifying images by place
67 types is a hard problem in that more often than not the training image data is inadequate to
68 provide a full visual representation of different place types. Solely relying on visual signals,
69 as most deep convolutional neural networks do, falls short in modeling the feature space
70 as a result. To give an intuitive example, facades of restaurants may vary substantially
71 based on the type of restaurant, the target customers, and the surrounding. Their facade
72 may be partially occluded by trees or cars, may be photographed from different angles and
73 at different times of the day, and the image may contain parts of other buildings. Put
74 differently, the principle of spatial heterogeneity implies that there is considerable variation
75 between places of the same type.

76 To address this problem and improve classification accuracy, we propose to go beyond
77 visual stimuli by incorporating spatial contextual information to help offset the visual
78 representational inadequacy. Although data availability is less of an issue nowadays, the biased
79 pattern in the data poses a real challenge, especially as models such as deep convolutional
80 neural networks take a very long time to train. Instead of fine-tuning the parameters (weights)
81 by collecting and labeling more unbiased data, which are very resource-consuming, we take
82 advantage of external information, namely spatial context. There are many different ways
83 one can model such context; in this work, we focus on the types of nearby places. We explore
84 and compare the value of three different kinds of spatial context, namely spatial relatedness,
85 spatial co-location, and spatial sequence pattern.

86 We combine these context models with state-of-the-art deep convolutional neural network
87 models using search re-ranking algorithms and Bayesian methods. The result shows that,
88 by considering more complex spatial contexts, we can improve the classification accuracy

² <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>

89 for different place types. In fact, our results demonstrate that a *spatially explicit* model
90 [9], i.e., taking nearby places into account when predicting the place type from an image,
91 improves the accuracy of leading image classification models by at least 40%. Aside from this
92 substantial increase in accuracy, we believe that our work also contributes to the broader
93 and ongoing discussion about the role of and need for theory, i.e., domain knowledge, in
94 machine learning. Finally, and as indicated in the title, our spatial context (*SC*) models,
95 can be added to any of the popular CNN-based computer vision models such as AlexNet,
96 ResNet, and DenseNet – abbreviated to *xNet* here.

97 The remainder of this paper is organized as follows. Section 2 provides an overview of
98 existing work on spatial context and methods for incorporating spatial information into
99 image classification models. Section 3 presents the image classification tasks and provides
100 information about the convolutional neural network models used in our study. Section 4
101 explains in detail three different levels of spatial context and ways to combine them in image
102 classification models. Section 5 presents the results. Finally, Section 6 concludes the research
103 and points to future directions.

104 **2** Related Work

105 There is a large body of work that utilizes spatial context to improve existing methods and
106 provide deeper insights into the rich semantics of contextual information more broadly. For
107 instance, spatial context has been recognized as a complementary source of information in
108 computational linguistics. By training word embeddings for different place types derived
109 from OpenStreetMap (OSM) and Google Places, Cocos and Callison-Burch [7] suggested that
110 spatial context provides useful information about semantic relatedness. In Points of Interest
111 (POI) recommendation, spatial context has been used to provide latent representations of POI,
112 to facilitate the prediction of future visitors [8], and to recommend similar places [34]. By
113 implementing an information theoretic and distance-lagged augmented spatial context, Yan
114 et al. [30] demonstrated that high-dimensional place type embeddings learned using spatial
115 contexts can reproduce human-level similarity judgments with high accuracy. The study
116 showed that such a spatially explicit Place2Vec model substantially outperforms Word2Vec-
117 based models that utilize a linguistic-style of context. Liu et al. [21] used spatial contexts to
118 measure traffic interactions in urban area. In object detection, Heitz and Koller [13] leveraged
119 spatial contexts in a probabilistic model to improve detection result. Likewise, by embracing
120 the idea that spatial context provides valuable extrinsic signals, our work analyzes different
121 kinds of spatial contexts and tests their ability to improve image classification of place types.

122 Existing work on image classification has realized the importance of including a geographic
123 component. One direction of research focused on enriching images with geospatial data.
124 Baatz et al. [1] took advantage of digital elevation models to help geo-localize images in
125 mountainous terrain. Lin et al. [20] made use of land cover survey data and learned the
126 complex translation relationship between ground level images and overhead imagery to extend
127 the reach of image geo-localization. Instead of estimating a precise geo-tag, Lee et al. [19]
128 trained deep convolutional neural networks to enrich a photo with geographic attributes such
129 as elevation and population density. Another direction of research (which is more similar to
130 our study) focused on utilizing geographic information to facilitate image classification. In
131 order to better understand scenes and improve object region recognition, Yu and Luo [31]
132 exploited information from seasons and location proximity of images using a probabilistic
133 graphical model. Berg et al. [2] combined one-vs-most image classifiers with spatiotemporal
134 class priors to address the problem of distinguishing images of highly similar bird species.

135 Tang et al. [29] encoded geographic features extracted from GPS information of images into
 136 convolutional neural networks to improve classification results.

137 Our work differs from the existing work in that we explicitly exploit the distributional
 138 semantics found in spatial context [30] to improve image classification. Following the linguistic
 139 mantra that one *shall know a word by the company it keeps*, we argue that one can know
 140 a place type by its neighborhood’s types. This raises the interesting question of how such
 141 a neighborhood should be defined. We will demonstrate different ways in which spatial
 142 contextual signals and visual signals can be combined. We will assess to what extent different
 143 kinds of spatial context, namely spatial relatedness, spatial co-location, and spatial sequence
 144 pattern, can provide such neighborhood information to benefit image classification.

145 **3 Image Classification**

146 In this section, we first describe the image classification task and the data we use. The task is
 147 similar to scene classification but we are specifically interested in classifying different business
 148 venues as opposed to natural environment. Then we explain four different deep convolutional
 149 neural networks that solely leverages the visual signals of images. These convolutional neural
 150 network models are later used as baselines for our experiment.

151 **3.1 Classification Task**

152 Our task is to classify images into one of the several candidate place types. Because we want
 153 to utilize the spatial context in which the image was taken, we need to make sure each image
 154 has a geographic identifier, e.g. geographic coordinates, so that we are able to determine its
 155 neighboring place and their types. In order to classify place types of images, we consider
 156 the scene categories provided by Zhou et al. [35] as they also provide pretrained models
 157 (Places365-CNN) that we can directly use.³ Without losing generality, we select 15 place
 158 types as our candidate class labels. The full list of class labels and their alignment with the
 159 categories in Places365-CNN is shown in Table 1. For each candidate class, we selected 50
 160 images taken in 8 states⁴ within the US by using Google Maps, Google Street View, and
 161 Yelp. These images include both indoor and outdoor views of each place type. Please note
 162 that classifying place types from facade and interior images is a hard problem and even the
 163 most sophisticated models only distinguish a relatively small number of place types so far
 164 which is nowhere near the approximately 420 types provided by sources such as Foursquare.
 165 Places365, for instance, offers 365 classes but many of these are scenes or landscape features,
 166 such as waves, and not POI type, such as cinemas, in the classical sense.

167 **3.2 Convolutional Neural Network Models**

168 To establish baselines for our study, we selected several state-of-the-art image classification
 169 models, namely deep convolutional neural networks. Unlike traditional image classification
 170 pipelines, CNNs extract features from images automatically based on the error messages that
 171 are backpropagated through the network, thus fewer heuristics and less manual labor are
 172 needed. Contrary to densely connected feedforward neural networks, CNN adopts parameter
 173 sharing to extract common patterns which help capture translation invariance and creates
 174 sparse connections which result in fewer parameters and being less prone to overfitting.

³ https://github.com/CSAILVision/places365/blob/master/categories_places365.txt

⁴ Arizona, Illinois, Nevada, North Carolina, Ohio, Pennsylvania, South Carolina, and Wisconsin

■ **Table 1** Class label alignment between Yelp and the Places365 model.

Class label	Places365-CNN category
Amusement Parks	amusement_park
Bakeries	bakery
Bookstores	bookstore
Churches	church
Cinema	movie_theater
Dance Clubs	discotheque
Drugstores	drugstore, pharmacy
Hospitals	hospital, hospital_room
Hotels	hotel, hotel_room
Jewelry	jewelry_shop
Libraries	library
Museums	museum, natural_history_museum, science_museum
Restaurants	fastfood_restaurant, restaurant, restaurant_kitchen, restaurant_patio
Shoe Stores	shoe_shop
Stadiums & Arenas	stadium

175 The architecture of CNNs has been revised numerous times and has become increasingly
 176 sophisticated since its first appearance about 30 years ago. These improvements in architecture
 177 have made CNN more powerful as can be seen in the ImageNet challenge. Some of the
 178 notable architectures include: LeNet [18], AlexNet [17], VGG [24], Inception [27], ResNet
 179 [12], and DenseNet [15]. We selected AlexNet, ResNet with 18 layers (ResNet18), ResNet
 180 with 50 layers (ResNet50), and DenseNet with 161 layers (DenseNet161). AlexNet is among
 181 the first deep neural networks that increased the classification accuracy on ImageNet by
 182 a significant amount compared with traditional classification approaches. By using skip
 183 connections to create residual blocks in the network, ResNet makes it easy to learn identity
 184 functions that help with the vanishing and exploding gradient problems when the network
 185 goes deeper. In DenseNet, a dense connectivity pattern is created by connecting every two
 186 layers so that the error signal can be directly propagated to earlier layers, parameter and
 187 computational efficiency can be increased, and low complexity features can be maintained
 188 [15]. These models were trained on 1.8 million images from the Places365-CNN dataset. We
 189 used the pretrained weights for these models.

190 **4 Spatial Contextual Information**

191 In this section, we introduce three different kinds of spatial contexts and explore ways in
 192 which we can combine them with the CNN models in order to improve image classification.
 193 The first type of spatial context is spatial relatedness, which measures the extend to which
 194 different place types relate with each other. The second type of spatial context is spatial
 195 co-location, which considers what place types tend to co-occur in space and the frequency
 196 they cluster with each other. The third type of spatial context is spatial sequence pattern
 197 which considers both spatial relatedness and spatial co-location. In addition, spatial sequence
 198 pattern considers the interaction between context place types and the inverse relationship
 199 between distance and contextual influence. We use POIs provided by Yelp as dataset. ⁵

⁵ <https://www.yelp.com/dataset>

200 **4.1 Spatial Relatedness**

201 Since the output of CNN is the probability score for each class label, it is possible to interpret
 202 our task as a ranking problem: given an image, rank the candidate class labels based upon
 203 the visual signal and spatial context signal. For the visual signal, we can obtain the ranking
 204 scores (probability scores) from the CNN architectures mentioned in Section 3. Since the
 205 original CNN models has 365 labels, we renormalize the probability scores for each candidate
 206 place type by the sum of the 15 candidate ranking scores so that they sum up to 1. This
 207 renormalization procedure is also applied to the other two spatial context methods explained
 208 in Section 4.2 and Section 4.3. We will refer to the renormalized scores as CNN scores in this
 209 study. For the spatial context signal, the ranking scores are calculated using the place type
 210 embeddings proposed in [30]. These embeddings capture the semantics of different place
 211 types and can be used to measure their similarity and relatedness. In this regard, the task is
 212 equivalent to a re-ranking problem, which adjusts the initial ranking provided by the visual
 213 signal using auxiliary knowledge, namely the spatial context signal. Intuitively, the extent
 214 to which the visual signals from the images match with different place types and the level
 215 of relevance of the surrounding place types with respect to candidate place types jointly
 216 determine the final result.

217 Inspired by search re-ranking algorithms in information retrieval, we use a *Linear Bimodal*
 218 *Fusion* (LBF) method (here essentially a 2-component convex combination), which linearly
 219 combines the ranking scores provided by the CNN model and the spatial relatedness scores,
 220 as shown in Equation 1.

$$221 \quad s_i = \omega^v s_i^v + \omega^r s_i^r \quad (1)$$

222 where s_i , s_i^v , and s_i^r are the LBF score, CNN score, and spatial relatedness score for place
 223 type i respectively, ω^v and ω^r are the weights for the CNN component and spatial relatedness
 224 component, and $\omega^v + \omega^r = 1$. The weights here are decided based on the relative performance
 225 of individual components. Specifically, the weight is determined using Equation 2.

$$226 \quad \omega^v = \frac{acc^v}{acc^v + acc^r} \quad (2)$$

227 where acc^v and acc^r are the accuracies for CNN and spatial relatedness measurements for
 228 the image classification task. Intuitively, this means that we have higher confidence if the
 229 component performs better on its own and want to reflect such confidence using the weight
 230 in the LBF score.

231 In order to calculate the spatial relatedness scores, we use cosine similarity to measure
 232 the extend to which each candidate class embedding is related with the spatial context
 233 embedding of an image in a high dimensional geospatial semantic feature space. Following
 234 the suggestions in [30], we use a concatenated vector of 350 dimensions (i.e., 70D vectors for
 235 each of 5 distance bins) as the place type embeddings. The candidate class embeddings can
 236 be retrieved directly. Then we search for the nearest n POIs based on the image location,
 237 determine the place types of these n POIs, and calculate the average of these place type
 238 embeddings as the final spatial context embeddings for images. The cosine similarity score
 239 sm_i is calculated between the spatial context embedding of an image and the embedding
 240 of each candidate place type class i . Because sm_i ranges from -1 to 1, we use min-max
 241 normalization to scale the values to $[0, 1]$. Finally, we apply the same renormalization as for
 242 the CNN score to turn the normalized score sm_i' into probability score, i.e. spatial relatedness
 243 score s_i^r .

244 Combining these normalizations together with Equation 1 and Equation 2, we are able to
 245 derive that $0 \leq s_i \leq 1$ and $\sum_{i=1}^N s_i = 1$ where $N = 15$ in our case. This means that the LBF
 246 score s_i can be considered a probability score.

247 4.2 Spatial Co-location

248 The spatial relatedness approach follows the assumption that relatedness implies likelihood
 249 which is reasonable in cases where similar place types cluster together, such as restaurant,
 250 bar, and hotel. However, in cases of high spatial heterogeneity, this assumption will fall short
 251 of correctly capturing the true likelihood. An example would be places of dissimilar types
 252 that co-occur, e.g., grocery stores and gas stations. Moreover, the LBF method can only
 253 capture a linear relationship between the two signals.

254 Following Berg et al.[2], we also test a Bayesian approach in which we assume there is a
 255 complex latent distribution of the data that facilitates our classification task. Intuitively,
 256 the CNN score gives us the probability of each candidate class t given the image I , i.e.,
 257 $P(t|I)$, and the spatial context informs us of the probability of each candidate class given its
 258 neighbors $c_1, c_2, c_3, \dots, c_n$, denoted as C , around the image location, i.e., $P(t|C)$. We would
 259 like to obtain the posterior probability of each candidate class given both the image and
 260 its spatial context, i.e., $P(t|I, C)$. Using Bayes' theorem, the posterior probability can be
 261 written as:

$$262 \quad P(t|I, C) = \frac{P(I, C|t)P(t)}{P(I, C)} \quad (3)$$

263 For variables I , C , and t , we construct their dependencies using a simple probabilistic
 264 graphical model, i.e., Bayesian network, which assumes that both the image I and the spatial
 265 context C are dependent on the place type t , which intuitively makes sense in that different
 266 place types will result in different images and different place types of their neighbors. We
 267 know that given information about the image I we are able to update our beliefs, i.e., the
 268 probability distributions, about the place type t . In addition, the changes in our beliefs about
 269 the place type t can influence the probability distributions of the spatial context C . However,
 270 if place type t is observed, the influence cannot flow between I and C , thus we are able to
 271 derive the conditional independence of I and C given t . So Equation 3 can be rewritten as:

$$272 \quad \begin{aligned} P(t|I, C) &= \frac{P(I|t)P(C|t)P(t)}{P(I, C)} \\ &= \frac{P(t|I)P(I)}{P(t)} \frac{P(t|C)P(C)}{P(t)} \frac{P(t)}{P(I, C)} \\ &\propto \frac{P(t|I)}{P(t)} P(t|C) \end{aligned} \quad (4)$$

273 in which we have dropped all the factors that are not dependent on t as they can be considered
 274 as normalizing constants for our probabilities. It follows that the posterior probability
 275 $P(t|I, C)$ can be computed using the CNN probability score $P(t|I)$, the spatial context prior
 276 $P(t|C)$, and the candidate class prior $P(t)$. Instead of estimating the distribution of spatial
 277 context priors, we take advantage of the spatial co-location patterns and calculate the prior
 278 probabilities using the Yelp POI data directly. As mentioned earlier, the spatial context
 279 C is composed of multiple individual context neighbors $c_1, c_2, c_3, \dots, c_n$; hence, we need to
 280 calculate $P(t|c_1, c_2, c_3, \dots, c_n)$. In order to simplify our calculation, we impose a bag-of-words
 281 assumption as well as a Naive Bayes assumption in the spatial co-location patterns. The
 282 bag-of-words assumption simplifies the model by assuming that the position (or the order) in

283 which different context POIs occur does not play a role. The Naive Bayes assumption implies
 284 that the only relationship is the pair-wise interaction between the candidate place type t
 285 and an individual neighbor's place type c_i and there is no interaction between neighboring
 286 places wrt. their types, i.e. $(c_i \perp\!\!\!\perp c_j | t)$ for all c_i, c_j . Using spatial co-location, we are able to
 287 calculate the conditional probability using place type co-location counts $P(c_i | t) = \frac{\text{count}(c_i, t)}{\text{count}(t)}$
 288 where $\text{count}(c_i, t)$ is the frequency that neighbor type c_i and candidate type t co-locate
 289 within a certain distance limit and $\text{count}(t)$ is the frequency of candidate type t in the study
 290 area. Combining all these components, we can derive:

$$\begin{aligned}
 P(t|C) &= P(t|c_1, c_2, \dots, c_n) \\
 &= \frac{P(t) \prod_{i=1}^n P(c_i|t)}{P(c_1, c_2, c_3, \dots, c_n)} \\
 &= \frac{P(t)}{P(c_1, c_2, c_3, \dots, c_n)} \frac{\prod_{i=1}^n \text{count}(c_i, t)}{\text{count}(t)^n}
 \end{aligned} \tag{5}$$

292 Using Equation 4 and Equation 5, we can derive the final formula for calculating $P(t|I, C)$
 293 shown in Equation 6. For the sake of numerical stability, we calculate the log probability
 294 $\log P(t|I, C)$ using the natural logarithm. Since the natural logarithm is a monotonically
 295 increasing function, it will not affect the final ranking of the classification results.

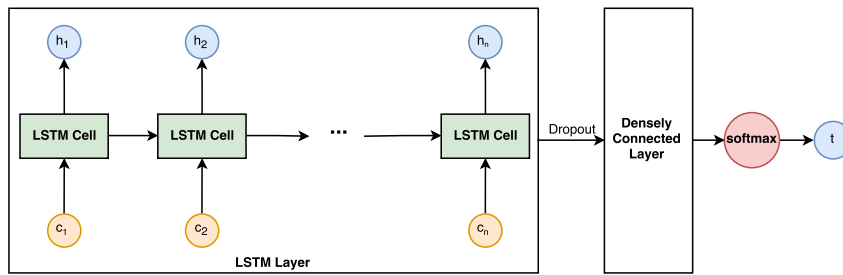
$$\begin{aligned}
 \log P(t|I, C) &\propto \log \left(\frac{P(t|I)}{P(t)} P(t|C) \right) \\
 &= \log \left(\frac{P(t|I)}{P(c_1, c_2, c_3, \dots, c_n)} \frac{\prod_{i=1}^n \text{count}(c_i, t)}{\text{count}(t)^n} \right) \\
 &\propto \log P(t|I) + \sum_{i=1}^n \log(\text{count}(c_i, t)) - n \log(\text{count}(t))
 \end{aligned} \tag{6}$$

297 where we also drop $P(c_1, c_2, c_3, \dots, c_n)$ as it does not depend on t , so it will not affect the
 298 result ranking. The log posterior probability is then used to generate the final ranking of
 299 candidate place types and produce the classification results.

300 4.3 Spatial Sequence Pattern

301 The spatial co-location approach follows the bag-of-words assumption that the position of
 302 spatial context POIs does not matter and the Naive Bayes assumption that the context
 303 neighbors are independent of each other. However, in many cases this assumption is too
 304 strong. In fact, numerous methods, such as Kriging and multiple-point geostatistics, have
 305 been devised to model geospatial proximity patterns and complex spatial interaction patterns.
 306 However, incorporating these complex spatial patterns in a multidimensional space would
 307 adversely affect the model complexity and make the distribution in Section 4.2 intractable.
 308 In order to strike the right balance between the complexity of model and the integrity of
 309 spatial context pattern, we propose to capture the spatial sequence pattern in our model by
 310 collapsing the 2D geographic space into a 1D sequence.

311 Specifically, we use the Long Short-Term Memory (LSTM) network model, a variant of
 312 recurrent neural network (RNN), in our study. Recurrent neural networks are frequently
 313 used models to capture the patterns in sequence or time series data. In theory, the naive
 314 recurrent neural networks can capture long term dependencies in the sequence, however,
 315 due to the vanishing and exploding gradient problem, they fail to do so in practice. LSTM
 316 is explicitly designed to solve the problem by maintaining a cell state and controlling the



■ **Figure 1** Structure of the LSTM.

317 input and output flow using forget gate, input gate, and output gate [14]. We use LSTM
 318 as a generative model in order to capture the latent distribution of place types using the
 319 spatial sequence pattern. In the training stage, the input is a sequence of context place
 320 types $c_1, c_2, c_3, \dots, c_n$ and the output is the place type t of the POI from which the context is
 321 created. The input sequence is ordered in a way so that the previous one is further away
 322 from the output than the next one in the collapsed 1D space. Image one would drive around
 323 a neighborhood before reaching a destination. For each of the POIs encountered during the
 324 route, one would update the beliefs about the neighborhood by considering the current POI
 325 and all previously seen POIs. Upon arriving at the destination, one would have a reasonable
 326 chance of guessing this final POI's type. The structure of the LSTM model is shown in
 327 Figure 1. We apply a dropout after the LSTM layer to avoid overfitting. After training
 328 the LSTM model on Yelp's POI dataset, we are able to obtain the spatial context prior
 329 $P(t|c_1, c_2, c_3, \dots, c_n)$ based on the spatial sequence pattern around the image locations in our
 330 test data. We specifically removed the image locations and their context in the training data.
 331 Similar to the spatial co-location approach, we use Bayesian inference and log probability to
 332 calculate the final result:

$$\begin{aligned}
 \log P(t|I, C) &\propto \log \left(\frac{P(t|I)}{P(t)} P(t|C) \right) \\
 &= \log P(t|I) + \log P(t|c_1, c_2, c_3, \dots, c_n) - \log P(t)
 \end{aligned}
 \tag{7}$$

334 where the candidate class prior $P(t)$ can be computed using the Yelp data. Since we use LSTM
 335 as a generative model, in the prediction phase, sampling strategies, such as greedy search,
 336 beam search, and random sampling, can be applied based on the distribution provided
 337 by the output of the LSTM prediction. However, we only generate the next prediction
 338 instead of a sequence, so we do not apply these sampling strategies. Instead, we make use of
 339 the hyperparameter *temperature* τ to adjust the probability scores returned by the LSTM
 340 model before combining them with the CNN model in a Bayesian manner. Including the
 341 hyperparameter τ , the softmax function in the LSTM model can be written as:

$$P(t_i|C) = \frac{\exp(\frac{\text{logit}_i}{\tau})}{\sum_{j=1}^N \exp(\frac{\text{logit}_j}{\tau})}
 \tag{8}$$

343 where logit_i is the logit output provided by LSTM before applying the softmax function and
 344 $N = 15$ in our case. Intuitively, when the temperature τ is high, i.e., $\tau \rightarrow \infty$, the probability
 345 distribution will become diffuse and $P(t_i|C)$ will have almost the same value for different t_i ;
 346 when τ is low, i.e., $\tau \rightarrow 0^+$, the distribution becomes peaky and the largest logit_i stands
 347 out to have a probability close to 1. This idea is closely related to the exploration and
 348 exploitation trade-off in many machine learning problems. The value of τ will affect the
 349 probability scores $P(t_i|C)$ but not the ranking of these probabilities.

350 In this study, we propose two ways to model the 2D geographic space as a 1D sequence.
 351 The first one is a distance-based ordering approach. For any given POI, we search for nearby
 352 POIs within a certain distance from it, choose the closest n POIs, and rearrange them by
 353 distance with descending order, thereby forming a 1D array. This distance-based method is
 354 isotropic in that it does not differentiate between directions while creating the sequence. The
 355 second method is a space filling curve-based approach. We utilize *Morton order* here which
 356 is also used in geohashing to encode coordinates into an indexing string that can preserve
 357 the locality of spatial locations. We use Morton order to encode the geographic locations of
 358 every POI and order them in a sequence based upon their encodings, i.e., indexing sequence.
 359 After obtaining the sequence, for each POI, we use the previous n POI in the sequence as
 360 the context sequence. Other space filling curves could be used in future work.

361 Because each POI can have multiple place types associated with it, e.g., restaurant and
 362 beer garden, the sequence of place types is usually not unique for the same sequence of POIs.
 363 As our LSTM input is a sequence of place *types*, we compute the Cartesian product of all
 364 POI type sets in the sequence of nearby places:

$$365 \quad T_{c_1} \times T_{c_2} \times T_{c_3} \times \dots \times T_{c_n} = \{(t_{c_1}, t_{c_2}, t_{c_3}, \dots, t_{c_n}) | \forall i = 1, 2, 3, \dots, n, t_{c_i} \in T_{c_i}\} \quad (9)$$

366 where T_{c_i} is the set of place types associated with POI c_i in the context sequence. In
 367 practice, however, we randomly sample a fixed number of place type sequences from each
 368 of the Cartesian product for the POI context sequence as the potential combinations grow
 369 exponentially with increasing context size.

370 **5 Experiment and Result**

371 In this section, we explain our experimental setup for the models described above, describe
 372 the metrics used to compare the model performance for place type image classification, and
 373 present the results and findings.

374 **5.1 Implementation Details**

375 For all three types of spatial context, we use 10 as the maximum number of context POIs
 376 and a distance limit of 1000m for the context POI search. For the spatial sequence pattern
 377 approach, we use a fixed sample size of 50 to sample from the Cartesian product of all POI
 378 type sets in the sequence.⁶ We use a one-layer LSTM with 64 hidden units. We train our
 379 LSTM model using the recommended Root Mean Square Propagation (RMSProp) optimizer
 380 with a learning rate of 0.005. A dropout ratio of 0.2 is applied in the LSTM and we run
 381 100 epochs. The same settings are used for all LSTM trainings in our experiment. The
 382 total number of POI in the dataset is 115,532, yielding more than 5 million unique training
 383 sequences.

384 For evaluation, we use three different metrics, namely Mean Reciprocal Rank (MRR),
 385 Accuracy@1, and Accuracy@5. Another common metric for image classification would also
 386 be Mean Average Precision (MAP), but since there is only one true label per type in our
 387 task, we use MMR instead.

⁶ The median for types per place in Yelp is 3.

5.2 Results

We run the 750 test images we collected, i.e., 50 images per each of 15 types, on the four CNN baseline models (AlexNet, ResNet18, ResNet50, and DenseNet161) as well as the combined models using our three different types of spatial context.⁷ In addition to the two methods for converting geographic space into 1D sequences in the spatial sequence pattern approach, we also test one model using random sequences with the same context count and distance limits. We did so to study whether results obtained using the LSTM would benefit from distance-based spatial contexts. A higher result for the spatial sequence based LSTM over the random LSTM would indicate that the network indeed picked up on the distance signal.

The hyperparameter τ can be adjusted; a value of 0.5 has been proposed as a good choice before. In order to test this and find the optimal temperature value, we run the combined model using spatial sequence patterns with three types of sequencing approaches, namely random sequence, distance-based sequence, and Morton order-based sequence.

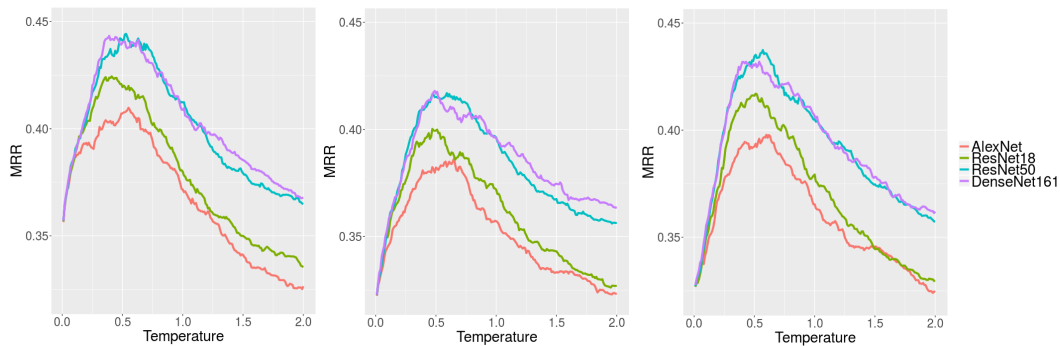


Figure 2 From left to right, MRR result using distance-based sequence, random sequence, and Morton code-based sequence with varying temperatures

We test temperature values ranging from 0.01 to 2 with a step of 0.01. We combine the spatial sequence pattern models with all CNN models. The MRR result with respect to temperature are shown in Figure 2. Although there are slight variations, the MRR curves all reach their peaks around a τ value of 0.5. This confirms the suggestion from the literature. Figure 3 shows selected example predictions. The results for MRR, Accuracy@1, and Accuracy@5 using the baseline models as well as our proposed, spatially explicit models are shown in Table 2, Table 3, and Table 4.⁸

As we can see, by incorporating spatial context in the image classification model, we are able to improve the classification result in general. However, integrating spatial relatedness using the LBF method does not seem to affect the result. This essentially confirms our aforementioned assumption that relatedness does not always imply likelihood. The benefit of incorporating spatial relatedness in cases of spatial homogeneity are likely to be offset by cases of high spatial heterogeneity in which spatial relatedness may have a negative effect as dissimilar places co-occur.

⁷ Transfer learning could be applied to fine tune the CNN models first, but we only have limited images and our hypothesis is that spatial context can be used as a powerful complement or alternative to the visual component for image classification.

⁸ The baseline models are not comparable with a random classifier which would yield an expected accuracy of 1/15 in this case, because the baseline CNN models have 365 unique labels and we choose 15 labels in our experiment.



■ **Figure 3** From left to right, images of a restaurant, a hotel, and a museum from Yelp, Google Street View, and Google Maps respectively. The first image is incorrectly classified as library using all 4 CNN models and it is correctly classified as restaurant using the spatial sequence pattern (distance) models. The second image is classified as hospital and library by the original CNN models and is classified as hotel by the spatial sequence pattern (distance) models. For the third image the correct label museum is in the third position in the label rankings of all 4 CNN models while, using the spatial sequence pattern (distance) models, ResNet18 and ResNet50 can correctly label it and in the label rankings of AlexNet and DenseNet161 museum is in the second position.

■ **Table 2** MRR result using baseline models and proposed combination models using different types of spatial context and sequences

MRR	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.27	0.28	0.31	0.31
Relatedness	0.27	0.28	0.31	0.32
Co-location	0.30	0.31	0.31	0.32
Sequence Pattern (Random)	0.38	0.40	0.42	0.42
Sequence Pattern (Distance)	0.41	0.42	0.44	0.44
Sequence Pattern (Morton order)	0.39	0.42	0.43	0.43

415 The Accuracy@1 measurement is improved by incorporating spatial co-location component
 416 in the models. This confirms our previous reasoning that considering the external signal,
 417 namely spatial contexts, and assuming a complex latent distribution of the data in a Bayesian
 418 manner improve image classification. However, for MRR the improvement is marginal and
 419 for Accuracy@5 there even is a decrease after incorporating the spatial co-location component
 420 because this type of spatial context falls short of taking into account the intricate *interactions*
 421 of different context neighbors. This shortcoming is not clear when only looking at the first
 422 few results in the ranking returned by the combined models, but it becomes clearer in later
 423 results in the ranking output, thus resulting in a decrease for Accuracy@5 and only a slight
 424 increase in the MRR measurement.

■ **Table 3** Accuracy@1 result using baseline models and proposed combination models using different types of spatial context and sequences

Accuracy@1	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.07	0.07	0.09	0.09
Relatedness	0.07	0.07	0.09	0.09
Co-location	0.15	0.17	0.17	0.17
Sequence Pattern (Random)	0.18	0.18	0.19	0.20
Sequence Pattern (Distance)	0.20	0.20	0.22	0.22
Sequence Pattern (Morton order)	0.19	0.20	0.22	0.22

425 The Bayesian combination model using spatial sequence patterns shows better overall
 426 results compared with the baseline models, the spatial relatedness model, and the spatial

■ **Table 4** Accuracy@5 result using baseline models and proposed combination models using different types of spatial context and sequences

Accuracy@5	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.50	0.56	0.59	0.60
Relatedness	0.52	0.56	0.58	0.59
Co-location	0.42	0.44	0.45	0.44
Sequence Pattern (Random)	0.65	0.69	0.73	0.73
Sequence Pattern (Distance)	0.67	0.70	0.73	0.75
Sequence Pattern (Morton order)	0.65	0.70	0.72	0.71

427 co-location model. This is because the spatial sequence patterns capture spatial interactions
 428 between the neighboring POIs that are neglected by the other models. From the result we
 429 can see that using a distance-based sequence is better than using a random sequence. To
 430 prevent confusion and to understand why the random model still performs relatively well, it
 431 is important to remember that this model utilizes spatial context. However, it does not utilize
 432 the distance signal within this context but merely the presence of neighboring POI. The
 433 results show that a richer spatially explicit context, one that comes with a notion of *distance*
 434 *decay*, indeed improves classification results. Interestingly, the sequence using Morton order,
 435 which is widely used in geohashing techniques, does not further improve the result compared
 436 to the distance-based sequence. There may be multiple reasons for this. First, we may have
 437 reached a ceiling of possible improvements by incorporating spatial contexts. Second, our
 438 Morton order implementation takes the 10 places that precede the target place in the index.
 439 This may result in directional effects. Finally, all space filling curves essentially introduce
 440 different ways to preserve local neighborhoods; utilizing another technique such as Hilbert
 441 curves may yield different results. Given that the Morton order-based sequence in many
 442 cases yield results of equal quality to the distance-based sequences, further work is needed to
 443 test the aforementioned ideas.

444 Summing up, the results demonstrate that incorporating a (distance-based) spatial context
 445 improves the MRR of state-of-the-art image classification systems by over **40%**. The results
 446 for Accuracy@1 are more than **doubled** which is of particular importance for humans as
 447 this measure only considers the first ranked result.

448 **6 Conclusion and Future Work**

449 In this work, we demonstrated that utilizing spatial contexts for classifying places based on
 450 images of their facades and interiors leads to substantial improvements, e.g., increasing MRR
 451 by over 40% and doubling Accuracy@1, compared to applying state-of-the-art computer
 452 vision models such as ResNet50 and DenseNet161 alone. These advances are especially
 453 significant as the classification of places based on their images remains a hard problem. One
 454 could argue that our proposal requires additional information, namely about the types of
 455 nearby places. However, such data are readily available for POI, and only a few nearby places
 456 are needed. Secondly, and as a task for future work, one could also modify our methods
 457 to work in a *drive-by-typing* mode in which previously seen places are classified, and these
 458 classification results together with their associated classification uncertainty are used to
 459 improve estimation of the currently seen place, thereby relaxing the need for POI datasets. In
 460 the future, we would like to apply transfer learning and experiment with other ways to encode
 461 spatial contexts, e.g., by testing different space-filling curves. We plan to develop models to
 462 directly capture 2D spatial patterns rather than using a 1D sequence as a proxy and test

463 whether spatial contexts also aid in recognizing objects beyond places and their facades.

464 — **References** —

- 465 **1** Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geo-
466 localization of images in mountainous terrain. In *Computer Vision–ECCV 2012*, pages
467 517–530. Springer, 2012.
- 468 **2** Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and
469 Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In
470 *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2019–
471 2026. IEEE, 2014.
- 472 **3** Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai.
473 Man is to computer programmer as woman is to homemaker? debiasing word embeddings.
474 In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.
- 475 **4** Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically
476 from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- 477 **5** Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Land use clas-
478 sification in remote sensing images by convolutional neural networks. *arXiv preprint*
479 *arXiv:1508.00092*, 2015.
- 480 **6** Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep
481 structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.
- 482 **7** Anne Cocos and Chris Callison-Burch. The language of place: Semantic value from geo-
483 spatial context. In *15th Conference of the European Chapter of the Association for Com-
484 putational Linguistics*, volume 2, pages 99–104, 2017.
- 485 **8** Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent
486 representation for predicting future visitors. In *AAAI*, pages 102–108, 2017.
- 487 **9** Michael F Goodchild and Donald G Janelle. Thinking spatially in the social sciences.
488 *Spatially integrated social science*, pages 3–22, 2004.
- 489 **10** Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-
490 digit number recognition from street view imagery using deep convolutional neural networks.
491 *arXiv preprint arXiv:1312.6082*, 2013.
- 492 **11** Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing
493 adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- 494 **12** Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for
495 image recognition. In *Proceedings of the IEEE conference on computer vision and pattern
496 recognition*, pages 770–778, 2016.
- 497 **13** Jeremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In
498 *European conference on computer vision*, pages 30–43. Springer, 2008.
- 499 **14** Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*,
500 9(8):1735–1780, 1997.
- 501 **15** Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely
502 connected convolutional networks. In *Proceedings of the IEEE conference on computer
503 vision and pattern recognition*, volume 1, page 3, 2017.
- 504 **16** Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn
505 to criticize! criticism for interpretability. In *Advances in Neural Information Processing
506 Systems*, pages 2280–2288, 2016.
- 507 **17** Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep
508 convolutional neural networks. In *Advances in neural information processing systems*, pages
509 1097–1105, 2012.
- 510 **18** Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning
511 applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

- 512 **19** Stefan Lee, Haipeng Zhang, and David J Crandall. Predicting geo-informative attributes
513 in large-scale image collections using convolutional neural networks. In *Applications of*
514 *Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 550–557. IEEE, 2015.
- 515 **20** Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In
516 *Computer Vision and Pattern Recognition*, pages 891–898. IEEE, 2013.
- 517 **21** Kang Liu, Song Gao, Peiyuan Qiu, Xiliang Liu, Bo Yan, and Feng Lu. Road2vec: Measuring
518 traffic interactions in urban road system from massive travel routes. *ISPRS International*
519 *Journal of Geo-Information*, 6(11):321, 2017.
- 520 **22** Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A
521 Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the*
522 *National Academy of Sciences*, 114(29):7571–7576, 2017.
- 523 **23** Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distil-
524 lation as a defense to adversarial perturbations against deep neural networks. In *Security*
525 *and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.
- 526 **24** Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale
527 image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- 528 **25** Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Explanations,
529 bias detection, adversarial examples and model criticism. *arXiv:1711.11443*, 2017.
- 530 **26** Wanxiao Sun, Volker Heidt, Peng Gong, and Gang Xu. Information fusion for rural land-
531 use classification with high-resolution satellite imagery. *IEEE Transactions on Geoscience*
532 *and Remote Sensing*, 41(4):883–890, 2003.
- 533 **27** Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir
534 Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper
535 with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- 536 **28** Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian
537 Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint*
538 *arXiv:1312.6199*, 2013.
- 539 **29** Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving im-
540 age classification with location context. In *Proceedings of the IEEE international conference*
541 *on computer vision*, pages 1008–1016, 2015.
- 542 **30** Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec-
543 reasoning about place type similarity and relatedness by learning embeddings from aug-
544 mented spatial contexts. *Proceedings of SIGSPATIAL*, 17:7–10, 2017.
- 545 **31** Jie Yu and Jiebo Luo. Leveraging probabilistic season and location context models for
546 scene understanding. In *Proceedings of the 2008 international conference on Content-based*
547 *image and video retrieval*, pages 169–178. ACM, 2008.
- 548 **32** Andi Zang, Runsheng Xu, Zichen Li, and David Doria. Lane boundary extraction from
549 satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on High-Precision*
550 *Maps and Intelligent Applications for Autonomous Vehicles*, page 1. ACM, 2017.
- 551 **33** Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also
552 like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv*
553 *preprint arXiv:1707.09457*, 2017.
- 554 **34** Shenglin Zhao, Tong Zhao, Irwin King, and Michael R Lyu. Geo-teaser: Geo-temporal
555 sequential embedding rank for point-of-interest recommendation. In *Proceedings of the*
556 *26th international conference on world wide web companion*, pages 153–162. International
557 World Wide Web Conferences Steering Committee, 2017.
- 558 **35** Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places:
559 A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis*
560 *and Machine Intelligence*, 2017.