Placepedia: Comprehensive Place Understanding with Multi-Faceted Annotations

Anonymous ECCV submission

Paper ID 3644

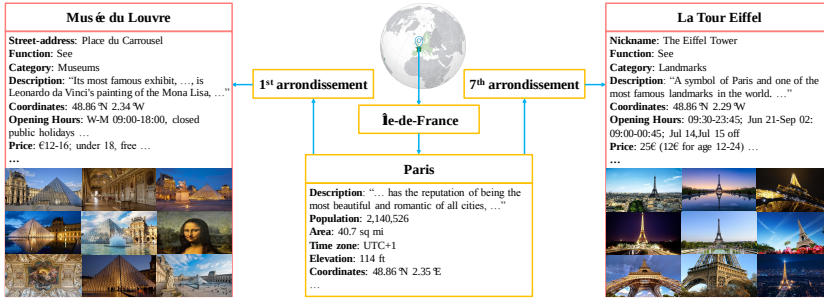

Fig. 1. Hierarchical structure of Placepedia with places from all over the world. Each place is associated with its *district*, *city/town/village*, *state/province*, *country*, *continent*, and a large amount of diverse photos. Both administrative areas and places have rich side information, *e.g. discription*, *population*, *category*, *function*, which allows various large-scale studies to be conducted on top of it

Abstract. Place is an important element in visual understanding. Given a photo of a building, people can often tell its functionality, *e.g.* a restaurant or a shop, its cultural style, *e.g.* Asian or European, as well as its economic type, *e.g.* industry-oriented or tourism oriented. While place recognition has been widely studied in previous work, there remains a long way towards comprehensive place understanding, which is far beyond categorizing a place with an image and requires information of multiple aspects. In this work, we contribute Placepedia, a large-scale place dataset with more than $35M$ photos from $240K$ unique places. Besides the photos, each place also comes with massive multi-faceted information, *e.g.* GDP, population, *etc.*., and labels at multiple levels, including function, city, country, *etc.*. This dataset, with its large amount of data and rich annotations, allows various studies to be conducted. Particularly, in our studies, 1) we develop PlaceNet, a unified framework for multi-level place recognition, and 2) a method for city embedding, which can produce a vector representation for a city that captures both visual and multi-faceted side information. Such studies not only reveal the key challenges in place understanding, but also allow us to establish the connections between visual observations and the underlying socioeconomic or cultural implications.

1 Introduction

Imagine that you are visiting a new country, and you are traveling among different cities. In each city, you will encounter countless places, and you may see a fancy building, experience some natural wild beauty, or enjoy the unique culture, *etc.*. All of these experiences impress you and lead you to a deeper understanding of the place. As we browse through a city, based on certain common visual elements therein, we implicitly establish connections between visual characteristics with other multi-faceted information, such as its function, socioeconomic status and culture. We therefore believe that it would be a rewarding adventure to move beyond conventional place categorization and explore the connections among different aspects of a place. It indicates that multi-dimension labels are essential for comprehensive place understanding. To support this exploration, a large-scale dataset that cover a diverse set of places with both images and comprehensive multi-faceted information is needed.

However, existing datasets for place understanding [31, 53, 34], as shown in Tab. 1 are subject to at least one of the following drawbacks: 1) *Limited Scale.* Some of them [33, 34] contain only several thousand images from one particular city. 2) *Restrictive Scope.* Most datasets are constructed for only one task, *e.g.* place retrieval [31] or scene recognition [53, 18]. 3) *Lack of Attributes.* These datasets often contain just a very limited set of attributes. For example, [31] contains just photographers and titles. Clearly, these datasets, due to their limitations in scale, diversity, and richness, are not able to support the development of comprehensive place understanding.

In this work, we develop *Placepedia*, a comprehensive place dataset that contains images for places of interest from all over the world with massive attributes, as shown in Fig. 1. *Placepedia* is distinguished in several aspects: **1)** *Large Scale.* It contains over $35M$ images from $240K$ places, several times larger than previous ones. **2)** *Comprehensive Annotations.* The places in Placepedia are tagged with categories, functions, administrative divisions at different levels, *e.g.* city and country, as well as lots of multi-faceted side information, *e.g.* descriptions and coordinates. **3)** *Public Availability.* Placepedia will be made public to the research community. We believe that it will greatly benefit the research on comprehensive place understanding and beyond.

Meanwhile, Placepedia also enables us to rigorously benchmark the performance of existing and future algorithms for place recognition. We create four benchmarks based on Placepedia in this paper, namely *place retrieval*, *place categorization*, *function categorization*, and *city/country recognition*. By comparing different methods and modeling choices on these benchmarks, we gain insights into their pros and cons, which we believe would inspire more effective techniques for place recognition. Furthermore, to provide a trigger for comprehensive place understanding, we develop *PlaceNet*, a unified deep network for multi-level place recognition. It simultaneously predicts place item, category, function, city, and country. Experiments show that by leveraging the multi-level annotations in Placepedia, PlaceNet can learn better representation of a place than previous works. We also leverage both visual and side information from Placepedia to

Table 1. Comparing Placepedia with other existing datasets. Placepedia offers the largest number of images and the richest information

	# places	# images	# categories	Meta data
Google Landmarks [31]	203,094	5,012,248	N/A	authors, titles
Places365 [53]	N/A	10,624,928	434	N/A
Holidays [18]	N/A	1,491	500	N/A
Oxford 5k [33]	11	5,062	N/A	N/A
Paris 6k [34]	11	6,412	N/A	N/A
SFLD [5]	N/A	1,700,000	N/A	coordinates
Pitts 250k [48]	N/A	254,064	N/A	coordinates
Google Street View [11]	10,343	62,058	N/A	coordinates, addresses, *etc.*
Tokyo 24/7 [47]	125	74,000	N/A	coordinates
Cambridge Landmarks [8]	6	>12,000	N/A	N/A
Vietnam Landscape[a]	103	118,000	N/A	N/A
Placepedia	**>240,000**	**>35,000,000**	**>3,000**	**divisions, descriptions, city info, *etc.***

[a] https://blog.facebit.net/2018/09/07/zalo-ai-challenge-problems-and-solutions

learn city embeddings, which demonstrate strong expressive power as well as the insights on *what distinguish a city*.

From the empirical studies on Placepedia, we see lots of challenges in performing place recognition. 1) The visual appearance can vary significantly due to the changes of angle, illumination, and other environmental factors. 2) A place may look completely different when viewed from inside and outside. 3) A big place, *e.g.* a university, usually consists of a number of small places that have nothing in common in appearance. All these problems remain open. We hope that Placepedia, with its large scale, high diversity and massive annotations, would provide a gold mine for the community to develop more expressive models to meet the aforementioned challenges.

Our contributions in this work can be summarized as below. **1)** We build Placepedia, a large-scale place dataset with comprehensive annotations in multiple aspects. To the best of our knowledge, Placepedia is the largest and the most comprehensive dataset for place understanding. **2)** We design four task-specific benchmarks on Placepedia *w.r.t.* the multi-faceted information. **3)** We conduct systematic studies on place recognition and city embedding, which demonstrate important challenges in place understanding as well as the connections between the visual characteristics and the underlying socioeconomic or cultural implications.

2 Related Work

Place Understanding Datasets. During the past decade, lots of place datasets were constructed to facilitate place-related studies. There are mainly three kinds of datasets. The first kind [31, 47, 5, 48, 11] focuses on the tasks of place recognition/retrieval, where images are labeled as particular place items, e.g. White House or Eiffel Tower. The second kind [53] targets place categorization or scene recognition. In these datasets, each image is attached with a place type, e.g. parks or museums. The third kind [33, 34, 18] is for object/image retrieval. The statistics is summarized in Tab. 1. Compared with these datasets, our Placepedia has much larger amount of image and context data, containing over $240K$ places

with 35 million images labeled with $3K$ categories. Hence, Placepedia can be used for all these tasks. Besides, the provision of hierarchical administrative divisions for places allows us to study place recognition in different scales, *e.g.* city or country recognition. Also, the function information (*See*, *Do*, *Sleep*, *Eat*, *Buy*, etc.) of places may lead to a new task, namely place function recognition.

Place Understanding Tasks. Lots of work aims at 2D place recognition [5, 48, 11, 47, 2, 4, 12, 21, 35, 37, 1, 43, 25, 29, 28, 41, 32, 55, 19, 16] or place retrieval [31, 46, 19, 36, 50, 10, 9, 44]. Given an image, the goal is to recognize what the particular place is or to retrieve images representing the same place. Scene recognition [53, 54, 52, 23, 51, 49], on the other hand, defines a diverse list of environment types as labels, *e.g.* nature, classroom, bar. And their job is to assign each image a scene type. [7, 39] collects images from several different cities, studies on distinguishing images of one city from others, and discovers what elements are characteristics of a given city. There also exist some other humanities-related studies. [42] classifies keywords of city description into *Economic*, *Cultural*, or *Political*, and then counts the occurrences of these three types to represent city branch. [26] uses satellite data from both daytime and nighttime to discover "ghost cities" which consist of abandoned buildings or housing structures which may hurt urbanization process. [20] collects place images from social media to extract human emotions at different places, to find out the relationship between human emotions and environment factors. [30] uses neural networks trained with map imagery to understand the connection among cities, transportation, and human health. Placepedia, with its large amount of data in both visual and textual domains, allows various studies to be conducted on top in large scale.

3 The Placepedia Dataset

We contribute Placepedia, a large-scale place dataset, to the community. Some example images along with labels are shown in Fig. 2. In this section, we introduce the procedure of building Placepedia. First, a hierarchical structure is organized to store places and their multi-level administrative divisions, where each place/division is associated with rich side information. With this structure, global places are connected and classified on different levels, which allows us to investigate numerous place-related issues, *e.g.* city recognition [7, 39], country recognition, and city embedding. With types (*e.g. park*, *airport*) provided we can explore tasks such as place categorization [53, 23]. With functions (*e.g. See*, *Sleep*) provided we are able to model the functionality of places. Second, we download place images from Google Image open source, which are cleaned automatically and manually.

3.1 Hierarchical Administrative Areas and Places

Place Collection. We collect place items with side information from Wikivoyage[1], a free worldwide travel guide website, through the public channel. Pages of

[1] https://en.wikivoyage.org/wiki/Destinations

Fig. 2. The text in red, orange, green, blue, purple represents place names, cities, countries/territories, categories, functions, respectively. We see that the appearance of a particular place can vary from: 1) daytime to nighttime, 2) different angles, and 3) inside and outside

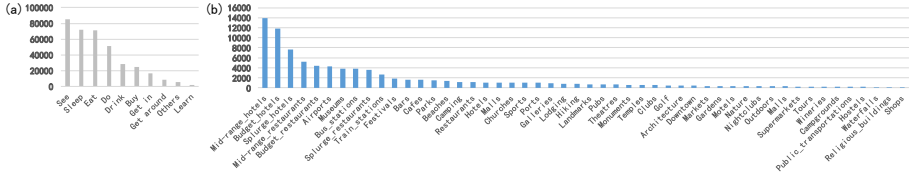

Fig. 3. (a) The number of places for top-10 functions. (b) The number of places for top-50 categories

Wikivoyage are organized in a hierarchical way, *i.e.*, each destination is obtained by walking through its *continent*, *country*, *state/province*, *city/town/village*, *district*, *etc.*. As illustrated in Fig. 1, these administrative areas serve as non-leaf nodes in Placepedia, and leaf nodes represent all the places. This process results in a list of $361,524$ places together with $24,333$ administrative areas.

Meta Data Collection. In Wikivoyage, all destinations are associated with some of the attributes below: *function*, *category*, *description*, *GPS coordinates*, *address*, *phone number*, *opening hour*, *price*, *homepage*, *wikipedia link*. The place number of top-10 functions and top-50 categories are shown in Fig. 3. Tab. 2 shows the definition for ten functions in Placepedia and Fig. 4 shows several examples of these ten *functions*. Function labels are the section names of places from Wikivoyage. Place functions serve as a good indicator for travelers to choose where to go. For example, some people love to go shopping when traveling; Some prefer to enjoy various flavors of food; Some people are addicted to distinctive landscapes. For administrative areas, Wikivoyage often lack meta data. Hence, we acquire the missing information by parsing their Wikipedia page. At last, the following attributes are extracted: *description*, *GDP*, *population density*,

Fig. 4. This figure shows ten *function* labels with five example places for each

population, *elevation*, *time zone*, *area*, *land area*, *water area*, *GPS coordinates*, *establish time*, *etc.*.

Place Cleaning. To refine the place list, we only keep places satisfying at least one of the two following criteria: 1) It has the attribute *GPS coordinates* or *address*; 2) It is identified as a *location* by Google Entity Recognition [2] or Stanford Entity Recognition [3]. After the removal, 44,997 items are deleted, and 316,527 valid place entities remain.

3.2 Place Images

Image Collection. We collect all place images from Google Image engine in the public domain. For each location, its name plus its country is used as the keyword for searching. To increase the probability that images are relevant to a particular location, we only download those whose stem words of image titles contain all stem words of the location name. By this process, a total of over $30M$ images are collected from Google Image.

Image Cleaning. There are 28,154 places containing Wikipedia links with 8,125,108 images. We use this subset to further study place-related tasks. Image set is refined by two stages. Firstly, we use Image Hashing technique to remove duplicate images. Secondly, we ask human annotators to remove irrelevant images for each place, including those: 1) whose main body represents another place; 2) that are selfie images with faces occupying a large proportion; 3) that are maps indicating the geolocation of the place. In total, 4,795,778 images are kept to form this subset. For those places without category labels, we manually annotate the labels for them. And after merging some similar labels, we obtain 50 categories.

Placepedia also helps solve some problems on place understanding, like label confusion and label noise. On one hand, all the labels are collected automatically

[2] https://cloud.google.com/natural-language
[3] https://nlp.stanford.edu/software/CRF-NER.html

Table 2. The description and examples for the 10 *function* labels of Places-Fine and Places-Coarse, which are collected from Wikivoyage

Label	Description	Category examples	Place examples
See	People can enjoy beautiful scenes, arts, and architectures therefrom.	Park, Tower, Museum, Gallery, Historical_site	Bergen Aquarium, Lake Parramatta
Do	People can do significant things such as reading books, watching movies, going on vacation, playing sports.	Library, Theater, Resort, Sport, Beach	Apollo Theater, Moray Golf Club
See & Do	People can *see* and *do* in these places. For instance, people can not only enjoy mountain landscape but also climb them.	Land_nature, Theater, Resort, Park, Island	Treasure Island, Sydney Opera House
Sleep	People can have a sleep there.	Splurge_hotel, Mid-range_hotel	Othon Palace Rio, Kviknes Hotel
Eat	People can eat food there.	Restaurant, Street	Louis' Lunch, Taste on Willis
Drink	People can drink something there.	Café, Pub, Street	The Oxford Bar, Cafe Brasilero
Get in	Places for intercity or intercountry transportation.	Airport, Train_station, Public_transportation	Treviso Airport, Kaohsiung Station
Get around	People can travel from one place to another inside a city or a town, such as bus stations, metros, and some ports.	Train_station, Public_transportation	Lujiazui Station, Rathen Ferry
Buy	People usually go shopping there.	Mall, Market, Street, Shop, Town, Square	Labrador Mall, Alexa Centre
Learn	People usually learn new knowledge or skills there.	University, Sport	Kyoto University, St. Clair College

from the Wikivoyage website. Since it is a popular website that provides worldwide travel guidance and the labels in Wikivoyage have been well organized, there are less label confusion in Placepedia dataset. On the other hand, we have manually checked the labels in Placepedia, which would significantly reduce label noise.

From the examples shown in Fig. 2, we observe that: 1) Images of places may look changeable from daytime to nighttime or during different seasons; 2) It can be significantly different viewed from multiple angles; 3) The appearances from inside and outside usually have little in common; 4) Some places such as universities span very large area and consist of different types of small places. These factors make place-related tasks very challenging. In the rest of this paper, we conduct a series of experiments to demonstrate important challenges in place understanding as well as strong expressive power of city embeddings.

4 Study on Comprehensive Place Understanding

This section introduces our exploration on comprehensive place understanding. Firstly, we carefully design benchmarks, and we evaluate the dataset with a lot of state-of-the-art models with different backbones for different tasks. Secondly, we develop a multi-task model, PlaceNet, which is trained to simultaneously predict place items, categories, functions, cities, and countries. This unified framework for place recognition can serve as a reasonable baseline for further studies in our Placepedia dataset. From the experimental results we also demonstrate the challenges of place recognition on multiple aspects.

4.1 Benchmarks

We build the following benchmarks based on the well-cleaned Placepedia subset, for evaluating different methods.

Datasets

- *Places-Coarse.* We select 200 places for validation and 400 places for testing, from 50 famous cities of 34 countries. The remained 27,554 places are used for training. For validation/testing set, we double checked the annotation results. Places without category labels are manually annotated. After merging similar items of labels, we obtain 50 categories and 10 functions. The training/validation/testing set have $3M/60K/120K$ images respectively, from 4,758 cities of 201 countries.
- *Places-Fine.* Places-Fine shares the same validation/testing set with Places-Coarse. For training set, we selected 400 places from the 50 cities of validation/testing places. Different from Places-Coarse, we also double checked the annotation of training data. The training/validation/testing set have $110K/\ 60K/120K$ images respectively, which are tagged with 50 categories, 10 functions, 50 cities, and 34 countries.

Tasks

- *Place Retrieval.* This task is to determine if two images belong to the same place. It is important when people want to find more photos of places they adore. For validation and testing set, 20 images for each place are selected as queries and the rest images form the gallery. Top-k retrieval accuracy is adopted to measure the performance of place retrieval, such that a successful retrieval is counted if at least one image of the same place has been found in the top-k retrieved results.
- *Place Categorization.* This task is to classify places into 50 place categories, e.g. *museums*, *parks*, *churches*, *temples*. For place categorization, we employ the standard top-k classification accuracy as evaluation metric.
- *Function Categorization.* This task is to classify places into 10 place functions: *See*, *Do*, *Sleep*, *Eat*, *Drink*, *See & Do*, *Get In*, *Get Around*, *Buy*, *Learn*. Again, we employ the standard top-k classification accuracy as evaluation metric.
- *City/Country Recognition.* This task is to classify places into 50 cities or 34 countries. The goal is to determine what city/country an image belongs to. Also, the standard top-k classification accuracy is applied as evaluation metric.

4.2 PlaceNet

We construct a CNN-based model to predict all tasks simultaneously. The training procedure performs in an iterative manner and the system is learned end-to-end. **Network Structures.** The network structure of PlaceNet is similar to ResNet50 [15], which has been demonstrated powerful in various vision tasks. As illustrated

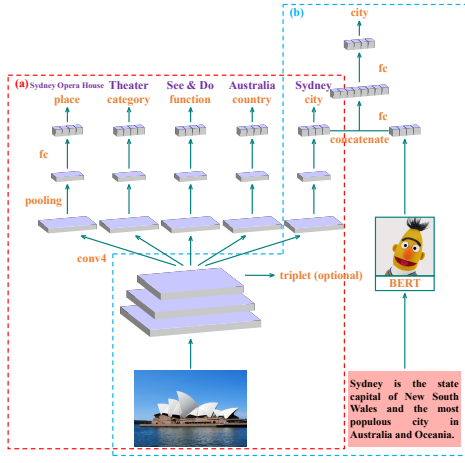

Fig. 5. (a) Pipeline of PlaceNet, which learns five tasks simultaneously. (b) Pipeline of city embedding, which learns city representations considering both vision and text information

in Figure 5 (a), the structures of PlaceNet below the last convolution layer are the same as ResNet50. The last convolution/pooling/fc layers are duplicated to five branches, namely, *place*, *category*, *function*, *city*, and *country*, which is carefully designed for places. Each branch contains two FC layers. Different loss functions and pooling methods are studied in this work.

Loss Functions. We study three losses for PlaceNet, namely, softmax loss, focal loss, and triplet loss. *Softmax loss* or *Focal loss* [24] is adopted to classify place, category, function, city, and country. To learn the metric described by place pairs, we employ *Triplet loss* [38], which enforces distance constraints among positive and negative samples. When using triplet loss, the network is optimized by a combination of $L_{softmax}$ and $L_{triplet}$.

Pooling Methods. We also study different pooling methods for PlaceNet, namely, average pooling, max pooling, spatial pyramid pooling [14].Spatial pyramid pooling (SPP) is used to learn multi-scale pooling, which is robust to object deformations and can augment data to confront overfitting.

4.3 Experimental Settings

Data. We use Place-Fine and Place-Coarse defined in Sec. 4.1 as our experimental datasets. Note that Place-Fine and Place-Coarse share the same validation data and testing data, while training data size of the latter is much larger.

Backbone Methods. Deep Convolutional Neural Networks (CNNs) [13, 17, 22] have shown the impressive power for classification and retrieval tasks. Here we choose four popular CNN architectures, **AlexNet** [22], **GoogLeNet** [45], **VGG16** [40], and **ResNet50** [15], then train them on Place-Fine to create backbone models.

Table 3. The experimental results for different methods on all tasks. We vary different pooling methods and loss functions for PlaceNet. Except for the last line, models are trained on Places-Fine. The figures in bold/blue indicate optimal/sub-optimal performance, respectively

		Place		Category		Function		City		Country	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Backbone	AlexNet	33.78	48.19	24.16	53.03	64.97	96.70	12.47	32.52	17.97	43.30
	GoogLeNet	53.48	66.23	26.01	54.81	65.69	97.20	16.34	37.19	20.98	46.43
	VGG16	43.84	59.03	26.89	55.68	65.97	97.11	18.65	41.13	24.86	51.35
	ResNet50	54.53	67.01	25.22	53.62	68.25	96.89	17.15	38.55	19.72	45.51
Pooling	Average	54.33	67.66	25.95	55.07	67.35	97.34	18.73	40.30	24.80	51.03
	Max	49.66	63.26	25.11	54.07	65.45	97.12	16.93	38.18	22.83	48.61
	SPP	28.18	45.55	27.21	53.86	67.02	96.37	15.36	34.48	21.00	43.08
Loss	Softmax	54.31	67.66	25.95	55.07	67.35	97.34	18.73	40.30	24.80	51.03
	Triplet	50.33	64.06	21.15	48.92	64.84	95.61	14.73	36.56	20.43	46.66
	Focal	55.03	67.38	25.27	55.48	67.62	97.53	18.67	40.87	24.73	51.46
PlaceNet on Places-Coarse		**67.85**	**79.35**	**40.42**	**68.98**	**75.48**	**97.58**	**29.25**	**53.47**	**35.83**	**63.78**

Training Details. We train each model for 90 epochs. For all tasks and all methods, the initial learning rate is set to be 0.5. And the learning rate is multiplied by 0.1 at epoch 63 and epoch 81. The weight decay is $1e^{-4}$. For the optimizer, we use stochastic gradient descent with 0.9 momentum. We also augment the data following the operation on ImageNet, including randomly cropping and horizontally flipping the images. All images are resized to 224×224 and normalized with mean $[123, 117, 109]$ and standard deviation $[58, 56, 58]$. Each model is pre-trained on ImageNet and then trained with our Placepedia Dataset in an end-to-end manner.

All experiments are conducted on Place-Fine. And we also train our PlaceNet on Place-Coarse to see if larger scale of datasets can further benefit the recognition performance.

4.4 Analysis on Recognition Results

Quantitative evaluations of different methods on the four benchmarks are provided. Table 3 summarizes the performance of different methods on all tasks. We first analyze the results on Places-Fine for all benchmark tasks.

Place Retrieval. PlaceNet with focal loss achieves the best retrieval results when evaluated using the top-1 accuracy. Some sample places with high/low accuracies are shown in Fig. 6 (a). We observe that: 1) Places with distinctive architectures can be easily recognized, *e.g. Banco de México* and *Temple of Hephaestus*. 2) For some parks, *e.g. Franklin Park*, there is usually no clear evidence to tell them from other parks. The same scenario can take place in categories such as gardens and churches. 3) Big places like *Fun Spot Orlando* may contain several small places, where their appearance may have nothing in common, which makes it very difficult to recognize. Places like resorts, towns, parks, and universities suffer the same issue.

Place Categorization. The best result is yielded by PlaceNet plus SPP. Some sample categories with high/low accuracies are shown in Fig. 6 (b). We observe that: 1) *Zoos* are the most distinctive. Intuitively, if animals are seen in a place,

Fig. 6. The 4 tables show the performance of 4 tasks, where each presents the most and the least accurate 5 classes. Below each table are 4 sets of examples, including 2 green/red dash boxes representing sample classes with high/low accuracies. Inside each dash box is the ground truth at the top and three images associated with predicted labels. Green/red solid boxes of images mean right/wrong predictions

that is probably a zoo. However, photos in zoos may be mistaken for taking in nature parks. 2) *Tombs* can be confused with *Pubs*, due to bad illumination condition.

Function Categorization. The best setting for learning the function of a place is to use ResNet models. Some sample functions with high/low accuracies are shown in Fig. 6 (c). 1) *See* is recognized with the highest accuracy. 2) Some examples of *Buy* are very difficult to identify, *e.g.* the third image in *Buy*. Even human cannot tell what a street is mainly used for. Is it for shopping, eating, or just for transportation? Same logic applies to shops. The images of *Eat* are often categorized as shops for *buy*ing or *drink*ing. One possible way to recognize the function of a shop is to extract and analyze its name, or to recognize and classify the food type therein. 3) Universities are often unrecognized either, due to its large area with various buildings/scenes.

City/Country Recognition. From Fig. 6 (d), we observe that: 1) Cities with long history (*e.g. Florence*, *Beijing*, *Cairo*) are more likely to distinguish from others, because they often preserve the oldest arts and architectures. 2) Travelers often conclude that Taiwan and Japan look quite alike. The results do show that places of *Taipei* may be regarded as in *Tokyo*. 3) Although places can be wrongly classified to another city, the prediction often belongs to the same country with the ground truth city. For instance, Florence and Milan are both in Italy; Beijing and Shanghai are both in China. The results of country recognition are not presented here. They demonstrate similar findings to city recognition.

To conclude, we see that place-related tasks are often very challenging: 1) Places of parks, gardens, churches, *etc.*are easy to classify; However, it is difficult to distinguish one park/garden/church from another. 2) Under bad environmental condition, photos can be extremely difficult to categorize; 3) To recognize the function of a street or a shop is non-trivial, *i.e.* it is hard to determine their use for people to have a dinner, take a drink, or go shopping. 4) Cities of long history such as Beijing and Florence are often recognized with a high accuracy. While images of others are more likely to be misclassified as similar ones inside and outside their countries. We hope that Placepedia with its well-defined benchmarks can foster more effective studies and thus benefit place recognition and retrieval. The last line of Tab. 3 shows that, to train PlaceNet on larger amount of data, we can further obtain performance gain, by 7 to 16 percent on different tasks.

5 Study on Multi-Faceted City Embedding

We embed each city in an expressive vector to understand places on a city level. Also, the connections between the visual characteristics and the underlying economic or cultural implications are studied therefrom.

5.1 City Embedding

City embedding is to use a vector to represent a city, the items of which indicate different aspects, *e.g.* the economy level, the cultural deposits, the politics atmosphere, *etc.*. In this study, cities are embedded from both visual and textual domains. 1) Visual representations of cities are obtained by extracting features from models supervised by city items. 2) Leading paragraphs collected from Wikipedia are used as the description of each city. [6] provided a pre-trained model on language understanding to embed the content of texts into numeric space. We use this model to extract the textual representations for all cities.

Network Structure. The model for city embedding is illustrated in Figure 5 (b). The input is constructed by concatenating visual and textual vectors. Two fully connected layers are then applied to learn city embedding representations. The corresponding activation functions are ReLU. At last, a classifier and cross entropy loss are used to supervise the learning procedure.

Representative Vectors. We train the network iteratively. The well-trained network is then used to extract the embedding vectors for all images. City embeddings are then acquired by averaging image embeddings city-wise.

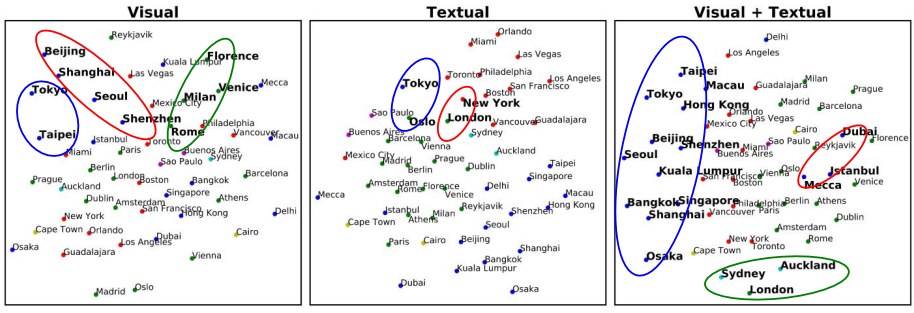

Fig. 7. These three figures show t-SNE representation for city embeddings using vision, text, and vision & text info, respectively. Points with the same color belong to the same continent. We can see that learning from both generates the best embedding results

5.2 Experimental Results

We analyze city embedding results from two aspects. Firstly, we compare the expressive power of embeddings using different information, namely vision, text, and vision & text, in order to see if learning from both can yield a better city representation. Secondly, we investigate the embedding results neuron-wise to explore what kinds of images can express economic/cultural/political levels of cities the most.

Visual and Text Embedding. In Fig. 7, we demonstrate three embedding results using t-SNE [27]. 1) The left graph shows embeddings using only visual features . We observe that it tends to cluster cities that are visually similar. For example, Tokyo looks like Taipei; Beijing, Shanghai, and Shenzhen are all from China, and Seoul shares lots of similar architectures with them; Florence, Venice, Milan, and Rome are all Italy cities. 2) The graph on the middle shows embeddings using only textual features. We can see that textual features usually express the functionality and geolocation of a city. For example, Tokyo and Oslo are both capitals; London and New York are both financial centers. However, they are not visually alike. Also, cities from the same continent are clustered. 3) The right graph shows embeddings learned from both visual and textual domains. They can express the resemblance visually and functionally. For example, cities from east/west-Asia are all clustered together, and cities from Commonwealth of Nations like Sydney, Auckland, and London, are also close to each other on the graph. From the comparison of these graphs, we conclude that learning embeddings from both vision and text content produces the most expressive power of cities.

Economic, Cultural, or Political. We follow the work in [42] to represent each city in three dimensions, namely *economy*, *culture*, and *politics*. In [42], word lists indicating *economy*, *culture*, *politics* are predefined. In this work, leading paragraphs of Wikipedia pages are adopted as our city description. For each city, we calculate the weights of *economic*, *cultural*, and *political* therefrom as in [42]. And we match each neuron to them using Pearson correlation [3], in

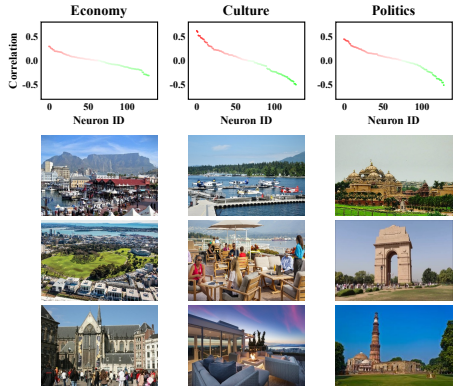

Fig. 8. Three graphs rank Pearson correlation based on neurons in terms of *economy*, *culture*, and *politics*. Below each presents top-3 places activating the neuron of the largest correlation value

order to quantify the connection between each neuron and them. Quinnipiac University[4] concludes that a correlation above 0.4 or below −0.4 can be viewed as a strong correlation. From Fig. 8, we see that neurons can express *culture* most confidently, with the highest correlation score larger than 0.6. This is consistent with our knowledge, *i.e.* culture usually is expressed from distinctive architectures or some unique human activities. Looking at the top-3 places that activate the most relevant neuron, we observe that: 1) Economy level is usually conveyed by a cluster of buildings or the crowd on streets, indicating a prosperous place; 2) Cultural atmosphere can be expressed by distinguished architecture styles and human activities; (3) Political elements are often related to temples, churches, and some historical sites, which usually indicate religious activities and politics-related historical movements.

6 Conclusion

In this work, we construct a large-scale place dataset which is comprehensively annotated with multiple aspects. To our knowledge, it is the largest place-related dataset available. To explore place understanding, we carefully build several benchmarks and study contemporary models. The experimental results show that there still remains lots of challenges in place recognition. To learn city embedding representations, we demonstrate that learning from both visual and textual domains can better characterize a city. The learned embeddings also demonstrate that *economic*, *cultural*, and *political* elements can be represented in different types of images. We hope that, with comprehensively annotated Placepedia contributed to the community, more powerful and robust systems will be developed to foster future place-related studies.

[4] http://faculty.quinnipiac.edu/ libarts/polsci/Statistics.html

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: CVPR. pp. 5297–5307 (2016) 4
2. Arandjelović, R., Zisserman, A.: Dislocation: Scalable descriptor distinctiveness for location recognition. In: Asian Conference on Computer Vision. pp. 188–204. Springer (2014) 4
3. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009) 13
4. Cao, S., Snavely, N.: Graph-based discriminative learning for location recognition. In: CVPR. pp. 700–707 (2013) 4
5. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., et al.: City-scale landmark identification on mobile devices. In: CVPR 2011. pp. 737–744. IEEE (2011) 3, 4
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 12
7. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes paris look like paris? (2012) 4
8. En, S., Lechervy, A., Jurie, F.: Rpnet: an end-to-end network for relative camera pose estimation. In: ECCV. pp. 0–0 (2018) 3
9. Gavves, E., Snoek, C.G.: Landmark image retrieval using visual synonyms. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1123–1126. ACM (2010) 4
10. Gavves, E., Snoek, C.G., Smeulders, A.W.: Visual synonyms for landmark image retrieval. Computer Vision and Image Understanding **116**(2), 238–249 (2012) 4
11. Gronat, P., Havlena, M., Sivic, J., Pajdla, T.: Building streetview datasets for place recognition and city reconstruction. Research Reports of CMP, Czech Technical University in Prague (2011) 3, 4
12. Gronat, P., Obozinski, G., Sivic, J., Pajdla, T.: Learning and calibrating per-location classifiers for visual place recognition. In: CVPR. pp. 907–914 (2013) 4
13. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: CVPR. pp. 1026–1034 (2015) 9
14. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015) 9
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 8, 9
16. Hong, Z., Petillot, Y., Lane, D., Miao, Y., Wang, S.: Textplace: Visual place recognition and topological localization through reading scene texts. In: ICCV 2019 (2019) 4
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018) 9
18. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: ECCV. pp. 304–317. Springer (2008) 2, 3
19. Johns, E., Yang, G.Z.: Ransac with 2d geometric cliques for image retrieval and place recognition. In: CVPR Workshop. pp. 4321–4329 (2015) 4

20. Kang, Y., Jia, Q., Gao, S., Zeng, X., Wang, Y., Angsuesser, S., Liu, Y., Ye, X., Fei, T.: Extracting human emotions at different places based on facial expressions and spatial clustering analysis. Transactions in GIS (2019) 4
21. Knopp, J., Sivic, J., Pajdla, T.: Avoiding confusing features in place recognition. In: ECCV. pp. 748–761. Springer (2010) 4
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012) 9
23. Li, Y., Crandall, D.J., Huttenlocher, D.P.: Landmark classification in large-scale image collections. In: 2009 IEEE 12th international conference on computer vision. pp. 1957–1964. IEEE (2009) 4
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: CVPR. pp. 2980–2988 (2017) 9
25. Lopez-Antequera, M., Gomez-Ojeda, R., Petkov, N., Gonzalez-Jimenez, J.: Appearance-invariant place recognition by discriminatively training a convolutional neural network. Pattern Recognition Letters **92**, 89–95 (2017) 4
26. Lu, H., Zhang, C., Liu, G., Ye, X., Miao, C.: Mapping china's ghost cities through the combination of nighttime satellite data and daytime satellite data. Remote Sensing **10**(7), 1037 (2018) 4
27. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008) 13
28. Milford, M., Shen, C., Lowry, S., Suenderhauf, N., Shirazi, S., Lin, G., Liu, F., Pepperell, E., Lerma, C., Upcroft, B., et al.: Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition. In: CVPR Workshops. pp. 18–25 (2015) 4
29. Mishkin, D., Perdoch, M., Matas, J.: Place recognition with wxbs retrieval. In: CVPR 2015 workshop on visual place recognition in changing environments. vol. 30 (2015) 4
30. Nice, K.A., Thompson, J., Wijnands, J.S., Aschwanden, G.D., Stevenson, M.: The'paris-end'of town? urban typology through machine learning. arXiv preprint arXiv:1910.03220 (2019) 4
31. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: CVPR. pp. 3456–3465 (2017) 2, 3, 4
32. Panphattarasap, P., Calway, A.: Visual place recognition using landmark distribution descriptors. In: Asian Conference on Computer Vision. pp. 487–502. Springer (2016) 4
33. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR 2007. pp. 1–8. IEEE (2007) 2, 3
34. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR 2008. pp. 1–8. IEEE (2008) 2, 3
35. Sattler, T., Havlena, M., Radenovic, F., Schindler, K., Pollefeys, M.: Hyperpoints and fine vocabularies for large-scale location recognition. In: CVPR. pp. 2102–2110 (2015) 4
36. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: BMVC. vol. 1, p. 4 (2012) 4
37. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: CVPR 2007. pp. 1–7. Citeseer (2007) 4
38. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015) 9

39. Shi, X., Khademi, S., van Gemert, J.: Deep visual city recognition visualization. arXiv preprint arXiv:1905.01932 (2019) 4
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9
41. Sizikova, E., Singh, V.K., Georgescu, B., Halber, M., Ma, K., Chen, T.: Enhancing place recognition using joint intensity-depth analysis and synthetic data. In: ECCV. pp. 901–908. Springer (2016) 4
42. Son, J.S., Thill, J.C.: Is your city economic, cultural, or political? recognition of city image based on multidimensional scaling of quantified web pages. In: Spatial Analysis and Location Modeling in Urban and Regional Systems, pp. 63–95. Springer (2018) 4, 13
43. Stumm, E., Mei, C., Lacroix, S., Nieto, J., Hutter, M., Siegwart, R.: Robust visual place recognition with graph kernels. In: CVPR. pp. 4535–4544 (2016) 4
44. Sun, X., Ji, R., Yao, H., Xu, P., Liu, T., Liu, X.: Place retrieval with graph-based place-view model. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. pp. 268–275. ACM (2008) 4
45. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015) 9
46. Teichmann, M., Araujo, A., Zhu, M., Sim, J.: Detect-to-retrieve: Efficient regional aggregation for image search. In: CVPR. pp. 5109–5118 (2019) 4
47. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: CVPR. pp. 1808–1817 (2015) 3, 4
48. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: CVPR. pp. 883–890 (2013) 3, 4
49. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008) 4
50. Wang, Y., Lin, X., Wu, L., Zhang, W.: Effective multi-query expansions: Robust landmark retrieval. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 79–88. ACM (2015) 4
51. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3485–3492. IEEE (2010) 4
52. Yang, J., Zhang, S., Wang, G., Li, M.: Scene and place recognition using a hierarchical latent topic model. Neurocomputing **148**, 578–586 (2015) 4
53. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 2, 3, 4
54. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in neural information processing systems. pp. 487–495 (2014) 4
55. Zhu, Y., Wang, J., Xie, L., Zheng, L.: Attention-based pyramid aggregation network for visual place recognition. In: 2018 ACM Multimedia Conference on Multimedia Conference. pp. 99–107. ACM (2018) 4