

Article

Extracting Representative Images of Tourist Attractions from Flickr by Combining an Improved Cluster Method and Multiple Deep Learning Models

Shanshan Han ¹ , Fu Ren ^{1,2,3}, Qingyun Du ^{1,2,3,4,*}  and Dawei Gui ^{5,*} 

¹ School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; hanshan@whu.edu.cn (S.H.); renfu@whu.edu.cn (F.R.)

² Key Laboratory of Geographic Information Systems, Ministry of Education, Wuhan University, Wuhan 430079, China

³ Key Laboratory of Digital Mapping and Land Information Application Engineering, National Administration of Surveying, Mapping and Geoinformation, Wuhan University, Wuhan 430079, China

⁴ Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

⁵ Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan 430079, China

* Correspondence: qydu@whu.edu.cn (Q.D.); guidawei@pric.org.cn (D.G.); Tel.: +86-27-8766-4557 (Q.D.); +86-027-6877-8030 (D.G.)

Received: 5 December 2019; Accepted: 29 January 2020; Published: 31 January 2020



Abstract: Extracting representative images of tourist attractions from geotagged photos is beneficial to many fields in tourist management, such as applications in touristic information systems. This task usually begins with clustering to extract tourist attractions from raw coordinates in geotagged photos. However, most existing cluster methods are limited in the accuracy and granularity of the places of interest, as well as in detecting distinct tags, due to its primary consideration of spatial relationships. After clustering, the challenge still exists for the task of extracting representative images within the geotagged base image data, because of the existence of noisy photos occupied by a large area proportion of humans and unrelated objects. In this paper, we propose a framework containing an improved cluster method and multiple neural network models to extract representative images of tourist attractions. We first propose a novel time- and user-constrained density-joinable cluster method (TU-DJ-Cluster), specific to photos with similar geotags to detect place-relevant tags. Then we merge and extend the clusters according to the similarity between pairs of tag embeddings, as trained from Word2Vec. Based on the clustering result, we filter noise images with Multilayer Perceptron and a single-shot multibox detector model, and further select representative images with the deep ranking model. We select Beijing as the study area. The quantitative and qualitative analysis, as well as the questionnaire results obtained from real-life tourists, demonstrate the effectiveness of this framework.

Keywords: tourist attractions; representative images; geotagged photos; Word2Vec; convolutional networks

1. Introduction

An increasing number of studies related to tourism geography have been conducted in recent years because the tourism industry is making a significant contribution to the global economy: The total spending on tourism abroad in 2016 reached \$1.23 trillion, and international tourist arrivals in 2017 reached 1.32 billion with growth at 4 % per year in eight years [1]. Among these studies, extracting representative images of tourist attractions is unending and practical research. It can provide informative descriptions about the tourist attractions [2]. Furthermore, it can be applied in building touristic information systems [3] and generating tourist maps [4], as well as providing image content to

some content-based tourist recommendations [5]. With the prevalence of image-based content sharing platforms, more and more researchers are inclined to extract tourist attractions from such platforms. Because the photos captured by various users can reflect the actual tourists' preferences more directly compared with few experts' subjective opinions [6]. Furthermore, as a kind of social media data, they can quickly discover newly emerging tourist attractions. Among these platforms, Flickr is one of the most popular platforms. It has more than 100 million registered users and is visited by over 60 million people per month [7]. Besides, it is also dominantly used in previous research related to extracting urban places, since its data is convenient to acquire and most of its contents tend to reflect more information about the surrounding locations compared with other platforms [8].

In contrast to studies that aim at inferring geo-location from images [9,10], representative image selection pays more attention to exploring users' visual preferences and visual similarity to images given a certain location. Therefore, it usually begins with clustering: Cluster the photos according to their coordinates, and then obtain semantic annotations by reverse geocoding [11,12] or selecting distinct tags with TF-IDF [6,13] and its variations [14]. However, most of the cluster methods can only generate a coarse-grained result, for instance, areas of interest. Such results may increase the difficulty of representative image selection. Another distinct way of tourist attraction extraction is to obtain the standard tourist attraction names from travel guide website or external gazetteers as keywords and query with them to get photos which have these tags [15,16]. Nevertheless, such a method also has some disadvantages. First, it cannot guarantee that all the photos related to certain tourist attractions are attached to the official names, because abbreviations, alternate spellings, and even misspelling also exist [17]. Therefore, it may cause low recall when using a single, standard name to retrieve photos. Second, there may exist photos attached to the name of some tourist attractions, but not located in the actual locations, due to mislabelling or other reasons, which can be considered as noise points [18]. Therefore, a clustering method that generates fine-grained results is needed for accurate results of representative image selection.

After obtaining the cluster results, the challenge still exists, due to the nature of the tourism photos. First, according to the definition in [19], a tourist attraction refers to a place of interest that offers leisure and amusement and attracts tourists to visit, which implies that photos of tourist attractions may contain many people. Besides, many tourists tend to take selfies in front of tourist attractions when traveling [2]. Such images are obviously not suitable to be the representative images of a tourist attraction, and thus, certain filtering processes are needed. However, it is inevitable that some types of places, such as bazaar and square, may have plenty of people. The crowd may be one of the essential components of such places' images. So the undifferentiated filter may fail to find the representative image for some types of tourist attractions. Second, the existence of irrelevant or noise photos in user-generated content makes more it difficult to find satisfactory images. Some tourists may take pictures of local objects within the tourist attractions, for instance, exhibits in a museum; some even take pictures of unrelated things, for instance, a cat or an apple in a museum. This image acquisition circumstances pose unique challenges for geotagged imagery for touristic means, which are underrepresented in the literature, and hence, addressed in this work.

Given the challenges described above, in this paper, we propose a framework that combines an improved cluster method and multiple neural network models to extract representative images of tourist attractions from Flickr. We first modified the density-joinable cluster [20] by adding the constraint of time and user threshold, which aims to detect place-relevant tags and simultaneously filter some tags related to temporary events occurred in a fixed place or frequently used by one user. Then considering the existence of alternate spellings of place names, we further merge and extend these tag-based clusters according to both spatial relationships among clusters and semantic similarity among tags. The semantic similarity is obtained from the cosine similarity of Word2Vec embeddings of tags, because Word2Vec can make words get closer in the semantic distance and be more applicable to calculate the similarity of different text granularities [21]. Based on the cluster results, we filter the noisy images with objects by Multilayer Perceptron (MLP) and those with humans by a single-shot

multibox detector (SSD) model [22]. Then the similarity of the filtered images is ranked by the deep ranking model [23], returning a ranked location list with representative images. We choose Beijing as the study area. The results of the proposed cluster methods and those of the existing methods are compared in clustering number, clustering accuracy and semantic distinctiveness, and the results of representative images selection are analyzed qualitatively. Additionally, a questionnaire is also conducted to evaluate whether the overall results meet the satisfaction of tourists in real life. A series of achieved results demonstrate the effectiveness of the framework.

The remainder of this paper is organized as follows. Section 2 reviews related work on clustering methods for geotagged photos, and selecting representative images for the extracted locations. Section 3 introduces the preliminary and the overall framework of extracting tourist attractions and their representative images. Section 4 describes the study area and discusses the implementation results and evaluation. Section 5 summarizes this paper and points out the directions for future work.

2. Related work

2.1. Geotagged Photo Clustering

Clustering is the premise and basis for representative image selection from geotagged photos. Among these clustering methods, density-based cluster methods are widely applied in clustering geotagged photos because they do not need to predefine the number of clusters and can filter noise points [11,24]. These density-based cluster methods include DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [6,8,24] and various modified DBSCAN, such as P-DBSCAN [25], a method that considers a pre-set radius with the minimum number of photo owners [11,26]. After clustering, some directly reverse geocode these clusters and identify the corresponding place names by using tools, such as Geonames [12] and Google Places API [11]. However, the diversity of reverse geocoding results makes it difficult to judge the accuracy, and sometimes positioning errors make it worse [27,28]. Other articles consider leveraging textual contents attached to geotagged photos to obtain more accurate place names, mainly by calculating TF-IDF or its variants of each cluster and find the representative tags as place names or information [14,29]. Nevertheless, most of the cluster methods can only generate a coarse-grained cluster result, for instance, areas of interest, which possibly contain more than one tourist attraction, especially in the area with high density uploaded photos. Such results are not beneficial to the further application, such as tourist attraction recommendations.

A small proportion of researchers choose to retrieve geotagged photos with standard names of tourist attractions [15,30]. This may cause low recall and contain noise. A few studies attempted to detect place tags in geotagged photos without referring to any gazetteer by clustering photos with the same tag and analyzing the spatial distribution [18]. As mentioned above, there are plenty of abbreviations and alternate spellings of standard place names in the tag sets, which are non-trivial to resolve. However, when merging the place tags, researchers in previous studies mainly consider the similarity of spatial distribution between tags, and few consider the semantic similarity of place tags. Furthermore, some tags related to events taken place in fixed places or frequently used by few users cannot be filtered. Therefore, further improvement is needed in clustering places with geotagged photos.

2.2. Representative Image Selection

Representative image selection is a popular, but also challenging study, due to the existence of noisy images in geotagged photos. A few studies used supervised learning methods to extract representative images of certain places. For instance, Crandall et al. [31] utilize an SVM model to distinguish tourist attraction photos from negative ones obtained from other locations. Similarly, Samany [32] applies a deep belief network to classify landmarks in Tehran, Iran. Kim et al. [33] seek another way to categorize and analyze the representative image of major components in each area of interest in Seoul with Inception v3 model that is pre-trained with ImageNet. However, labeling

images for supervised learning costs extensive manual labor [34]. Besides, it is hardly possible to train classifiers for every tourist attraction in the world [30]. Therefore, a more common approach is to compare image properties and find similar images after clustering or extracting places from geotags and tags. Among those image properties, SIFT is frequently used [18,31]. Other properties are also used, including GIST [29,35], color histograms [36], etc. For better representation, some studies may combine more than one image property [36,37]. However, the performance may be limited by the representation of these hand-crafted features to a great extent [23]. With the development of convolutional neural networks, researchers gradually try to leverage convolutional-based models to finish this task. For instance, Ding and Fan [38] combine SURF (an algorithm similar to SIFT) and LIFT (a deep-learning model) to find representative images and match untagged images to them.

For selecting better representative images, some filtering preprocess also attempted in previous studies. Most of them target to filter images with humans, by either applying a sophisticated library (such as OpenCV) [8] or training a deep learning model for image classification [5]. However, all of the previous studies conducted the undifferentiated filter process, which may fail to find the representative image for some tourist attractions. In addition, apart from images with humans, few consider other types of noise images, such as artificial images (e.g., a logo) and images with objects (e.g., an apple) [39]. In summary, more effort is required to tap the potential of convolutional-based models to apply in representative image selection for tourist attraction.

3. The Overall Framework

3.1. Preliminary

We define the set of photos in a certain study area as $P = \{p_1, p_2, \dots, p_{|P|}\}$, where $\forall p_i \in P$ consists of a tuple of attributes, represented as $p_i = (id_{p_i}, t_{p_i}, l_{p_i}, u_{p_i}, X_{p_i})$. These attributes include the unique photo ID id_{p_i} , taken time t_{p_i} , taken location l_{p_i} (represented by latitude lat_{p_i} and longitude lon_{p_i}), user who contributes this photo u_{p_i} , and a list of tags $X_{p_i} = [x_1, x_2, \dots, x_{|X_{p_i}|}]$. Note that the number of tags in X_{p_i} could be zero or any positive integer, and a tag x can be attached to one or more than one photo. We represent the set of tags as $X = \{x_1, x_2, \dots, x_{|X|}\}$, and the subset of photos that are attached to a specific tag x as $P_x = \cup_{p_i \in P, x \in X} p_i$.

Our goal is to detect the set of place-relevant tags, and further merge and extend these clusters; based on the cluster results, find the representative images of each tourist attraction. Figure 1 shows the overall framework, and each step is illustrated in detail in the following sections.

3.2. Data Acquisition

Harvesting data is the first step of the framework. As mentioned above, the Flickr geotagged photo dataset is an optimal choice for this study, due to its several advantages. Apart from Flickr APIs, the datasets can also be conveniently obtained from Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) hosted on Amazon AWS. As a part of the Yahoo Webscope program [40], it provides approximate 100 million public Flickr photos, each including user id, longitude, latitude, user tags, capture time, capture device, photo/video page URL, license URL, etc. This provides an adequate amount of data under a Creative Commons Attribution License and can free researchers from troublesome data crawling work. We leverage geotagged photos from YFCC100M, whose coordinates are bounded in the study area and taken within a certain time. The main features we use in this paper contain: Line number (the unique identification of each geotagged photo), user id (the unique identification of each user), capture time, geotag (longitude and latitude), user tags and the images themselves.

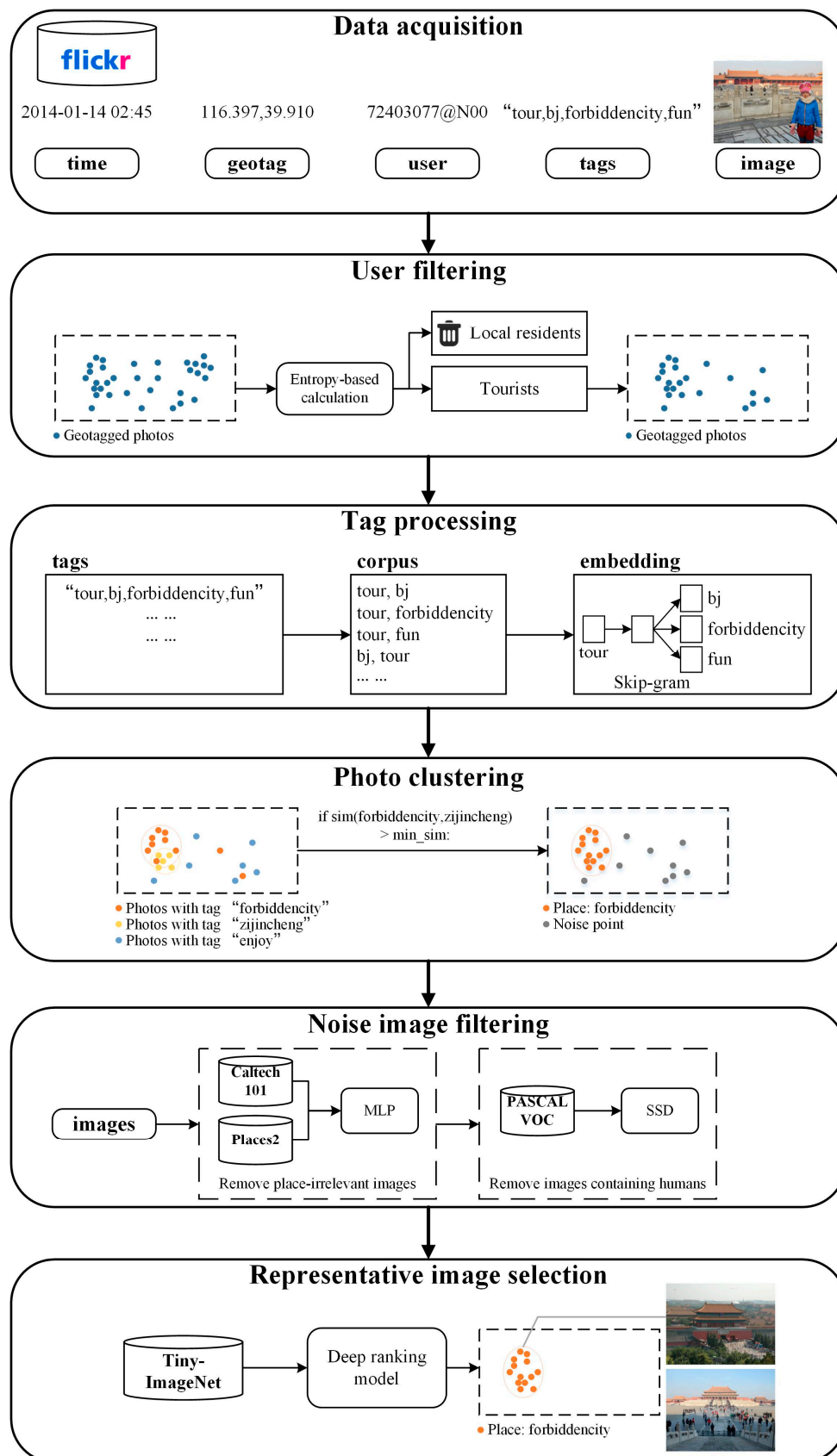


Figure 1. The framework of representative image extraction of tourist attractions from Flickr (These photos in Figure 1 are all licensed by: <http://creativecommons.org/licenses/by-nc-sa/2.0/>).

3.3. User Filtering

Because we aim to extract tourist attractions in this study, geotagged photos uploaded by natives should be removed since most of their check-in records are about daily life and events unrelated to tourism. Similar to the study conducted by Sun et al. [24], we leverage an entropy-based method to distinguish tourists from natives in the tourist destination, formulated as Equation (1):

$$E(u) = - \sum_{m \in \text{Mon}(u)} P_m(u) \cdot \log P_m(u) \quad (1)$$

$$P_m(u) = \frac{D_m(u)}{\sum_{m \in \text{Mon}(u)} D_m(u)} \quad (2)$$

In Equation (2), $D_m(u)$ is the number of days that user u have stayed in the study area in Month m , and $\text{Mon}(u)$ is the total number of months that user u have stayed in this study area. $P_m(u)$ is the proportion of the number of days in Month m and the total number of days that user u has stayed in the study area. Intuitively, the larger the value of $E(u)$ is, the more dispersed the user u 's visiting distribution is, the less likely he/she is a tourist. So we define a threshold E to remove geotagged photos of the user u if the value of $E(u)$ for user u is larger than E .

3.4. Tag Processing

Preprocessing of tags and Word2Vec training is needed before detecting place tags and clustering places, which resolves common lingual ambiguities, such as white spaces, word separation and capitalization, and regularizes terms. After that, we leverage Word2Vec to extract semantic relationships, where all tags in the study area form the corpus, and tags in each photo form one sentence.

In order to extract the intended semantics, we adapt the word exclusion threshold as a user constraint (i.e., tags that are used by less than a minimum number of users will not be trained). Also, we define the word neighbourhood to encapsulate all phrases attached to the same photo. We leverage Skip-gram in Word2Vec to train the tag sets, which mainly aims to maximize the log-likelihood of the contextual word given the center word, formulated as Equation (3):

$$\Gamma = \frac{1}{|X|} \sum_{t=1}^{|X|} \sum_{x_c \in C(x_t)} \log p(x_c|x_t) \quad (3)$$

where x_t represents the given word, and $C(x_t)$ represents the contents of x_t , and $x_c \in C(x_t)$ (where x_t is exclusive) represents a neighbouring word. Skip-gram defines $p(x_c|x_t)$ using the softmax function. However, the cost of computing Equation (1) is impractically large when using the softmax function. Therefore, the hierarchical softmax function and negative sampling are proposed as two computationally efficient approximation algorithms in Equation (3). In this paper, the hierarchical softmax function is used to improve efficiency, which utilizes a binary Huffman tree to represent the output layer with words and explicitly represents the relative probabilities of the child nodes for each node [41].

3.5. Photo Clustering

Photo clustering includes place tag extraction, merge, and cluster extension. The process of place tag extraction and merge begins with arbitrary tag x that is not yet processed. If the number of photos containing tag x (given as $|P_x|$) is less than the minimum number of photos min_pts then the tag will be marked as a noise tag and continue to process the next tag; otherwise, cluster the photos with TU-DJ-Cluster. It is a modified density-joinable cluster method [20], which is further constrained by a time threshold Δt and a minimum number of users and is specific to geotagged photos. The main process of TU-DJ-Cluster is illustrated in Figure 2: (a) extract all points with the same tag as the clustering dataset per time, where different colors represent photos taken by different

users; (b) calculate the neighbourhood of each point within a radius of eps ; (c) mark points with no neighbourhood as noise points, and join those points with at least one common point; (d) after generating an initial cluster result, further judge whether each cluster meets the conditions of the minimum time threshold and minimum users. If not, mark them as noise points.

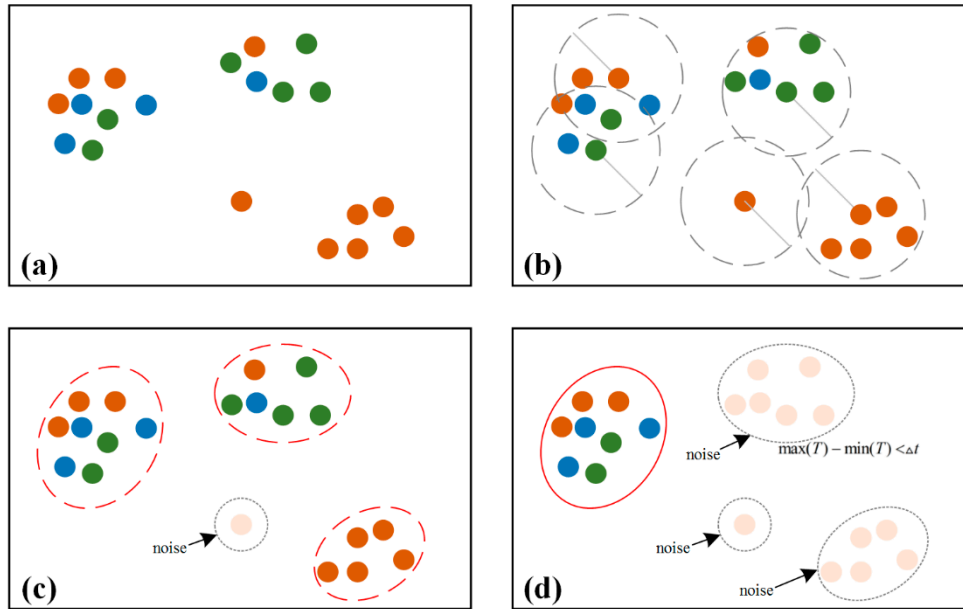


Figure 2. Illustration of TU-DJ-Cluster: (a) the dataset; (b) the process of calculating the neighbourhood; (c) initial cluster results; (d) final cluster results and noise points.

After clustering P_x with TU-DJ-Cluster, we will get the cluster results C_x . If no cluster is generated, mark the tag x as noise tag. Otherwise, loop through these clusters and determine whether there is a cluster that the number of photos accounting for the total number of photos $|P_x|$ is larger than the minimum proportion p_pro . If the cluster c_x^i that matches the above condition exists, then mark tag x as a place tag and create the convex hull with the points in c_x^i .

We further merge some convex hulls according to spatial relationships and semantic similarity of tags. If two convex hulls have the overlay part and the similarity value of their place tags is larger than the minimum threshold min_sim , then merge them. As Equation (4) shows, cosine similarity is used to calculate the similarity of two tags x_i and x_j :

$$similarity(x_i, x_j) = \frac{e_{x_i}^T \cdot e_{x_j}}{\|e_{x_i}\| \|e_{x_j}\|} \quad (4)$$

where e_{x_i} and e_{x_j} represent the embedding of tags x_i and x_j , respectively, which are obtained from the above Word2Vec training. After processing all the convex hulls, we obtain a set of processed convex hulls with different semantics CH_x^i .

The above cluster results only contain a small proportion of photos that are attached with place-relevant tags, because a subset of location-related photos is not tagged accordingly. Therefore, we continue to classify the unprocessed photos according to spatial relationships and semantic similarity to improve recall. The nature of photo acquisition of touristic places causes such photos to be captured within or near the place. So, we create a buffer with radius r for each convex hull in CH_x^i generated by the above steps for further use. Additionally, previous studies show that there is a correlation between tags and geotags [42], so we assume that photos taken in the adjacent location are inclined to assign similar tags. We judge if the unclassified photos are located within any convex hull in CH_x^b , and if

there exists any attached tag whose similarity with the name of the convex hull is larger than min_sim . The final output is a set of clusters that represent tourist attractions with different semantics.

3.6. Noise Image Filtering

Images that are place-irrelevant or are occupied by a large area proportion of humans are removed with multiple pre-trained models. Inspired by the study conducted by Zhang et al. [39], we also use the Caltech 101 dataset (an object image dataset) [43], and the Places2 dataset (a scene image dataset with most place types) [44] to train a binary classifier of place-relevant images and place-irrelevant images. Both datasets complement one another for the target binary classification: Caltech 101 depicts individual, human-made objects, whereas, Places2 explicitly shows geographically-locatable landscapes. For training, we randomly select about 4,000 images from each dataset to transform into 2,048 dimension features and feed into Multilayer Perceptron (MLP), and about 2,000 images to evaluate the accuracy. The final classification accuracy reaches to 98.68%.

Next, we apply a single-shot multibox detector (SSD) model [22] to detect persons in images. It is a convolutional-based object detection model, pre-trained on PASCAL Visual Object Classes (VOC) dataset. We assume that if a more substantial proportion of an image is occupied by at least one person, it is more likely to be a tourist's selfie in front of a tourist attraction. Examples are shown in Figure 3. Although both images show the same tourist attraction (The Great Wall in Beijing, China) and are both detected to have two persons, Figure 3a seems more likely to be the representative image for this tourist attraction than Figure 3b. Given this assumption, we detect each image and filter it if there is a person whose minimum bounding rectangle's area covers over 10% of this image.

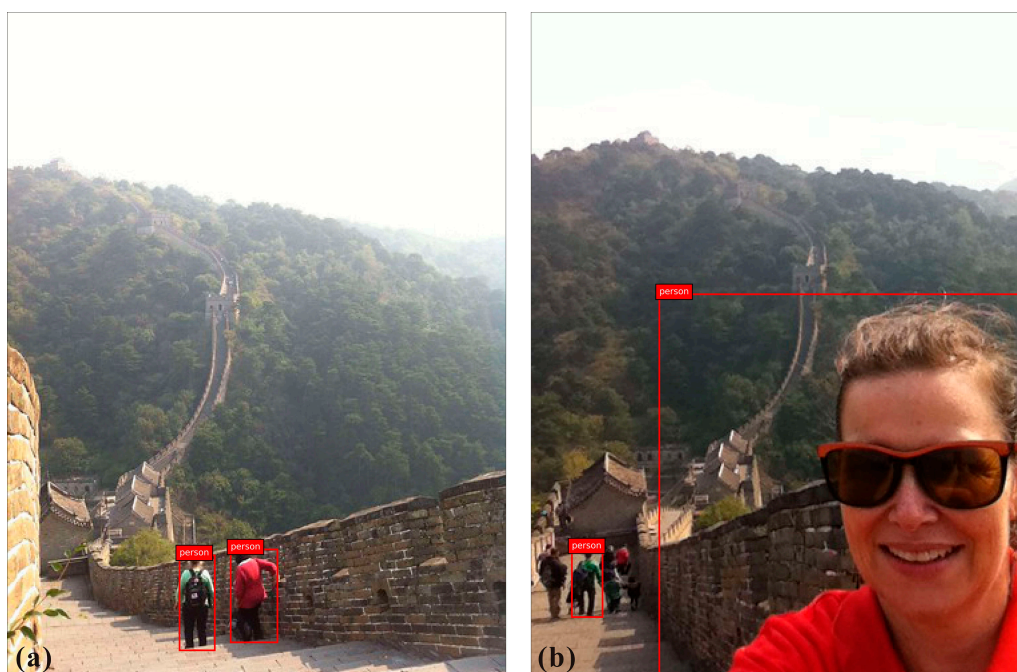


Figure 3. Examples of (a) small proportion and (b) large proportion that detected humans occupy in images (Figure 3a,b were drawn the object detection boxes, and are both licensed by: <http://creativecommons.org/licenses/by-nc-sa/2.0/>).

3.7. Representative Image Selection

After removing noise photos, we train a deep ranking model and find the most representative images of each tourist attraction. The deep ranking model is a convolutional model focusing on fine-grained visual similarity, which is different from most existing models that only focus on category-level similarity [23]. As shown in Figure 4, the model can integrate a commonly-used

convolutional network (ConvNet), such as VGG nets [45] and ResNet [46] with low-resolution paths and normalize their output features. Image triplets, including anchor image, positive image, and negative image, are fed independently into three networks with the same architecture and shared parameters. These embedding outputs of the inputs are leveraged to evaluate the hinge loss, by back-propagating the gradients to the lower layers to optimizing their parameters and minimizing the hinge loss.

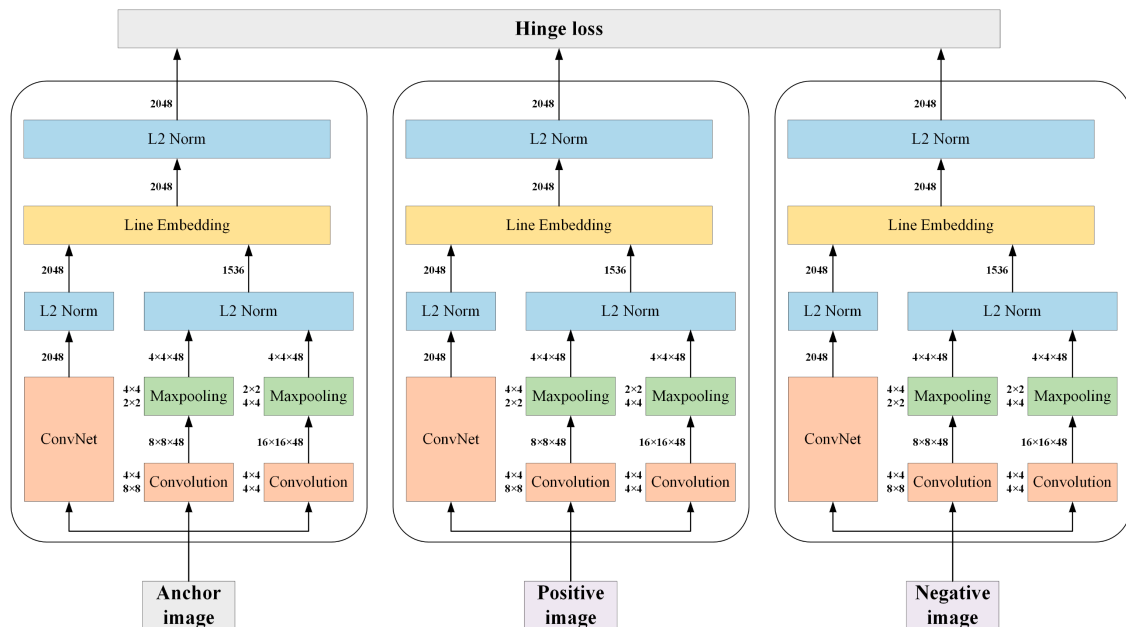


Figure 4. The architecture of the deep ranking model.

In our study, we leverage ResNet as ConvNet in the model and Tiny-ImageNet [47] as the training dataset. For each image in the training dataset, we randomly select one image in the same category as the positive image and one image in any other category as the negative one to create the triplet input. To accelerate the training process, we initialize the ConvNet part of the model with ImageNet weights. After training, we obtain the model weights and transfer them to our dataset.

4. Experimental results

4.1. Study Area

We select Beijing as the study area to verify the framework. Beijing is the capital of China, and also the second-largest city in China. It has abundant tourism resources, and every year it has attracted many tourists at home and abroad [48]. The number of raw images bounded in Beijing is 145,397, and the number of users is 2,846. After filtering users, as Section 3.3 described, the number of images has reduced to 140,891, and the number of users is 2,750. Figure 5 has shown photo distribution in Beijing.

4.2. Result of Place-Relevant Tag Detection

Before applying Word2Vec in tag set, we have analyzed the frequency distribution of tags used in the study (Figure 6a), as well as the number of users using these tags (Figure 6b), and represent them as log-log plots. The plots reveal that they both approximately follow a power-law distribution similar to the word frequency distribution in natural language, indicating that it is applicable to leverage Word2Vec to embed these tags with the limitation condition of user counts. We have set the minimum number of users as three, and the embedding size as 200. After filtering tags, the number of tags reduces from 19,469 to 2,845.

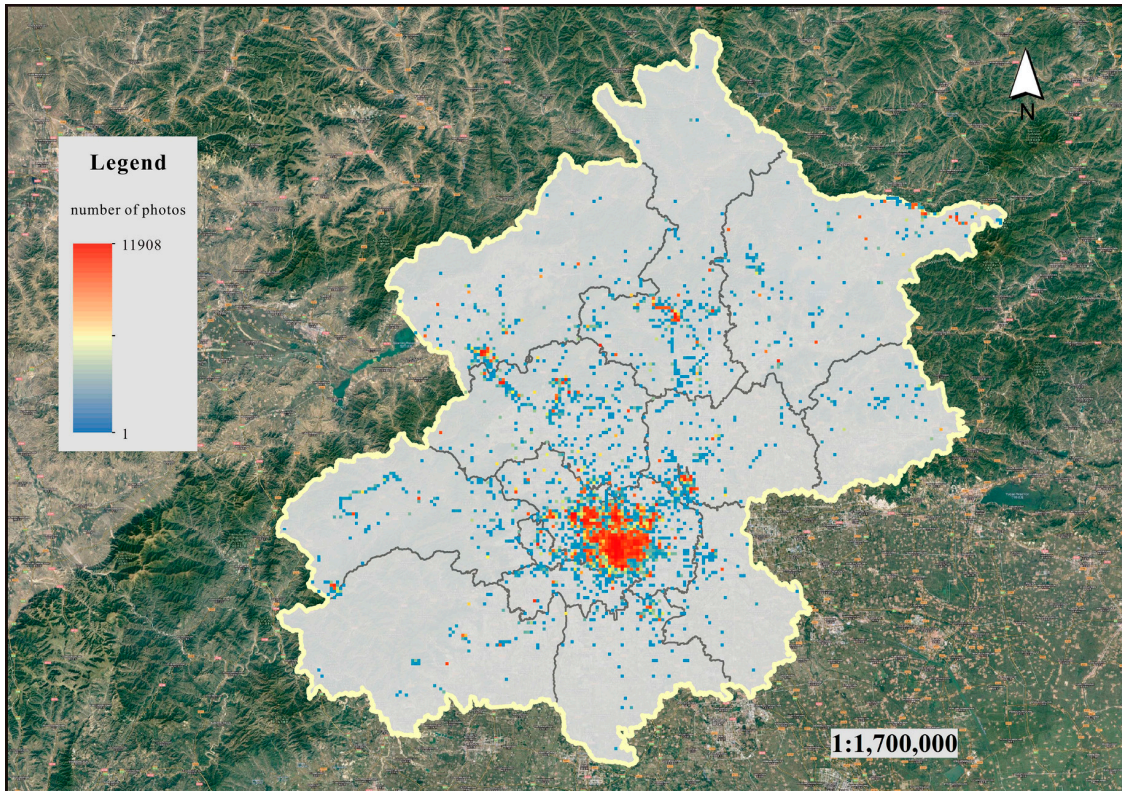


Figure 5. Geotagged photo distribution in the study area: Beijing.

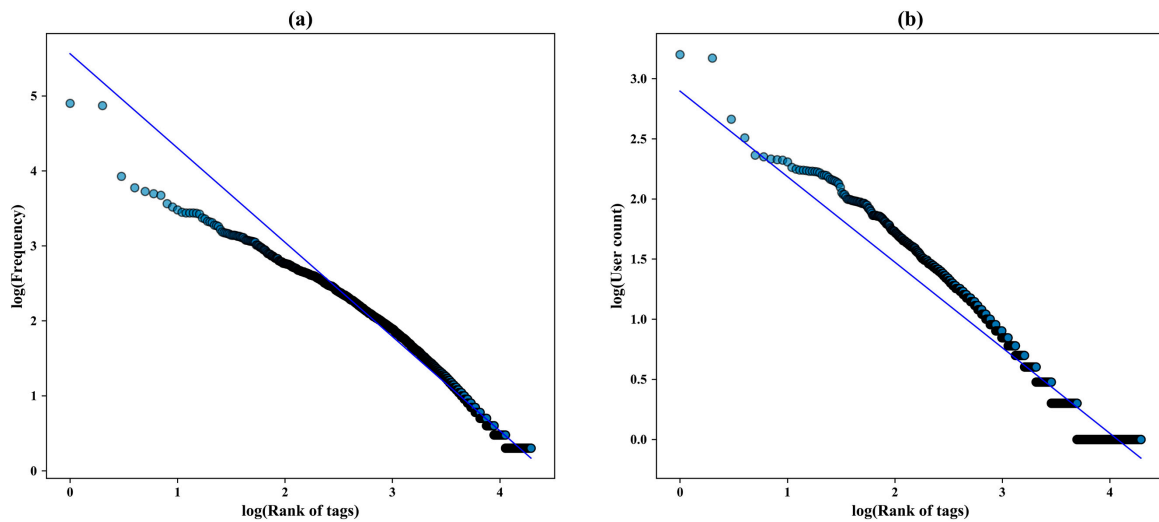


Figure 6. Log-log plots of tags used in the study: (a) Log-log plot on the frequency distribution of tags. (b) Log-log plot on the number of users using these tags.

To evaluate the ability to filter place-irrelevant tags of TU-DJ-Cluster, we compare it with a density-joinable cluster without the constraint of time and user, which can be considered as DBSCAN where $MinPts$ is 1 to some extent. We replace TU-DJ-Cluster with it in the framework of photo clustering. Table 1 shows the values of the parameters for all methods in this experiment. Meanwhile, the baseline method of DBSCAN set both min_users and Δt as zero, representing that there is no restriction on the number of users and time for clustering.

Table 1. The value of TU-DJ-Cluster’s parameters.

Parameter	Value	Notes
<i>eps</i> (meter)	500	Radius of TU-DJ-Cluster
<i>min_pts</i>	30	Minimum number of points for clustering
<i>min_users</i>	2	Minimum number of users
Δt (day)	180	Time threshold
<i>p_pro</i>	0.6	Minimum proportion of point number
<i>min_sim</i>	0.5	Minimum similarity for the merge of tags
<i>r</i> (meter)	300	Radius of buffer for cluster extension

Table 2 lists the detection results of both methods. TU-DJ-Cluster has detected 131 place-relevant tags, while DBSCAN has detected 385 without the constraint of time and user. To better validate the accuracy of place-relevant tag detection, we invite volunteers who are familiar with Beijing to manually mark place-relevant tags, and further use to calculate recall of TU-DJ-Cluster and DBSCAN, respectively, which is defined as the proportion of the number of true place-relevant tags and the number of detected tags, or we can regard it as the hitting ratio. We can see that the hitting ratio of TU-DJ-Cluster is much larger than DBSCAN, which is over 85% of the detected tags are true positive values. Although TU-DJ-Cluster has missed some true place-relevant tags (52 fewer than DBSCAN’s), many of them can be merged with the detected tags in further cluster extension, because the majority of them are alternate spellings or misspelling of the detected tags, which are used by few users. On the contrary, the false-positive values detected by DBSCAN have generated many trivial clusters. Figure 7 shows some misidentification results of DBSCAN when detecting place-relevant tags. It has detected “midi” (Figure 7a, a famous music festival held in Haidian Park, Beijing) and “cnbloggercon” (Figure 7b, a conference related to China’s Blogger) as a place-relevant tag, and also tags related to personal places, such as “office” and “home”, which we do not show in this Figure; Our preferred TU-DJ-Cluster algorithm, on the hand, naturally filters out those semantically-irrelevant tags (see in Supplementary Materials).

Table 2. Place-relevant tag detection results of two methods.

Method	Number of Detected Tags	Number of True Place-Relevant Tags	Hitting Ratio (%)	Number of Clusters after Merge
TU-DJ-Cluster	131	112	85.49	30
DBSCAN	385	164	42.60	123

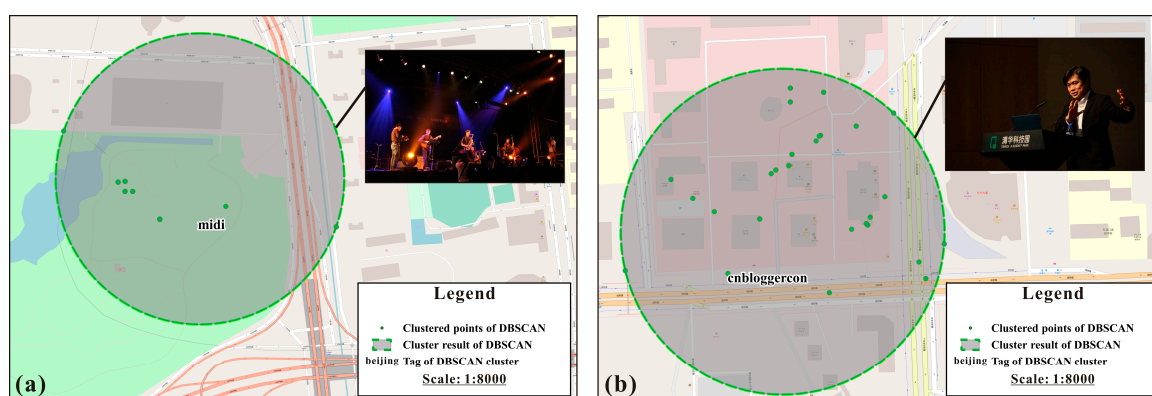


Figure 7. Some misdetection results of DBSCAN: (a) Midi; (b) cnbloggercon (The photos in Figure 7a,b are licensed by: <http://creativecommons.org/licenses/by-nc-nd/2.0/>).

To prove the effectiveness of applying Word2Vec in tag processing and similarity calculation, we also list some results of high similarity among place-relevant tags, while merging the semantic

convex hulls and show in Table 3. The analysis shows that the processing detects and merges synonymous place-relevant tags because the synonymous tags are more likely to have high similarity, including English name and “Pinyin” of a certain tourist attraction (for instance, “altarofheaven” and “tiantanpark”, “oldsummerpalace” and “yuanmingyuan”, etc.), abbreviations (for instance, both “nationalcentreforthetheperformingarts” and “ncpa” can represent National Centre for the Performing Arts.) and alternate names (for instance, both “birdsnest” and “nationalstadium” can represent Beijing National Stadium).

Table 3. Some results of high similarity among place-relevant tags.

Place-Relevant Tag 1	Place-Relevant Tag 2	Similarity
tvcc	cctvbuilding	0.77
oldsummerpalace	yuanmingyuan	0.70
nationalcentreforthetheperformingarts	ncpa	0.67
birdsnest	nationalstadium	0.62
altarofheaven	tiantanpark	0.63
lama	yonghegong	0.59

4.3. Result of Photo Clustering

Following the parameters and process above, we obtain the overall clustering result of our framework. The result contains a total number of 30 clusters, most of which are located in Dongcheng District and Xicheng District, including Tiananmen, the Forbidden City, Wangfujing, Jingshan Park, Drum Tower, etc., and are shown in Figure 8.

For better illustration, we compare our framework with P-DBSCAN and TF-IDF-UF, which we follow the process in the studies of Kennedy et al. [14] and Vu et al. [26]. As expected, P-DBSCAN results extract fewer clusters (16 clusters) with less distinctiveness than TU-DJ-Clustering, which is qualitatively presented in the OpenStreetMap graphics of Figure 9. Both P-DBSCAN and our method have successfully detected the same places of interest, including the Old Summer Palace (also known as “Yuanminyuan”), the art district “798” and the Summer Palace (also known as “Yiheyuan”; Figure 9a). However, because of the unbalanced distribution of point density within these places, the result of P-DBSCAN in Figure 9a does not include the southwest part of it, which is a part of the Summer Palace, shown in OSM. Moreover, because these points located in the northwest part of the P-DBSCAN result get closer to the high-density area, they are included in the cluster, where we randomly check the content of some photos and find that they are not semantically relevant to the Summer Palace. Consequently, although both of them successfully detect the same place of interest, a clustering result that has considered the semantical difference of photos can undoubtedly obtain a fine-grained clustering result and benefit further application.

Figure 9b compares the cluster results of TU-DJ-Cluster and P-DBSCAN around the area of the Forbidden City. Our method has detected different places of interest in this area, while P-DBSCAN has clustered such a wide range of areas into a cluster. Even if we have tested several combinations of the parameters of P-DBSCAN during the experiment, most of them tend to cluster these different places of interest into the same cluster. One possible reason is that these popular tourist attractions of Beijing densely locate in the area around the Forbidden City, which causes a relatively high density of geotagged photos and makes P-DBSCAN difficult to distinguish them. Also, with TF-IDF-UF method, it chooses “beijing”, a relatively unrepresentative tag for this cluster. Such a cluster result may have a bad influence on further applications, such as tourist attraction recommendation. The comparison shows the superiority of our method in detecting fine-grained places of interest and extracting accurate and representative tags to these places of interest over the traditional P-DBSCAN method.



Figure 8. Cluster results of TU-DJ-Cluster (These photos in Figure 8 are all licensed under the Creative Commons Attribution License).

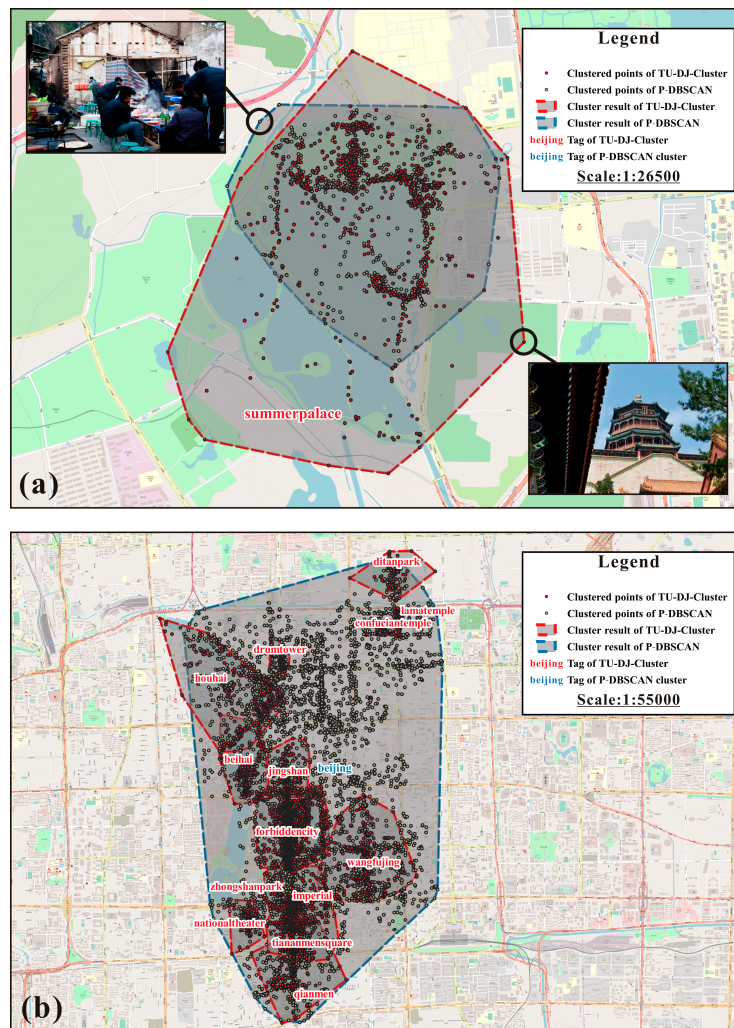


Figure 9. The comparison of results of TU-DJ-Cluster and P-DBSCAN: (a) Summer Palace; (b) area around the Forbidden City (The photos in Figure 9a are licensed under the Creative Commons Attribution License).

4.4. Result of Representative Image Selection

Based on the above cluster results of TU-DJ-Cluster, we further collect the corresponding images in each cluster to filter and find representative images. We exhibit the overall filtering result of each cluster with a stacked bar chart in Figures 10 and 11. Figure 10 is the absolute number of images, and these tourist attractions are sorted by the total number of images, which reflects the popularity of each tourist attraction to some extent. As Figure 10 shows, the Forbidden City is the most popular tourist attractions, since the number of images far exceeds others. The following are the Olympic Park, the Summer Palace, and Tiananmen Square. Figure 11 shows the proportion of different types of image contents. Capital Museum, Zoo, and Zhongshan Park attract tourists mainly by historical relics, pandas, and tulip, respectively. Tourist attractions like them have a relatively high proportion of images related to objects. This result indicates that tourists are more fond of taking photos of objects when visiting these types of tourist attractions. On the contrary, images with humans are dominant in tourist attractions, such as Wangfujing and Ditan Park. Regarding Wangfujing, it is easy to explain because it is a shopping area with a massive flow of people. Ditan Park equally attracts tourists by its vibrant temple fairs, and thus, has many images containing humans. In addition, the chart indicates that tourist attractions with magnificent appearance can appeal to tourists to take more overall photos of them because images related to scenes account for over 60% with tourist attractions like CCTV

Building, National Theater, and Yuanmingyuan. To sum up, tourists do show different preferences when taking photos of different types of tourist attractions, and the difficulty of representative image selection also varies from different types of tourist attractions.

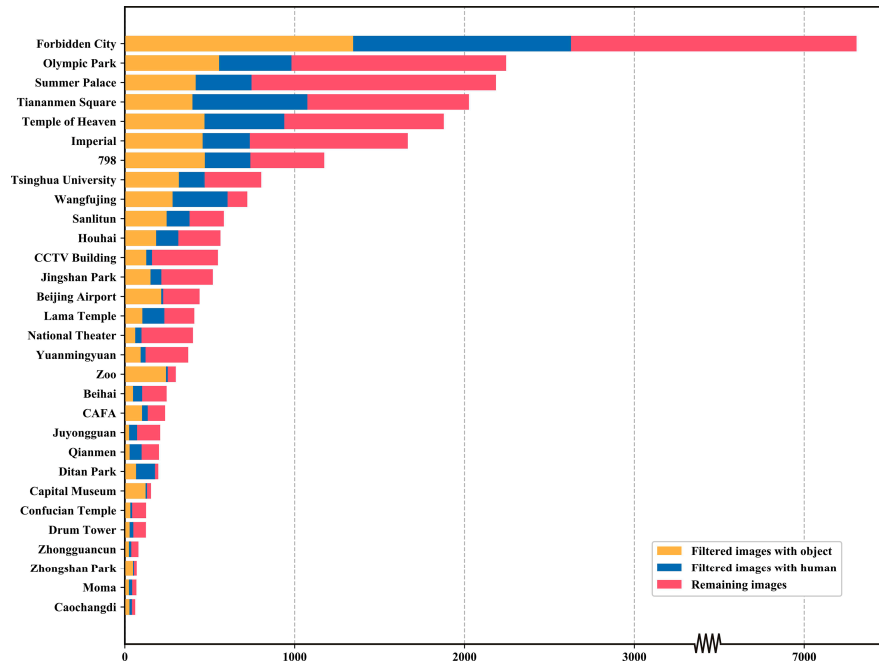


Figure 10. The number of filtered images and the remaining images of each tourist attraction.

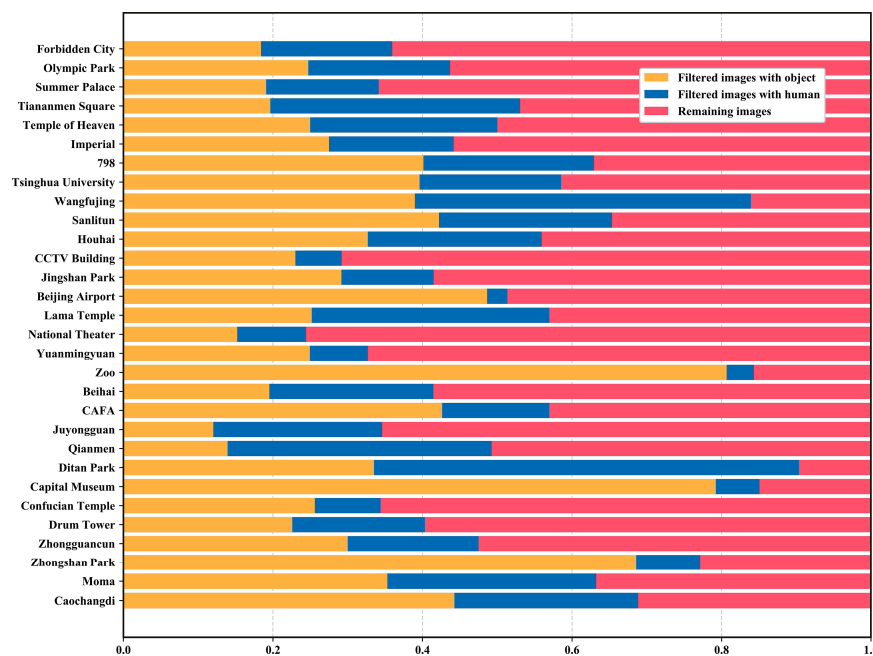


Figure 11. The number of filtered images and the remaining images of each tourist attraction.

We select the top five tourist attractions which have the most number of photos and analyze their results of representative image selection. We compare the result of our representative image selection framework with that of random selection (without noise image filtering process), shown in Figure 12. We can infer from the random selection result that the image set biased- or non-representative photo perspectives (2-a and 5-c in Figure 12 for instance), local parts of this tourist attraction (1-b and 3-b

in Figure 12) and even some noisy images (1-a and 5-b in Figure 12). In addition, although the noise image filtering process has been done, some unrelated images still exist, which increases the difficulty of ranking and selection for the deep ranking model. Our framework can still select the images taken from the most common and representative angles of view with the overall look of a certain tourist attraction. Although some representative images show the visual diversity, they reflect the diverse visual preferences of different users to some extent. For instance, different from the other representative images, 1-f and 3-f in Figure 12 show one of the Palaces in the Forbidden City and Marble Boat in the Summer Palace, respectively.































	Result of random selection			Result of our deep learning framework		
Forbidden City	 (1-a)	 (1-b)	 (1-c)	 (1-d)	 (1-e)	 (1-f)
Tiananmen	 (2-a)	 (2-b)	 (2-c)	 (2-d)	 (2-e)	 (2-f)
Summer Palace	 (3-a)	 (3-b)	 (3-c)	 (3-d)	 (3-e)	 (3-f)
Temple of Heaven	 (4-a)	 (4-b)	 (4-c)	 (4-d)	 (4-e)	 (4-f)
Olympic Park	 (5-a)	 (5-b)	 (5-c)	 (5-d)	 (5-e)	 (5-f)

Figure 12. Representative image selection results of random selection and our framework (The photos in Figure 12 are all licensed under the Creative Commons Attribution License).

4.5. Result of Users' Satisfaction

For better evaluating the overall framework results, we conducted a questionnaire based on the simple tourist map we created, where Baidu Map is the base map and extracted tourist attractions' locations, and representative images are shown (Figure 13). Eighty volunteers participated in the survey, including people who lived in Beijing, have visited Beijing before, or are potential tourists to Beijing in the future (Note that most tourist attractions in Beijing are famous enough, and therefore, most people in China are familiar with them to a different degree). According to the tourist map, each volunteer rated three items based on a Likert scale from 1 (strongly disagree) to 5 (strongly agree), including: (1) Integrity: To what extent do you think the extracted results can cover Beijing's famous tourist attractions (Q1); (2) representativeness: To what extent do you think the selected images represent the tourist attractions (Q2); (3) attractiveness: To what extent do you think adding representative images can attract you more to visit the tourist attractions (Q3).

The statistics results of the questionnaire are shown in Table 4, where the integer number represents the rating number of people for each option. Of all the three criteria, most volunteers chose "agree", and the following are "neutral" or "strong agree". The high average ratings indicate the high users' satisfaction, especially in the criteria of representativeness (about 3.88 out of 5), revealing that the framework of representative image selection is effective. From what has been analyzed above,

the overall framework has the potential to apply in tourism applications and meet the satisfaction of tourists in real life.



Figure 13. The screenshot of the tourist map.

Table 4. Rating results of Q1, Q2, and Q3 in the questionnaire.

Criteria	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	Mean
integrity	5	7	14	41	13	3.63
representativeness	3	3	9	51	14	3.88
attractiveness	5	4	11	49	11	3.71

5. Conclusions

In this paper, we propose a framework containing an improved cluster method and multiple neural network models to extract representative images of tourist attractions. Leveraging Flickr Creative Commons 100 Million Dataset, we choose Beijing as the study area to evaluate our framework. Then we filter the dataset with an entropy-based method to remove certain photos uploaded by natives. We improve a density-based cluster by adding the constraint of time and user number threshold (TU-DJ-Cluster) to extract place-relevant tags, and further merge and extend them according to the spatial relationship between convex hulls generated by these place-relevant tags and semantic similarity between tag embeddings obtained from Word2Vec training. By comparing the extraction result of DBSCAN, TU-DJ-Cluster extracts place-relevant tags and simultaneously filters unimportant tags unrelated to tourist attractions. Besides, the clustering results of our framework are superior to P-DBSCAN, whether in the number of clusters or the accuracy of clustering boundaries. After that, we further select representative images for each tourist attraction, by first filtering noise images with pre-trained MLP and SSD model and then ranking the remaining images with the deep ranking model. The comparative analysis further demonstrates the effectiveness of filtering irrelevant images and selecting representative images of this framework. A questionnaire is also conducted to evaluate users' satisfaction with the overall results. The high rating scores indicate that the results of our framework are effective in extracting tourist attractions and can meet real-life tourists' requirements.

Though the results are satisfactory, some efforts still should be made to improve our framework. For instance, even though noise images are filtered prior to their importance ranking, some thematically-unrelated images remain. This influences the ranking results, due to the diverse visual preferences of different users. Besides, the deep ranking model used in this paper calculates the similarity from embeddings of the whole images, while using convolutional models based on point

detector and descriptor may provide a more accurate selection result, due to the difficulty to select mostly outdoor scene images from noisy geotagged images. In future work, we will attempt to extract places of interest directly from photos or videos with unsupervised or semi-supervised deep learning methods. Furthermore, we try to analyze the visual contents of images taken from tourists to infer their preferences and apply in further applications, such as recommendation systems.

Supplementary Materials: Supplementary materials are available online at <http://www.mdpi.com/2220-9964/9/2/81/s1>.

Author Contributions: Conceptualization, Qingyun Du and Shanshan Han; Data curation, Shanshan Han; Formal analysis, Shanshan Han; Methodology, Shanshan Han; Software, Shanshan Han and Dawei Gui; Supervision, Fu Ren and Qingyun Du; Visualization, Shanshan Han and Dawei Gui; Writing—Original Draft Preparation, Shanshan Han; Writing—Review and Editing, Shanshan Han, Dawei Gui, Qingyun Du and Fu Ren. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2016YFC0803106, and the National Natural Science Foundation of China, grant number No. 41571438.

Acknowledgments: We thank YFCC100M for licensing the image dataset under a Creative Commons Attribution, and some images in these figures were clipped according to cartographic needs.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UNTWO. UNTWO Annual Report 2017. Available online: <https://www.unwto.org/global/publication/unwto-annual-report-2017> (accessed on 16 January 2020).
2. Wang, S.; Wang, Y.; Tang, J.; Shu, K.; Ranganath, S.; Liu, H. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 391–400.
3. Chen, W.-C.; Battestini, A.; Gelfand, N.; Setlur, V. Visual summaries of popular landmarks from community photo collections. In Proceedings of the 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–4 November 2009; pp. 1248–1255.
4. Kozaki, Y.; Wang, Y.; Kawai, Y. Generating Pictorial Maps for Tourists using Flickr Photo Data. In Proceedings of the 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 9–12 October 2018; pp. 403–407.
5. Zhang, Z.; Zou, C.; Ding, R.; Chen, Z. VCG: Exploiting visual contents and geographical influence for Point-of-Interest recommendation. *Neurocomputing* **2019**, *357*, 53–65. [CrossRef]
6. Zhou, X.; Xu, C.; Kimmons, B. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Comput. Environ. Urban Syst.* **2015**, *54*, 144–153. [CrossRef]
7. Flickr. Work at Flickr. Available online: <https://www.flickr.com/jobs/> (accessed on 26 December 2019).
8. Hu, Y.; Gao, S.; Janowicz, K.; Yu, B.; Li, W.; Prasad, S. Extracting and understanding urban areas of interest using geotagged photos. *Comput. Environ. Urban Syst.* **2015**, *54*, 240–254. [CrossRef]
9. Weyand, T.; Kostrikov, I.; Philbin, J. Planet-photo geolocation with convolutional neural networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 37–55.
10. Da Cunha, K.B.; Maggi, L.; Teichrieb, V.; Lima, J.P.; Quintino, J.P.; da Silva, F.Q.; Santos, A.L.; Pinho, H. Patch PlaNet: Landmark Recognition with Patch Classification Using Convolutional Neural Networks. In Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 29 October–1 November 2018; pp. 126–133.
11. Majid, A.; Chen, L.; Chen, G.; Mirza, H.; Hussain, I.; Woodward, J. A Context-aware Personalized Travel Recommendation System Based on Geotagged Social Media Data Mining. *Int. J. Geogr. Inf. Sci.* **2012**, *27*, 1–23. [CrossRef]
12. Cai, G.; Lee, K.; Lee, I. Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos. *Expert Syst. Appl.* **2018**, *94*, 32–40. [CrossRef]
13. Xia, P.; Zhou, H. A Novel Popular Tourist Attraction Discovering Approach Based on Geo-Tagged Social Media Big Data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 216.

14. Kennedy, L.; Naaman, M.; Ahern, S.; Nair, R.; Rattenbury, T. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In Proceedings of the 15th ACM international conference on Multimedia, Augsburg, Germany, 25–29 September 2007; pp. 631–640.
15. Abbasi, R.; Chernov, S.; Nejdil, W.; Paiu, R.; Staab, S. Exploiting Flickr Tags and Groups for Finding Landmark Photos. In Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Toulouse, France, 6–9 April 2009; pp. 654–661.
16. Gao, Y.; Tang, J.; Hong, R.; Dai, Q.; Chua, T.-S.; Jain, R. W2Go: A travel guidance system by automatic landmark ranking. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 123–132.
17. Luo, J.; Joshi, D.; Yu, J.; Gallagher, A. Geotagging in multimedia and computer vision—A survey. *Multimed. Tools Appl.* **2011**, *51*, 187–211. [[CrossRef](#)]
18. Liang, C.-K.; Hsieh, Y.-T.; Chuang, T.-J.; Wang, Y.; Weng, M.-F.; Chuang, Y.-Y. Learning landmarks by exploiting social media. In Proceedings of the 16th international conference on Advances in Multimedia Modeling, Chongqing, China, 6–8 January 2010; pp. 207–217.
19. Wikipedia. Tourist Attraction. Available online: https://en.wikipedia.org/wiki/Tourist_attraction (accessed on 16 January 2020).
20. Zhou, C.; Frankowski, D.; Ludford, P.; Shekhar, S.; Terveen, L. Discovering personal gazetteers: An interactive clustering approach. In Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems, Washington, DC, USA, 12–13 November 2004; pp. 266–273.
21. Li, Q.; Li, S.; Zhang, S.; Hu, J.; Hu, J. A Review of Text Corpus-Based Tourism Big Data Mining. *Appl. Sci.* **2019**, *9*, 3300. [[CrossRef](#)]
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
23. Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; Wu, Y. Learning Fine-Grained Image Similarity with Deep Ranking. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1386–1393.
24. Sun, Y.; Fan, H.; Bakillah, M.; Zipf, A. Road-based travel recommendation using geo-tagged images. *Comput. Environ. Urban Syst.* **2015**, *53*, 110–122. [[CrossRef](#)]
25. Kisilevich, S.; Mansmann, F.; Keim, D. P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, Washington, DC, USA, 21–23 June 2010; p. 38.
26. Vu, H.Q.; Gang, L.; Law, R.; Ye, B.H. Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tour. Manag.* **2015**, *46*, 222–232. [[CrossRef](#)]
27. McKenzie, G.; Janowicz, K. Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Comput. Environ. Urban Syst.* **2015**, *54*, 1–13. [[CrossRef](#)]
28. Lin, W.; Hong, Y.; Zhou, H.; Xia, P.; Ran, L. A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 2225–2246.
29. Cao, L.; Luo, J.; Gallagher, A.; Jin, X.; Han, J.; Huang, T.S. A worldwide tourism recommendation system based on geotagged web photos. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 2274–2277.
30. Kennedy, L.S.; Naaman, M. Generating diverse and representative image search results for landmarks. In Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 297–306.
31. Crandall, D.J.; Backstrom, L.; Huttenlocher, D.; Kleinberg, J. Mapping the world’s photos. In Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, 20–24 April 2009; pp. 761–770.
32. Samany, N.N. Automatic landmark extraction from geo-tagged social media photos using deep neural network. *Cities* **2019**, *93*, 1–12. [[CrossRef](#)]
33. Kim, D.; Kang, Y.; Park, Y.; Kim, N.; Lee, J.; Cho, N. Analysis of Tourists’ Image of Seoul with Geotagged Photos using Convolutional Neural Networks. In Proceedings of the ICA, Aachen, Germany, 9–13 September 2019.

34. Crandall, D.J.; Li, Y.; Lee, S.; Huttenlocher, D.P. Recognizing landmarks in large-scale social image collections. In *Large-Scale Visual Geo-Localization*; Springer: Berlin, Germany, 2016; pp. 121–144.
35. Ji, R.; Duan, L.; Chen, J.; Yang, S.; Yao, H.; Huang, T.; Gao, W. Learning the trip suggestion from landmark photos on the web. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2485–2488.
36. Kawakubo, H.; Yanai, K. Geovisualrank: A ranking method of geotagged images considering visual similarity and geo-location proximity. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 69–70.
37. Ma, X.; Zhao, Y.; Qian, X.; Tang, Y.Y. Multi-source fusion based geo-tagging for web images. *Multimed. Tools Appl.* **2018**, *77*, 16399–16417. [[CrossRef](#)]
38. Ding, X.; Fan, H. Exploring the Distribution Patterns of Flickr Photos. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 418. [[CrossRef](#)]
39. Zhang, F.; Zhou, B.; Ratti, C.; Liu, Y. Discovering place-informative scenes and objects using social media photos. *Roy. Soc. Open Sci.* **2019**, *6*, 181375. [[CrossRef](#)]
40. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.-J. YFCC100M: The new data in multimedia research. *arXiv* **2015**, arXiv:1503.01817. [[CrossRef](#)]
41. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
42. Lee, S.S.; Won, D.; McLeod, D. Tag-geotag correlation in social networks. In Proceedings of the 2008 ACM Workshop on Search in Social Media, Napa Valley, CA, USA, 30 October 2008; pp. 59–66.
43. Fei-Fei, L.; Fergus, R.; Perona, P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Und.* **2007**, *106*, 59–70. [[CrossRef](#)]
44. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 Million Image Database for Scene Recognition. *IEEE T. Pattern Anal.* **2018**, *40*, 1452–1464. [[CrossRef](#)]
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Yao, L.; Miller, J. Tiny imagenet classification with convolutional neural networks. *CS 231N* **2015**, *2*, 8.
48. Wikivoyage. Beijing. Available online: <https://en.wikivoyage.org/wiki/Beijing#Q956> (accessed on 20 October 2019).



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).