Article

Unsupervised Visual Representation Learning for Indoor Scenes with a Siamese ConvNet and Graph Constraints

Mengyun Liu [1], Ruizhi Chen [1,2,*], Haojun Ai [3,*], Yujin Chen [1] and Deren Li [1,2]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China; amylmy@whu.edu.cn (M.L.); yujin.chen@whu.edu.cn (Y.C.); drli@whu.edu.cn (D.L.)

[2] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

[3] School of Computer Science, Wuhan University, Wuhan 430072, China;

[*] Correspondence: ruizhi.chen@whu.edu.cn (R.C.); aihj@whu.edu.cn (H.A.); Tel.: +027-8773-1869-802 (R.C.)

Abstract: Indoor scene recognition has great significance for intelligent applications such as mobile robots, location-based services (LBS) and so on. Wherever we are or whatever we do, we are under a specific scene. The human brain can easily discern a scene with a quick glance. However, for a machine to achieve this purpose, on one hand, it often requires plenty of well-annotated data which is time-consuming and labor-intensive. On the other hand, it is hard to learn effective visual representations due to large intra-category variation and inter-categories similarity of indoor scenes. To solve these problems, in this paper, we adopted an unsupervised visual representation learning method which can learn from unlabeled data with a Siamese Convolutional Neural Network (Siamese ConvNet) and graph-based constraints. Specifically, we first mined relationships between unlabeled samples with a graph structure. And then, these relationships can be used as supervision for representation learning with a Siamese network. In this method, firstly, a *k*-NN graph would be constructed by taking each image as a node in the graph and its *k* nearest neighbors are linked to form the edges. Then, with this graph, cycle consistency and geodesic distance would be considered as criteria for positive and negative pairs mining respectively. In other words, by detecting cycles in the graph, images with large differences but in the same cycle can be considered as same category (positive pairs). By computing geodesic distance instead of Euclidean distance from one node to another, two nodes with large geodesic distance can be regarded as in different categories (negative pairs). After that, visual representations of indoor scenes can be learned by a Siamese network in an unsupervised manner with the mined pairs as inputs. In order to evaluate the proposed method, we tested it on two scene-centric datasets, *MIT67* and *Places365*. Experiments with different number of categories have been conducted to excavate the potential of proposed method. The results demonstrated that semantic visual representations for indoor scenes can be learned in this unsupervised manner. In addition, with the learned visual representations, indoor scene recognition models trained with the learned representations and a few of labeled samples can achieve competitive performance compared to the state-of-the-art approaches.

Keywords: indoor scene recognition; unsupervised representation learning; Siamese network; graph constraints

1. Introduction

Scene recognition is a well-known task in computer vision field. It can be divided as outdoor scenario and indoor scenario. Comparing to outdoor scenario, indoor scene recognition is far more difficult due to the diversity of intra-categories and similarity of inter-categories [1]. Despite of the challenges in indoor scene recognition, it is still extremely important since recognizing an indoor scene efficiently and appropriately is a significant perception ability for indoor-based applications such as indoor pedestrian localization / navigation [2-6], indoor mobile robot [7-9] and human activity analysis [10-12], etc. In recent years, deep convolutional neural networks (DCNNs) have achieved vast success on different computer vision tasks including scene recognition. Even there are no indoor scene recognition model trained from scratch with labeled image data, this task still benefits a lot from pre-trained models in a transfer learning way [13,14]. A trained model on ImageNet with DCNNs [15-17] still can act as a feature extractor in an indoor scene recognition task. Or by fine-tuning, the weights of the trained model can be adapt to suit for the goal task.

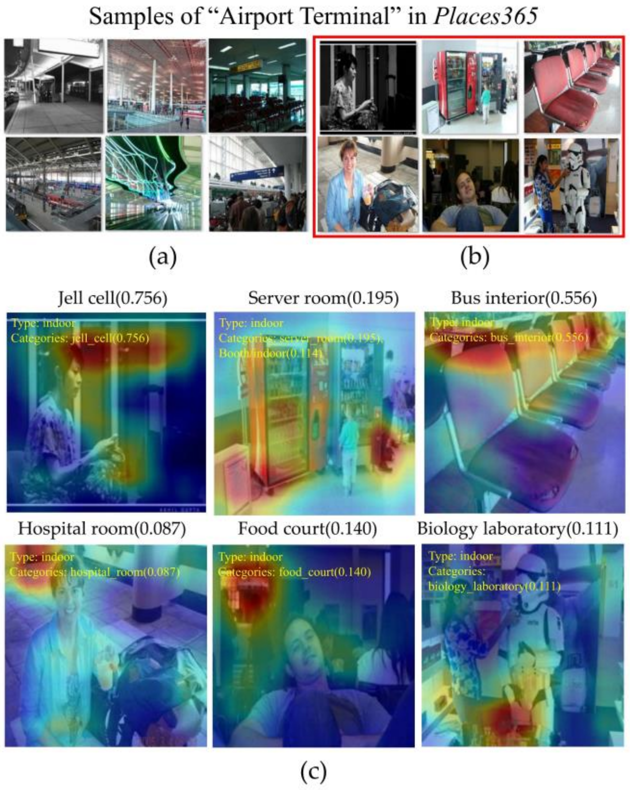

Samples of "Airport Terminal" in *Places365*
(a)
(b)
Jell cell(0.756)
Type: indoor
Categories: jell_cell(0.756)
Server room(0.195)
Type: indoor
Categories: server_room(0.195), Booth/indoor(0.114)
Bus interior(0.556)
Type: indoor
Categories: bus_interior(0.556)
Hospital room(0.087)
Type: indoor
Categories: hospital_room(0.087)
Food court(0.140)
Type: indoor
Categories: food_court(0.140)
Biology laboratory(0.111)
Type: indoor
Categories: biology_laboratory(0.111)
(c)

Figure 1. Comparison of training samples which labeled as "airport terminal" in *Places365*. In this figure, (a) demonstrates samples with good annotations and (b) illustrates samples that are hard to be confirmed. (c) is recognition results and class activation map (CAM) [25] of samples in (b). None of them have a right prediction with the trained model.

However, most of previous practices rely on large-scale and well-annotated datasets. Even image data can be obtained easily with web-crawling, the labeling process is extremely time-consuming and labor-intensive. Datasets such as *ImageNet* [15] and *Places365* [19] used a crowdsourcing platform named Amazon Mechanical Turk (AMT) [20] which requires manual verification and the quality of annotations cannot be guaranteed. As shown in Figure 1, although all of these images belong to "airport terminal" in *Places365*, it is still arduously to discriminate that they are in same category even with human brain. With the trained model provided by [19,21], Figure 1(c) has illustrated the recognition results and class activation map (CAM) [25] of poor samples in Figure 1(b). It is obvious that none of them give a correct recognition result. Although, effective features for indoor scene recognition can be learned through hierarchical nonlinear mappings with DCNNs. It is still hard to apply features extracted from unlabeled data to a recognition model directly.

The goal of representation learning is to transform the input data into representations that are more suitable for a given task [22]. Therefore, in this paper, we thought about learning the visual representations from unlabeled data to facilitate the indoor scene recognition task. The main idea of this method is to mine the relationships between unlabeled data and find positive pairs and negative pairs with a graph structure. Then, these learned verification signals can be used to learn visual representations which will facilitate the target task. The unsupervised representation learning process in this work also can be taken as a pre-training stage which adopted by many deep neural network training practices. The difference between this work and others is that our goal is to consider more about semantic-level relationships between indoor scene images. To this end, we adopted a graph structure and then using constraints of this graph to mine relationships between unlabeled data. In this method, each node in the graph represents an image in dataset, and each edge denotes a high similarity of two images. The weight of an edge is determined by the value of similarity. In this paper, Euclidean distance or L_2 distance is utilized to measure the similarity. The reason of using a graph structure instead of others is that images in same category with large difference can be distinguished by cycle detection in the graph, while images in different categories with small L_2 distance can be figured out by computing the weighted path length between them. Therefore, the proposed method can effectively overcome the large intra-categories diversity and small inter-categories diversity problems of indoor scenes.

After mining the relationships between unlabeled data and obtaining positive and negative pairs, a Siamese network [23] are taken into consideration for following training. A Siamese network often contains two identical sub-networks (like twins) which share the weights among them. Different structures of sub-networks can be chosen to form a Siamese network. However, our goal is to learn visual representations. For this purpose, a CNN structure is adopted in this work. The advantage of using this architecture is that the goal of a Siamese network is to distinguish whether two inputs are of the same class or not. It will give results about how similar or different between the two inputs. This matched well with the positive and negative pairs we mined in previous procedures, since the mined pairs can be used as supervisions in training process. After training the Siamese network, a binary classification model can be acquired to measure the similarity of two inputs. Meanwhile, effective data descriptors or visual representations are learned which can be used for further applications such as indoor scene recognition. Taking the trained model as a feature extractor and followed by a simple classifier can train an effective indoor scene recognition model.

As described above, the main contributions of this paper can be concluded as follows.

(1) Supervisory signals, which are positive and negative pairs, are mined from unlabeled samples with graph constraints in this work, which can effectively solve the intra-diversity and inter-similarity problems in indoor circumstances.

(2) A Siamese ConvNet is adopted to learn visual representations in an unsupervised manner, which reduces the requirements for large amounts of well-annotated data compare to other deep learning practices.

(3) With visual representations learned from unlabeled data, we show how indoor scene based applications can benefit from these. We evaluate the performance of indoor scene recognition model which trained with the learned representations and a few of labeled data.

(4) Larger indoor scene datasets extracted from two scene-centric datasets have been adopted to investigate properties of indoor scenes. While, other works only consider indoor scene problems in a relatively smaller dataset such as MIT67.

The rest of this paper is organized as follows. Related work of indoor scenes and unsupervised visual representation learning are reviewed in Section 2. Overview and methodology are given in Section 3, which including graph construction, positive / negative pairs mining and Siamese network training. Experiments and results are presented in Section 4. Section 5 discusses the results and is followed by Section 6 to conclude the whole work.

2. Related Work

Scene recognition is a valuable research topic and recent years have witnessed a rapid growth of related works on it. Before deep learning technique became prevailing in computer vision field, most of these works are based on handcraft features. For instance, [26] proposed to use GIST descriptors to recognize over 60 places both in indoor and outdoor scenarios. While [27] adopted probability density response maps (PDRMs) and bagged LDA classifier to build a model for 10-scenes recognition. [28] designed a color descriptors for images and [29] used objects as scene attributes directly. In addition, [30] use a multispectral SIFT to classify scenes with RGB and near-infrared images. [31] proposed a latent variable model for scene recognition which represented a scene as a collection of regions while [32] proposed a discriminative latent topic model for scene recognition based on spatial layout and scene elements. [33] proposed a SPMSM model for scene recognition which was augmented an image representation on semantic probability simplex with a rough encoding of spatial information. [34] also has developed an image representation for scene recognition with the response maps of objects part filters. From these works, it can be found that local and global feature descriptors have been designed to learn scene representations. However, for one hand, they still cannot achieve fair performance in indoor scenes. For another hand, they need full-supervisions in training process, which are unable to make good use of large amount unlabeled data acquired from internet and our daily life. Thus, it is necessary to design an unsupervised representation learning method for indoor scenes. Hence, in this paper, we focus on learning semantic representations for indoor scenes in an unsupervised manner. Previous research related to the proposed method can be divided into two parts. Indoor scene recognition and unsupervised representation learning.

For indoor scene recognition, one of the most representative work is the creation of indoor dataset *MIT67* [35]. It contains 67 categories in 5 big scene groups including store, home, public spaces, leisure and working place. In following research of indoor scene recognition, this dataset has been widely adopted as a benchmark for it is well annotated and contains a wide range of indoor categories. They also proposed a method to combine GIST descriptors and ROIs (Regions of Interest) to solve the recognition problem which has considered the mechanism of how human beings recognize indoor scenes. However, this method required manually annotated positions of ROIs and need to create a visual vocabulary in advance. In later research, Kaerwong et.al. [36] considered a different situation which scene images were gradually obtained during long-term operation. They designed an incremental learning framework which based on n-value self-organizing and incremental neural network (n-SOINN). Other works including [37-40] also aimed to exploit real world attributions of indoor scenes and apply them to classify cluttered indoor scenes. [37] proposed a method using common objects as an intermediate semantic representation. In this work, they enhanced the performance of indoor scene recognition with contextual relations of objects in a scene. [38] adopted a BoW(Bag-of-Words) scheme to learn representations and then utilize Nearest-Neighbor (NN) classifiers based on metric functions. [39] used dense-SIFT descriptors and an encoding method which combine saliency-driven perceptual pooling with simple spatial pooling. And [40] adopted weighted hypergraph to represent the connectivity among images according to statistics of objects appearing in the same image. However, all these works are based on handcraft features. As deep networks become an overwhelming technique in the field of computer vision, automatically feature extraction methods with CNN models were adopted by more and more researchers. It is an end-to-end representation learning method for many vision task. Thus, methods with deep features for indoor scene tasks became prevalent [21,41-42]. Among them, Zhou et al. [21] trained deep models with a large-scale scene dataset which including both indoor and outdoor scenes. They demonstrated that they can achieve higher accuracy with Places-CNN features (68.24%) compared to ImageNet-CNN features (56.79%) on *MIT67* dataset. Khan et al. [41] also applied CNN features to categorize indoor scenes. The difference between their work and others is they proposed a method to encode the features into a number of multiple codebooks to overcome large variations in scene layouts. Although they achieved an accuracy with 71.8%, large amount of semantically labeled elements were needed in this work. In [42], CNN features and a sparse coding method was adopted to recognize indoor scenes. They replaced the traditional feature extraction method with CNNs and get better performance on *MIT67* dataset. Although these works show a great progress on indoor scene recognition problems, they still need a large amount of human-labeled data. Whereas, in this paper, we focus on learning visual representations for indoor scenes in an unsupervised way. Based on previous work, we combined the local feature inction methods with a graph structure to mine latent relationships between unlabeled samples. Then a Siamese network [24] was adopted for visual representation learning with the mined formation.

For unsupervised representation learning, related work can be found in [45-54]. The goal of representation learning is to reconstruct distribution of input data to facilitate following learning processes [44]. In other words, good representations can make a task easier and vice versa. Earlier works can be traced back to greedy layerwise unsupervised pre-training with Restricted Boltzmann Machines (RBM) [45] and autoencoders [46]. The pretraining procedure acted as an initial learning step which overcame the difficulties when learning a deep network. After that, Vincent et al. [47] enhanced these works by making the learned representations robust to partial corruption of the input pattern. Different from them, Bosch et al. [48] proposed a latent generative model while Srivastava et al. [49] extended representation learning to multi-modal inputs. Nevertheless, these works relied on a single representation learning method may only learn the low-level features of inputs. To learn high-level features, Le et al. [50] adopted a locally connected sparse autoencoder network to learn from large-scale unlabeled data and showed that class-specific feature detectors can be learned from this method. Other works related to visual representation learning including [51-56]. They advocated that learning visual representations with very few labeled data or unlabeled data like humans. Jeff et al. [51] investigated and visualized the semantic clustering of deep convolutional features of different tasks and proposed a method to transfer visual representations from one task to a related one with few training samples. [52] learned visual representations with unlabeled videos from web. Their idea was to utilize relationships of two patches in adjacent frames and use a Siamese-triplet network to train CNN representations. While [53] learned visual representations with an algorithm driven by context-based pixel prediction and proposed context encoders to generate the contents of an arbitrary image region conditioned on its surroundings. [54] also adopted spatial context as supervisory signal to learn visual representations. They extracted random pairs of patches from each image and train a CNN to predict relative position of the patches. [55] proposed graph-based consistent constraints to learn the visual representation in an unsupervised manner with mining positive and negative image pairs. However, this work put more emphasis on simple object-centric datasets. Fabio et al. [56] discussed the invariant representations from unsupervised learning and sample complexity related to good representations. Furthermore, unsupervised visual representation learning also was widely used in remote sensing area for image classification or object detection [57-60]. Cheriyadat et al. [57] adopted sparse encoding method for dense low-level feature descriptors, and then learned the visual representations with statistics of the sparse features. Zhang et al. [58] proposed a learning method with saliency detection. While Hu et al. [59] presented an improved unsupervised feature learning method with spectral clustering. In [60], Tao et al. learned the feature representations adaptively from unlabeled data with stacked sparse autoencoder.

Previous works have laid a solid foundation for the proposed method in both theory and practice. Comparing to them, firstly, we focus on indoor scene problems instead of objects, which is extremely challenging but meaningful. Secondly, to exploit the real-world pattern, we not only evaluated the proposed method on *MIT67* but also extracted the indoor part of *Places365* which has far more samples and categories then *MIT67*. Thirdly, we adopted a graph structure to organize the unlabeled sample and then mined positive pairs and negative pairs for Siamese network training to learn the indoor scene representations.

3. Overview and Methodology

In this section, an overview contains key procedures of the proposed method will be illustrated first. Then, details and related methods of these procedures will be described comprehensively. There are three parts in methodology, including graph construction, positive and negative pairs mining and Siamese network training.

3.1. Overview

An overview of proposed method has been illustrated in Figure 2. The main target of this work is to learn indoor scene representations in an unsupervised way. To achieve that, mining supervisory information from unlabeled images is demanded. Thus, in this paper, a graph structure is adopted to represent unlabeled data for mining the relationships between them. This process consists of three parts. The first is graph construction. The second is mining the positive pairs of images (two images in same category), and the third is mining the negative pairs (two images in different categories).

After that, these pairs will be utilized as inputs for a Siamese network training. The training goal of a Siamese network is to discriminate whether two images are in same categories or not. The whole learning process does not need any image labels of categories but do need labels which indicate whether two inputs are in same class or not.

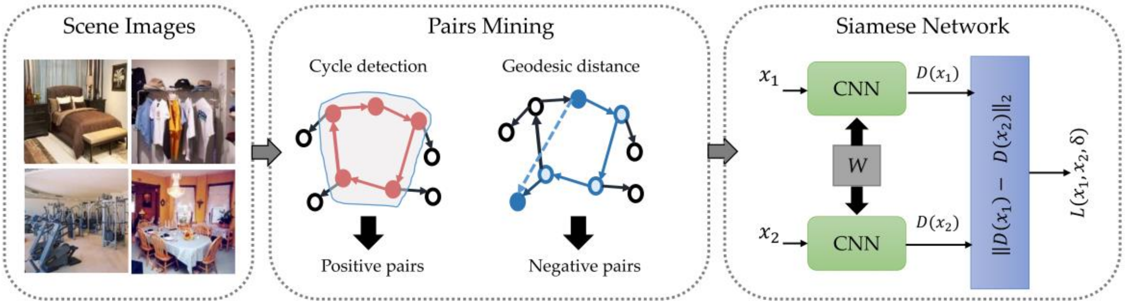


Figure 2. Overview of proposed method. The most import module of this method is pairs mining and Siamese network training. To mine positive and negative pairs, a *k*-NN graph need to be built first. The outputs of pairs mining is the inputs of Siamese network training.

3.2. Graph Construction

To mine information from unlabeled data, a *k*-NN graph is adopted. In this graph, each node represents an image and the edges formed by connections of their *k* nearest neighbors. The weights of edges are determined by Euclidean distance of original features extracted from images. Thus, in order to form the graph, original feature extraction methods will be briefly described first. Then followed by an illustration about the graph structure.

3.2.1. Original Feature Extraction

The whole procedure for original feature extraction has been depicted in Figure 3. It includes two steps, local feature extraction and feature encoding. Whereas, there are a wide range of methods to accomplish both of these two steps.

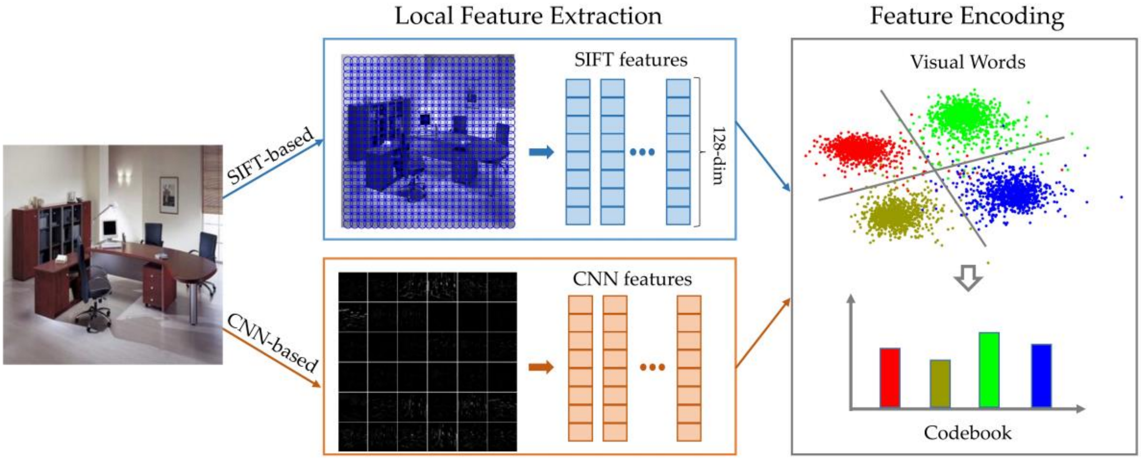


Figure 3. Procedure of local feature extraction and encoding for computing original features of images. SIFT-based method and CNN-based method are given for comparison. In practice, one of these two methods will be adopted for local feature extraction.

For local feature extraction, as discussed in [64], the most prevailing methods in recent years can be divided as SIFT-based and CNN-based.

- **SIFT-based features**: SIFT is a widely used local feature descriptor for vision tasks such as image matching. A standard version of SIFT descriptor is 128-dim. Dense SIFT is proposed to simplify the computation of SIFT. Instead of computing SIFT descriptors at detected key points, dense SIFT computes SIFT with dense patches. With such a computing process, dense SIFT can be faster than SIFT.

- **CNN-based features**: CNN features are intermediate outputs of pre-trained convolutional neural networks including outputs from both fully-connected layers (FC-layers) and convolutional layers (Conv-layers). Taking AlexNet [16] as an example, features from FC6 or FC7 is 4096-dim. Since FC-layers have a global receptive field, it often can be taken as global features. Whereas, outputs of Conv-layers are computed by convolutional filters which have smaller receptive fields but densely applied on the whole image. Thus, feature maps are produced by a Conv-layer can be seen as dense local features.

For feature encoding, it is used to to aggregate dense local features to global features to represent an image. Classical coding methods include Bag of Words (BoW) [61], Fisher Vector (FV) [62], Vector of Locally Aggregated Descriptors (VLAD) [63] and so on.

- **BoW**: The main idea of BoW is to create a codebook of visual words, which records number of occurrences of these visual words, but not positions. The visual words can be obtained by clustering method such as *k*-Means, which the center of a cluster is a visual word and *k* is the number of visual words.
- **FV**: Different from BoW, FV is a statistics capturing the distribution of local descriptors. Suppose that the local descriptors are independent and identically distributed, FV of an image can be computed as Equation (1), which stands for the sum of normalized gradient statistic of local descriptors.

$$I_{FV} = \sum_{t=1}^{T} F_\lambda^{-\frac{1}{2}} G_\lambda^X = \sum_{t=1}^{T} F_\lambda^{-1/2} \nabla_\lambda log u_\lambda(X) = \sum_{t=1}^{T} F_\lambda^{-1/2} \nabla_\lambda log u_\lambda(x_t) \tag{1}$$

Among them, $\mathrm{X} = \{x_t, \mathrm{t} = 1, \dots \mathrm{T}\}$ is a sample contains T descriptors. $G_\lambda^X = \nabla_\lambda log u_\lambda(x_t)$ is the gradient vector of likelihood function, and F_λ denotes the Fisher information matrix, which can be computed as $F_\lambda = E_{x \sim u_\lambda}[G_\lambda^X G_\lambda^{X'}]$. When the distribution obeys a Gaussian Mixture Model (GMM) and the parameters are $\lambda = \{\omega_k,\ \mu_k, \Sigma_k, \mathrm{k} = 1, \dots, \mathrm{K}\}$, $u_\lambda(x) = \mathrm{P}(x|\lambda)$ can be computed as Equation (2).

$$u_\lambda(x) = \sum_{k=1}^{K} \omega_k \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)' \Sigma_k^{-1} (x - \mu_k)\right\} \tag{2}$$

- **VLAD**: VLAD can be seen as a simplified version of FV. Similar to BoW, it quantized local descriptors to its nearest visual word, and then records cumulative residuals of all descriptors with this visual word. Suppose set $\{\mu_1,\ \mu_2,\ \dots, \mu_K\}$ of centroids is learned with k-Means, VLAD descriptor V from a set of descriptors $\{x_1, x_2, \dots, x_T\}$ can be computed as follows.

$$i = \arg min_j \left\| x_t - \mu_j \right\| \tag{3}$$

$$v_i := v_i + x_t - \mu_i \tag{4}$$

$$V = [{v_1}^T, \dots, {v_K}^T] \tag{5}$$

In a nutshell, the main idea of these three encoding methods is to create codebooks for visual words. BoW encodes the 0-order statistics of the descriptors. FV extends the BoW by encoding high-order statistics and VLAD is a simplified version of FV.

3.2.2. Graph Structure

As mentioned in previous part, a *k*-NN graph has been adopted in this paper to represent relationships of unlabeled data. Supposed that a *k*-NN graph is denoted as $G = (V, E)$, then each node in this graph represents an image sample and can be denoted as $v \in V = \{I_1, I_2, \dots, I_N\}$. Each edge represents a link between images and $I_i \rightarrow I_j$ represents I_j is one of *k* nearest neighbors of I_i. Then, the graph structure can be determined by the value of *k* in procedure of finding the *k* nearest neighbors. Here, we use Euclidean distance as a measurement to find the *k* nearest neighbors of a sample and also represent the weight of an edge. For two N-dim vectors $X = \{x_1, x_2, \dots, x_N\}$ and $Y = \{y_1, y_2, \dots, y_N\}$, Euclidean distance is computed as Equation (6). Then we take the top k samples with the smallest distance as nearest neighbors of a sample and connect them with each other.

$$d_{xy} = \sqrt{\sum_{k=1}^{N}(x_k - y_k)^2} \tag{6}$$

An example of the structure for k-NN graph with a 5-length circle can be illustrated as Figure 4(a). In this graph, value of k is set to 4 and $I_1 \sim I_5$ represents image samples. Besides, Figure 4(b) is an example to demonstrate different kinds of distances in the graph, which d_{12}, d_{23}, d_{34}, d_{45}, d_{15} denotes Euclidean distances between two nodes and the weighted path $I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow I_4 \rightarrow I_5$ is geodesic distance. The Euclidean distance and geodesic distance for I_1 to I_5 are d_{15} and $g_{15} = d_{12} + d_{23} + d_{34} + d_{45}$ respectively.

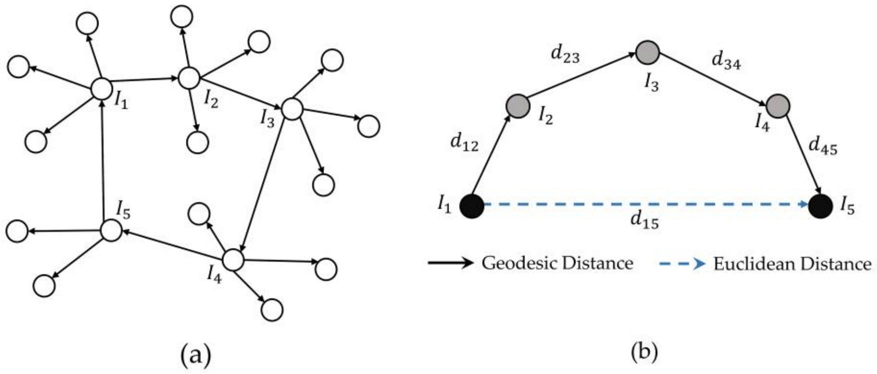


Figure 4. k-NN graph structure. (a) is an example when k = 3, and there is a 5-length circle in this graph. (b) is a demonstration for Geodesic distance and Euclidean distance, which d_{12}, d_{23}, d_{34}, d_{45}, d_{15} are Euclidean distance between two nodes. The weighted path from I_1 to I_5 is geodesic distance between I_1 and I_5.

3.3 Pairs Mining

After organizing samples as a graph, properties of this graph can be applied to mine positive and negative pairs for future training. Based on methods proposed in [55], cycle consistency criterion and geodesic distance are adopted to mine positive and negative pairs in this paper.

3.3.1. Positive Pairs Mining with Cycle Consistency

Positive pairs are images which supposed in same categories. Although clustering or matching algorithms can be used to find similar image samples in an unlabeled dataset, it is still a great challenge in indoor scene circumstances due to large intra-diversity. Therefore, it is important to find a method that not only can mined image pairs with similar appearance, but also pairs in same category but with large variations. In this paper, cycle consistency criterion is adopted to achieve this goal. This process can be depicted as follows.

Supposed that k-nearest neighbors of an image I_x were denoted as $N_k^1(I_x)$. If image $I_y \in N_k^1(I_x)$ and its k-nearest neighbors is $N_k^1(I_y) = \{I_y^i, i = 1, \dots k\}$, then we take $N_k^2(I_x) = N_k^1(I_y)$ as 2-order of k-nearest neighbors of I_x. Therefore, t-order of k-nearest neighbors of I_x can be represented as $N_k^t(I_x)$. Then, I_x can be identified in a t-length cycle if I_x belongs to its own t-order of k-nearest neighbors, which can be represented as follows.

$$I_x \in N_k^t(I_x) \tag{7}$$

Figure 5 has shown examples for detected cycles with 6-length. From this figure, it can be found that images even with large difference (such as I_1 and I_4) still can be matched as positive pairs by detecting cycles in a graph.

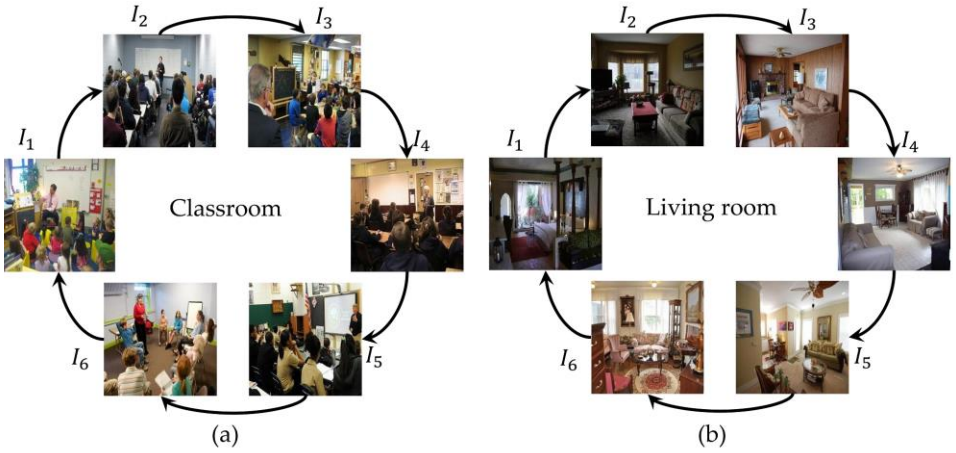


Figure 5. Example of detected cycles for indoor scenes. (a) are samples both from "classroom" and (b) are samples from living room. It is easy to see that even images in same category may have large appearance difference and they can be mined by a cycle detection approach.

3.3.2. Negative Pairs Mining by Geodesic Distance

Negative pairs represent images in different scene categories. Generally, images with large distance are taken as negative pairs. However, for indoor scene images, samples in different categories may have small Euclidean distance in feature space. Especially when two scenes share some of objects they contained (such as computer room and office). This phenomenon is so-called large inter-similarity. It makes these images easy to be omitted in negative mining process. Whereas, to learn an effective visual representations for indoor scenes, images in different classes but have similar appearance are significantly in following training process. To solve this problem, geodesic distance is adopted to mine the negative pairs. Different from Euclidean distance which represents distance between two nodes, geodesic distance is the distance of shortest path between two nodes. As demonstrated in Figure 4(b), given a shortest path from node I_i to I_j, the geodesic distance g_{ij} between them can be computed as sum of edge weights along this path. To detect shortest paths between any two nodes, Dijkstra [67] or Floyd-Warshall Algorithm [43] can be employed.

3.4. Siamese Network Training

After mining positive and negative pairs from unlabeled data, they can be acted as supervisory information to learn indoor scene representations. Since the training goal of a Siamese network is to guess whether the inputs are in same category or not, which is consistent with the mined image pairs. Thus, a Siamese architecture will be used to learn the visual representations in this paper.

3.4.1. Siamese Architecture

The original design of Siamese architecture [23] is shown as Figure 6. It consists of two subnetworks which has same architectures but accept different inputs. The weights W will shared by these two subnetworks. The training goal of a Siamese network is to guess whether the inputs are in same category or not. Let X_1 and X_2 represent two inputs, $y \in \{0, 1\}$ denotes whether they are in same category. If they belong to same category ($y = 1$), the training process will make $||G_w(X_1) - G_w(X_2)||$ small and if the inputs belong to different category ($y = 0$), the training process will make $||G_w(X_1) - G_w(X_2)||$ large in turn.

When use $P(X_1\, o\, X_2)$ denotes the possibility that X_1 and X_2 share the same class. The loss function for general network training can be denoted as Equation 8, where $\lambda\|W\|_2$ is the weight decay term which is used to reduce noise and improve the generalization of model.

$$L(X_1, X_2, y) = y \cdot log\big(P(X_1\, o\, X_2)\big) + (1 - y) \cdot log\big(1 - P(X_1\, o\, X_2)\big) + \lambda\|W\|_2 \tag{8}$$

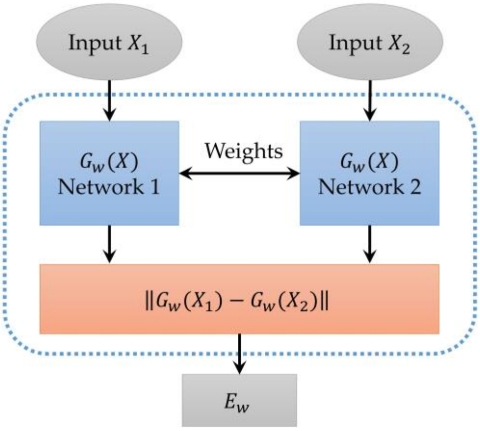


Figure 6. Siamese architecture. X_1 and X_2 are two different inputs. The purpose of this architecture is to detect whether the inputs is in same category or not. The goal of training is to find appropriate parameters W for mapping functions G_w.

3.4.2. Siamese ConvNet

In order to train Siamese network with image data, two convolutional networks (ConvNets) with same architectures are used to be the subnetworks. For simplify, a two-stream AlexNet [16] is adopted to implement the twin networks in this paper. The data flow of this Siamese ConvNet is demonstrated in Figure 7. The labels of inputs contain 0 and 1 representing two images are in same category or not respectively. For the specification of ConvNet, the dimension of inputs is $3 \times 227 \times 227$ and followed by conv1 ($96 \times 55 \times 55$), pool1 ($96 \times 27 \times 27$), norm1 ($96 \times 27 \times 27$), conv2 ($256 \times 27 \times 27$), pool2 ($256 \times 13 \times 13$), norm2 ($256 \times 13 \times 13$), conv3 ($384 \times 13 \times 13$), conv4 ($384 \times 13 \times 13$), conv5 ($256 \times 13 \times 13$), pool5 ($384 \times 6 \times 6$), fc6 (4096×1) and fc7 (4096×1). The weights will be shared by the two ConvNets and the outputs of fc7 will be concatenated. Then two fully-connected layers fc8 (64×1) and fc9 (2×1) are followed.

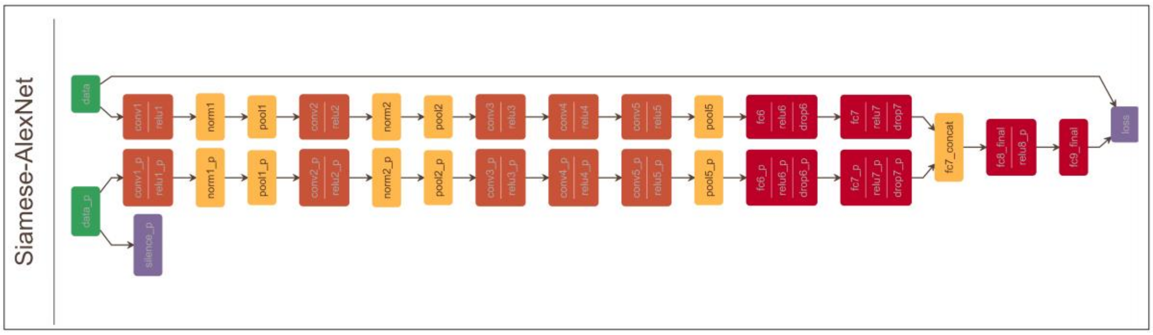


Figure 7. Data flow diagram of Siamese-AlexNet (Siamese with ConvNets). The twin networks share weights between convolutional layers and the outputs of fc7 is concatenated as fc7_concat. Then two fully-connected layers fc8 (64 × 1) and fc9 (2 × 1) are followed.

For other components of Siamese ConvNet is as follows.

- **Loss function**: Softmax is adopted to computer the probabilities of outputs by Equation (9). In this circumstance, softmax loss is equivalent to cross-entropy since calculation of cross-entropy can be denoted as Equation (10). Combined with Equation (8), the form of final objective on binary classifier can be expressed as Equation (11).

$$S_i = \frac{e^{a_i}}{\sum_{k=1}^{T} e^{a_k}} \tag{9}$$

$$L = -\sum_{i=1}^{T} y_i log P_i \tag{10}$$

$$L\left({X_1}^{(i)}, {X_2}^{(i)}\right) = y_i \cdot log\, S_i + (1 - y_i) \cdot log(1 - S_i) + \lambda \|W\|_2 \tag{11}$$

- **Optimization**: The optimization method is stochastic gradient descent (SGD) which is combined with a standard backpropagation algorithm. The gradient is accumulative across the two-stream networks with shared weights. A fixed mini-batch size is used with a learning rate η and momentum μ. Then the optimization process can be denoted as (12).

$$\theta \leftarrow \theta - \eta(\nabla_\theta L(x;\theta) - \lambda\theta) + \mu\theta' \tag{12}$$

Where x is input vector, θ is parameter vector need to be updated, θ' is update of last epoch, λ is regularization strength and $L(x;\theta)$ is loss function need to be minimized.

4. Experiments and Results

In this section, experiments for the proposed method have be demonstrated. Different subsets of *MIT67* [34] and *Places365* [19] were adopted to evaluate the performance of positive and negative pairs mining, visual representation learning and indoor scene recognition. A brief introduction of datasets and setup of experiments will been described first. Then, results and corresponding analysis will be presented respectively.

4.1. Experiment Setup

4.1.1. Datasets

We evaluated the performance of proposed method in two scene-centric datasets, *MIT67* and the indoor part of *Places365*. For *MIT67*, there are 67 indoor scene categories in this dataset, which distribute in 5 macro-categories including stores, home, public, leisure and working place. The total number of samples is 15620. For *Places365*, the original dataset contains scenes both from indoor, nature and urban scenario. Our task is to recognize indoor scenes. To this end, only indoor part of this dataset was considered in our experiments. There are 160 categories of indoor scenes in this dataset and we named it as *Places160*. In addition, there are two versions for this dataset, challenge version and standard version. For challenge version, the size of each category range from 3,068 to 40,000. For standard version, the range is 3,068 to 5,000.

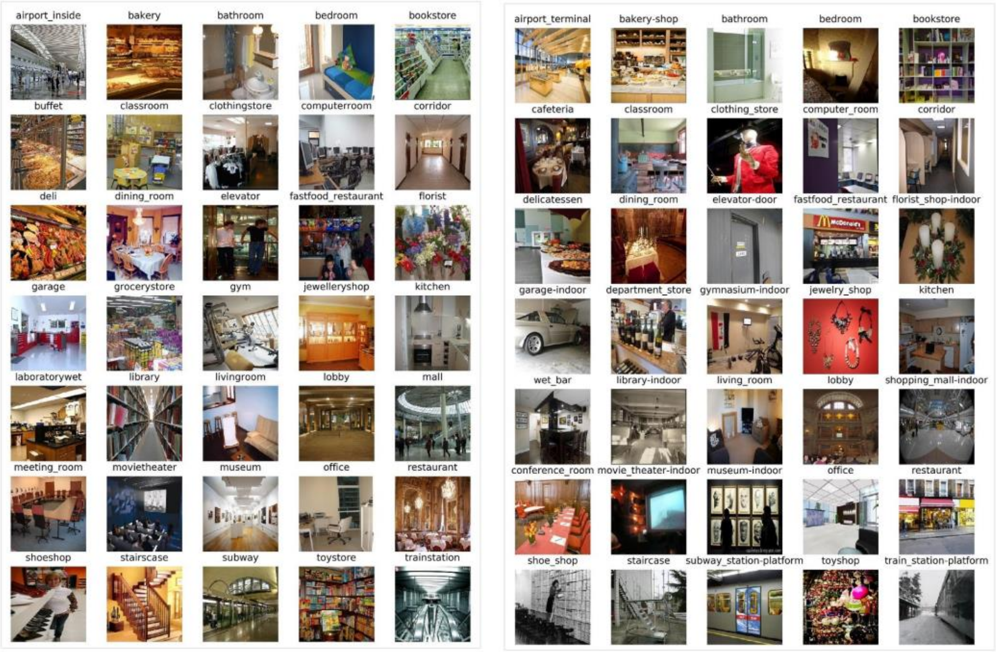
To better evaluate the performance of proposed method, the experiments are conducted in three settings. Considering the overlapping of the two datasets and their potential benefits for geolocation-based services, we evaluated data with 10, 35 and 67 categories respectively. And there are 7 groups of subset for comparison experiments in total.

(1) 10 categories. In this setting, there are 10 overlapped categories of *MIT67* and *Places160* and three groups of data would be considered, MIT-10, Places-10 and MIT-Places-10. MIT-10 and Places-10 only contains samples in corresponding dataset, while MIT-Places-10 is a mixed one.

(2) 35 categories. There are 35 overlapped categories of *MIT67* and *Places160*. Similar to (1), three groups of data would be evaluated, MIT-35, Places-35 and MIT-Places-35.

(3) 67 categories. In order to compare with the-state-of-art methods, we also considered *MIT67* which is a standard dataset has been adopted by a lot of previous works.

To be noted that, for overlapping datasets that contain both samples form *MIT67* and *Places365* (such as MIT-Places-10 and MIT-Places-35), some of scene names may different. We combined them by the names used in *MIT67* in our experiments. For dataset with 10 categories, they are subsets of 35 categories. Indoor scene categories in these datasets including airport inside, bedroom, bookstore, classroom, computer room, corridor, living room, office, mall and restaurant. For 35 categories, Figure 8 has illustrated samples from each category in both MIT-35 and Places-35. Generally speaking, data in Places-35 are more cluttered than MIT-35. Besides, the size of sample in Places-35 is also far more large than MIT-35. Figure 9 has shown a comparison for size of these two dataset.



(a) MIT-35 (b) Places-35

Figure 8. Example of image samples. (a) is samples from MIT-35 and (b) is samples from Places-35.

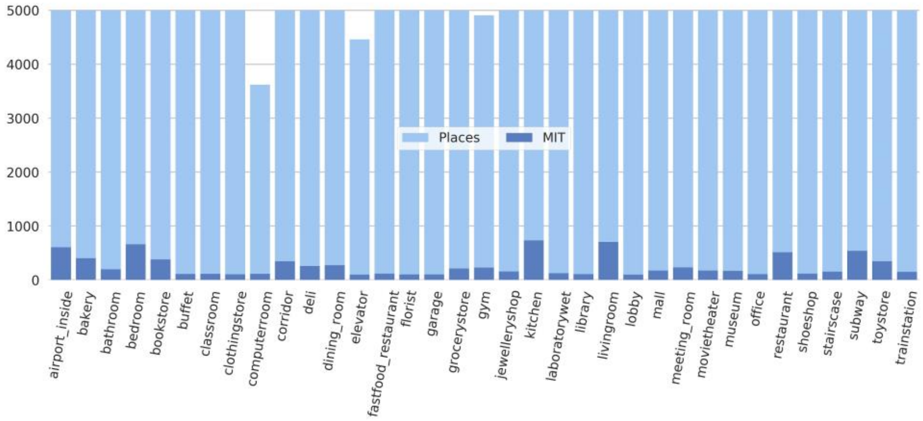


Figure 9. Comparison of data size for MIT-35 and Places-35. For better illustration, only standard version of *Places365* are considered.

4.1.2. Experiment Setup

The experiments in this paper were conducted on Ubuntu 16.04 system with two NVIDIA Titan X GPUs (12GB RAM for each). Deep learning framework Caffe [66] was adopted to train the Siamese-AlexNet. All images were resized to 256 by 256 pixels for convenience. As mentioned before, in order to construct the *k*-NN graph, original features need to be extracted from images. There are two ways to accomplish this: SIFT-based and CNN-based methods. In order to decrease influence by a pre-trained model and well prove the effectiveness of proposed method, dense SIFT was adopted to extract the local features in this period. VLFeat [65] is adopted for this process. After that, Fisher-Vector (FV) followed by [62] is used for feature encoding. Then Euclidean distances are used to find *k*-nearest neighbors of an image sample and weight the edges of constructed graph. For cycle detection in weighted direct graph, a shortest path algorithm named Floyd-Warshall [43] is employed. In order to validate the effectiveness of the trained Siamese network, CNN features were extracted to train an indoor scene recognition model with a simple SVM classifier.

4.2. Results of Pairs Mining

In this part, positive and negative pairs were mined respectively. Before constructing the graph, dataset for training would be randomly disrupted.

4.2.1. Positive Pairs Mining

In order to find best *k* value for graph construction, experiments were conducted with different data and mining results were demonstrated with different cycle lengths. Firstly, experiments on standard benchmark dataset MIT-67 was conducted to investigate the relationships between *k* value and pair amounts. As shown in Figure 10(a), for a fixed cycle length, when *k* increased, the number of mined positive pairs also increased. It is easy to understand because the links of *k*-NN graph would be increased when *k* increased. However, in Figure 10(b), it can be found that for most values of cycle length, the accuracy of positive pairs mining were decreased as *k* grown up. This phenomena is caused by redundant links between images in different categories when increasing the value of *k*. Therefore, the value of k cannot be too small or too large. If too small, there would be no enough pairs mined from the graph. On the contrary, if too large, noises will be brought in which may lead to a lower mining accuracy.

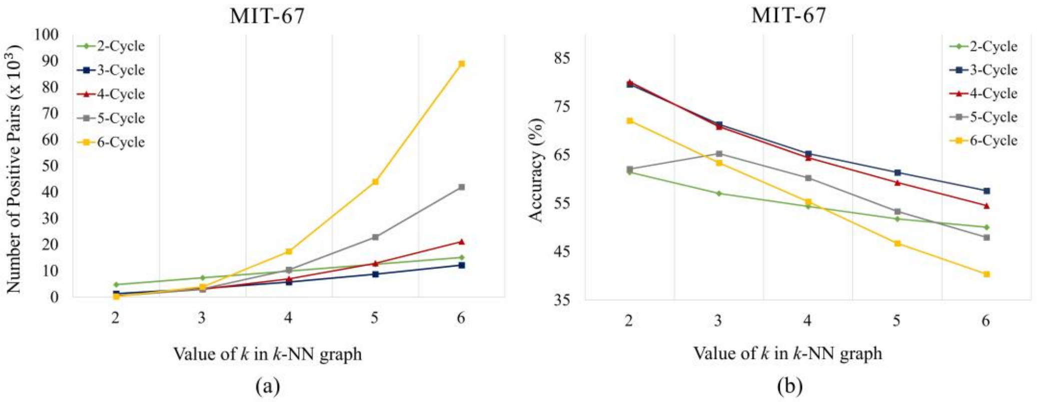


Figure 10. Results of positive pairs mining for MIT-67 with different *k* and cycle length. (a) Numbers of mined positive pairs; (b) Accuracy of mined positive pairs.

Furtherly, results for other groups of data also has been demonstrated to discover which *k* value and cycle length can achieve best performance for positive pairs mining. As illustrated in Figure 11, it is obvious that 4-Cycle (in red line) can achieve higher accuracy in most circumstances. The reason for this phenomenon is that noise may be bought in when circle length is too large. And when circle length is too small, samples in same categories may be omitted, especially for those images which have large appearance variations. Besides, it can be found that when *k* = 2, data except MIT-10 can achieve highest accuracy. However, number of mined positive pairs are too small in this circumstance. Relatively, when *k* = 3, no matter the amount of mined pairs or accuracy can achieve favorable performance. Thus, *k* = 3 and cycle length = 4 would be selected for following experiments. For comparison between different data groups, accuracy for positive pairs mining with different data has been illustrated in Table 1 with parameters selected.

Beyond this, it is also can be found in Figure 11 that when cycle length was set to 2 (2-Cycle in the figure), this method became a direct matching process. In this circumstance, mining accuracy would be lower than most cases with longer cycle length. Thus, it can be seen that positive pair mining method utilized in this paper is superior to a directed matching method from these experiments.

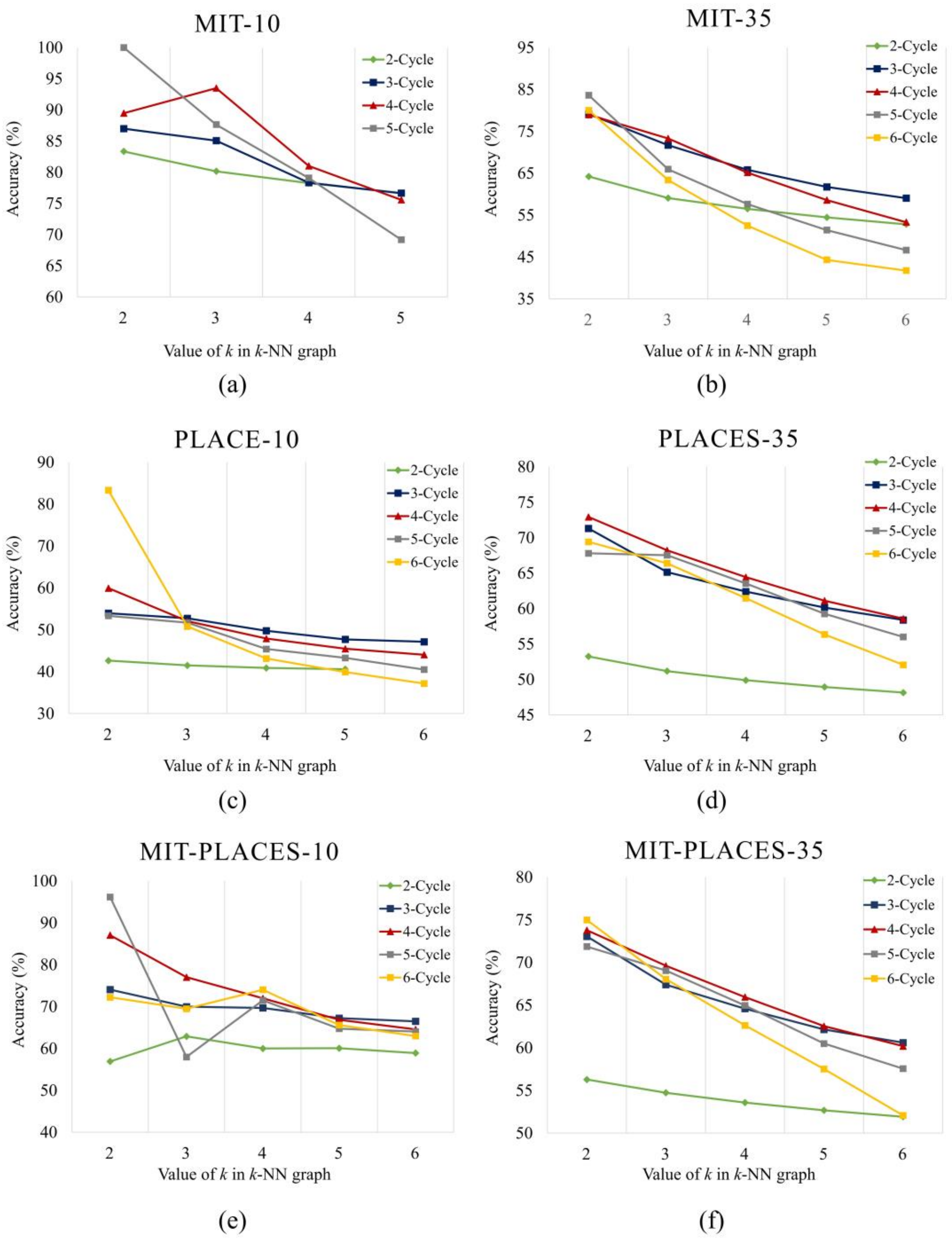


Figure 11. Accuracy of positive pairs mining for different groups of data. Subgraphs from (a) to (f) are mining accuracy for MIT-10, MIT-35, Places-10, Places-35, MIT-Places-10 and MIT-Places-35, respectively.

Table 1. Comparisons of accuracy (%) for positive pairs mining when $k = 3$ and cycle length = 4.

Number of Categories	MIT	Places	MIT-Places
10	93.4932	52.0602	77.0240
35	73.2836	68.2141	69.5983
67	70.8526	/	/

4.2.2. Negative Pairs Mining

For negative pairs mining, in order to evaluate performance of geodesic distance adopted in this paper, two other methods random sampling and Euclidean distance have been investigated for comparison. For random sampling, two samples were supposed to chosen from a dataset, then we computed the probability that they are not in same categories. For Euclidean distance and geodesic distance based methods, a same threshold was set to measure whether two samples in same category or not. In order to decrease computation, a batch size with 1000 was set to mine the negative pairs.

Table 2 has demonstrated the accuracy for these three methods and 7 groups of data. In this table, it can be found that even with random selecting, accuracy of negative pairs mining was still high. Since for a multi-categories dataset, most samples are in different categories. Therefore, with a specific data size, the more kinds of categories, the higher probability of sampling two images as a negative pair. While, with certain kinds of categories, if there are more samples in each categories, the higher accuracy can be achieved for negative pairs mining. Comparing three kinds of methods, it can be seen that the differences between their accuracy are subtle. For most circumstance, geodesic distance can achieve a slightly higher accuracy comparing to random sampling and Euclidean distance. Except for Places-35, both Euclidean distance and geodesic distance got lower accuracy comparing to random sampling. And for MIT-Places-35, geodesic distance was less accurate than Euclidean distance by 0.03%.

Even though, geodesic distance still can be taken as a better choice for negative mining for it has considered images pairs with similar appearance but belongs to different categories. It is easy to find negative pairs in a dataset with lots of categories. However, it cannot guarantee significant information has been obtained which may play an important role in indoor scene circumstances. Without negative pairs which have similar appearance but in same categories, the following learning process still cannot learn good representations for indoor scenes.

Table 2. Comparisons of accuracy (%) for different negative pairs mining methods.

Data	Random Sampling	Euclidean Distance	Geodesic Distance
MIT-10	85.8745	86.6667	87.0667
Places-10	89.9273	89.9417	90.0625
MIT-Places-10	89.8917	90.1519	90.6885
MIT-35	95.3423	95.5125	95.9875
Places-35	97.1358	97.0814	97.1140
MIT-Places-35	97.1309	97.2818	97.2497
MIT-67	97.5768	97.7429	98.0643

In general, due to the complexity of indoor scene data, mining positive and negative pairs from a graph structure is a feasible method. High-level semantic information for unlabeled scene data can facilitate following visual representation process. Figure 12 has shown some examples of mined pairs. it also can be found that challenge pairs such as classroom and shopping mall can be mined as positive pairs. And computer-room and office can be distinguished in different classes even they share same objects such as chair and computer.

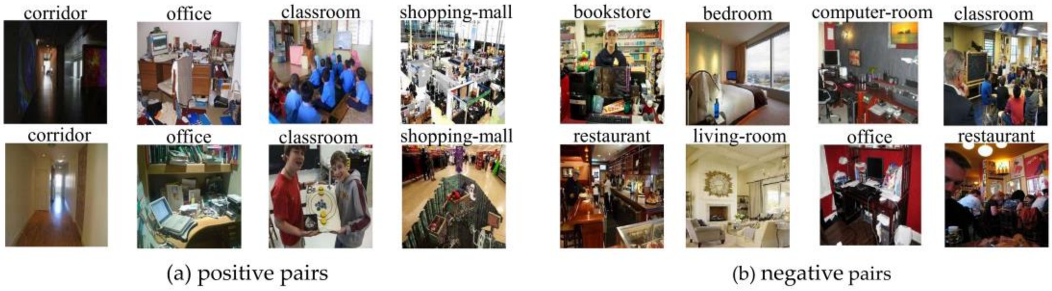


Figure 12. Examples of mined pairs. (a) are positive pairs and it can be seen that some of them have large variations. (b) are negative pairs while some of them have a similar appearance.

4.3. Results of Unsupervised Visual Representation Learning

As depicted in previous part, mined pairs would be used as inputs for indoor scene representation learning. In this experiment, AlexNet was adopted as architecture of subnetwork. A whole structure and data flow of Siamese-AlexNet has been illustrated in Figure 7. In order to decrease training time and avoid over fitting, a trained model of AlexNet which provided by Caffe has been adopted for initialization. In this case, datasets including MIT-10, Places-10, MIT-Places-10, MIT-35 and MIT-67 can quickly converge with 5k iterations. Whereas, due to the complexity of Places-35 and MIT-Places-35, iterations for these two datasets were set to 20k.

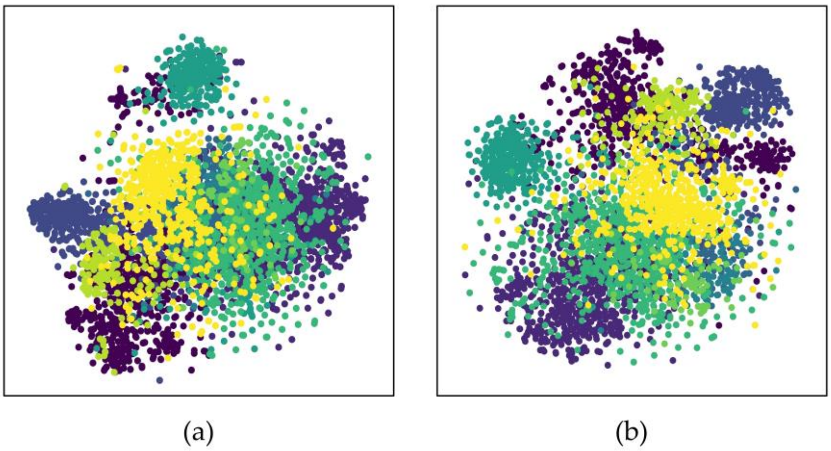


Figure 13. Comparison of feature embedding. (a) is embedding result of model trained with MIT-10. (b) is embedding with benchmark model bvlc_alexnet.

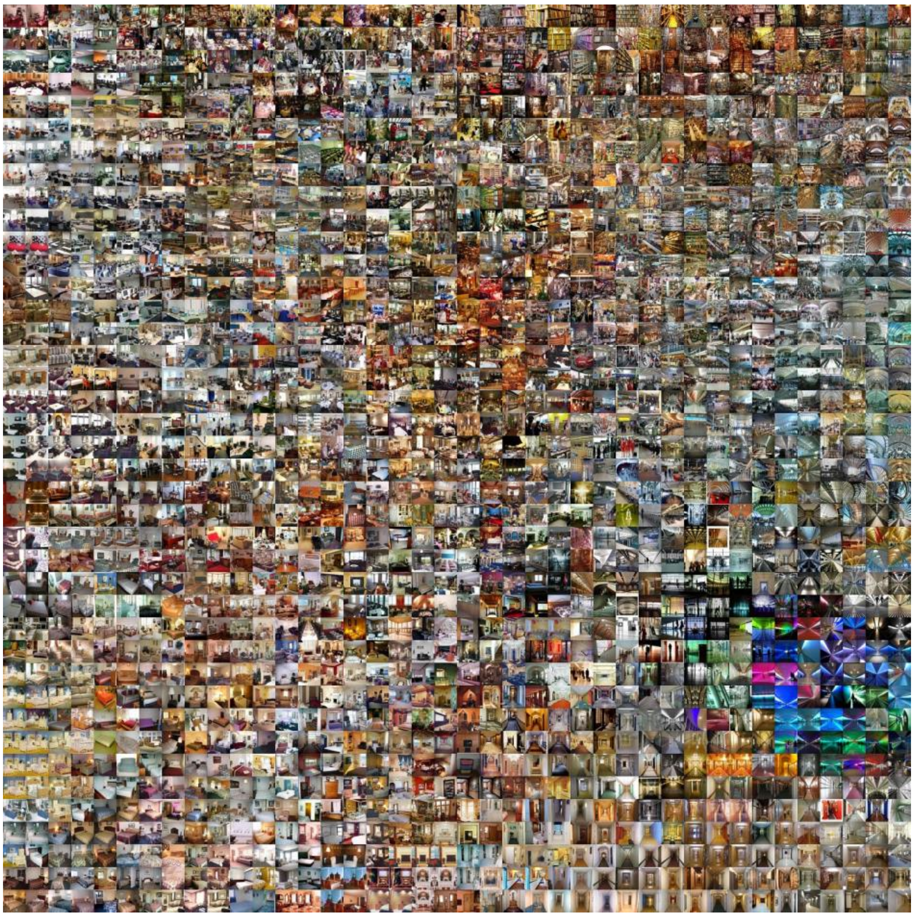


Figure 14. t-SNE visualization with images based on features extracted from the model trained with MIT-10. 4096-dimensional features has been utilized for this illustration.

In order to validate representations learned with the proposed method, examples of feature embedding with t-SNE [68] have been demonstrated in Figure 13. With dimension reduction, learned features can be visualized in two dimension. Figure 13(a) is embedding result with model trained on MIT-10, while Figure 13(b) is based on benchmark model provided by Caffe (bvlc_alexnet). From this comparison, it can be found that our model has a considerable effect with benchmark even it was trained in an unsupervised manner. For a more intuitive illustration, an embedding with images which were rearranged by similarity of fc7 outputs has been shown in Figure 14.

4.4. Results of Indoor Scene Recognition

To further evaluate the learned representations, experiments for indoor scene recognition have been conducted with CNN features extracted from trained Siamese ConvNet. Along with a few of labeled data as supervision, an indoor scene recognition model can be trained with CNN features and a simple SVM classifier. Recognition accuracy of 10-categories and 35-categories datasets have been shown in Table 3, and related confusion matrices have been demonstrated in Figure 15. While results for MIT-67 have been illustrated separately in Figure 16 for a closer look.

Table 3. Accuracy for indoor scene recognition with different datasets.

Dataset	MIT-10	Places-10	MIT-Places-10	MIT-35	Places-35	MIT-Places-35
Accuracy (%)	75.10	61.30	64.00	56.71	44.14	53.60

In Table 3 and Figure 15, it can be seen that models trained with data from *Places365* got poor performance than *MIT67*. This phenomenon was caused by the diversity and complexity of *Places365*. Thus, a mixed dataset achieved a higher accuracy than data from *Places365* and lower accuracy than data from *MIT67*. Moreover, by comparing results form 10-classes and 35-classes indoor scenes, it is also can be found that when number of categories increased, it would be more difficult to distinguished an indoor scene from others.

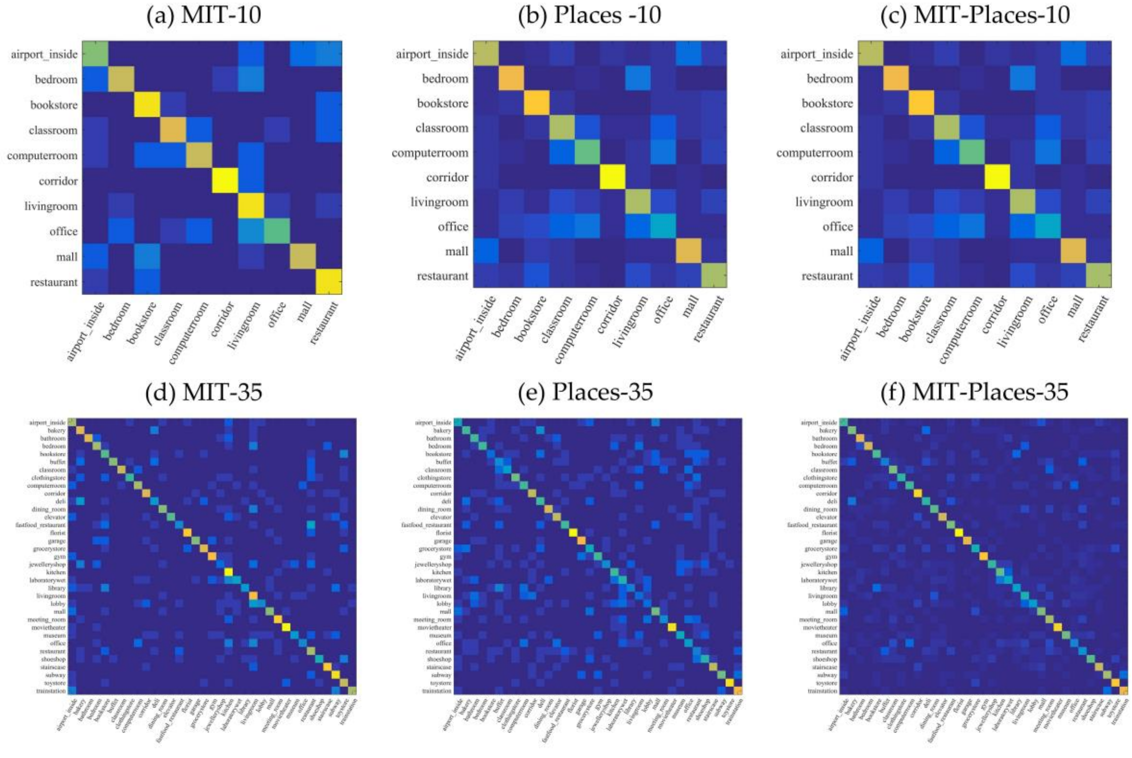


Figure 15. Confusion matrices for 10-categories datasets and 35 categories datasets. (a)-(c) are results of 10-categories while (d)-(f) are results of 35-categories.

For MIT-67, in Figure 16, it can be seen that scenes with similar-looking are prone to obtain lower classification accuracy, such as “library” vs “bookstore “, “dining room” vs ”living room" and “fast food restaurant” vs “bakery”.

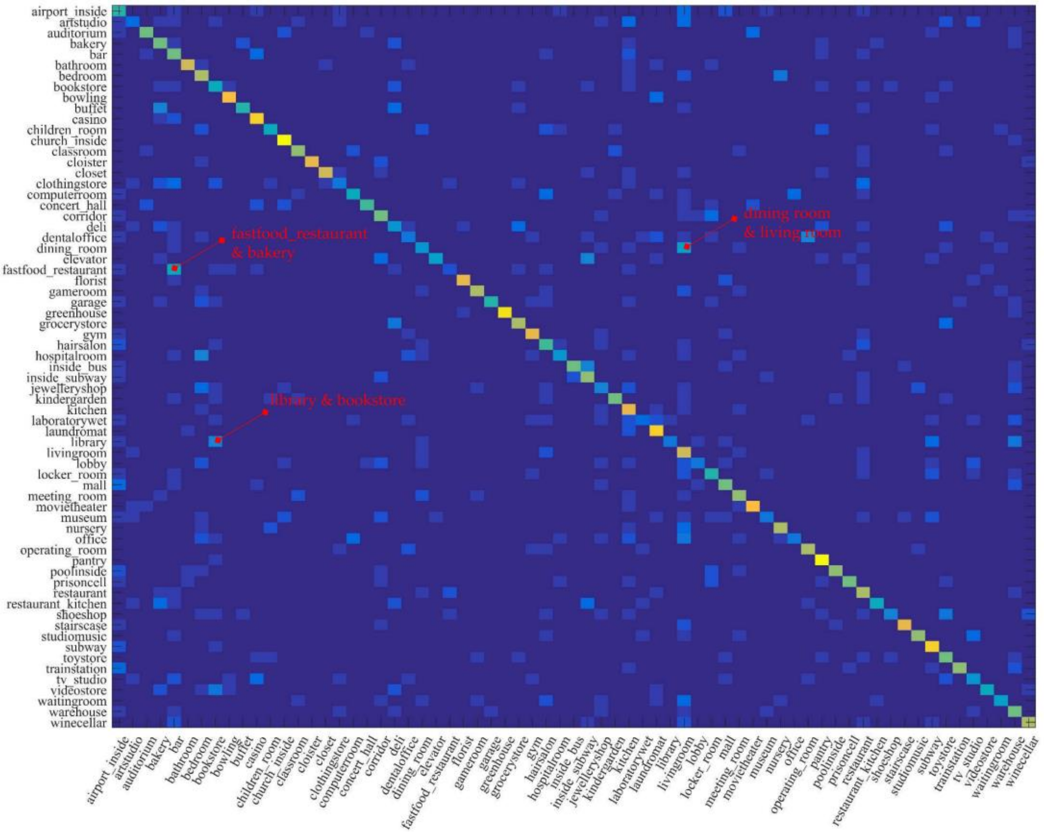


Figure 16. Confusion matrix for MIT-67. Examples of indoor scenes that are easy to confuse have been annotated, such as fastfood_restaurant with bakery, dining room with living room and library with bookstore.

Furthermore, in order to compare indoor scene recognition performance with methods of state-of-the-art, Table 4 has illustrated accuracy of different methods with MIT-67 dataset. In this table, it can be seen that, even without labeled data for representation learning, our methods still can achieve a 52.61% accuracy which is higher than traditional methods with manual features and supervision methods. Although, CNN features based methods can obtain higher accuracy than other methods, they require large scale well-annotated datasets which may contain at least millions of instances. This may restrict their utilization potential for indoor scene related applications.

Table 4. Comparison of indoor scene recognition accuracy on MIT-67.

Method	Accuracy (%)	Method	Accuracy (%)
ROI+GIST (09’) [35]	26.10	Embedding+Hypergraph (14’) [40]	39.05
Object Banks (10') [29]	37.60	GGM (14’) [39]	41.15
NN+BoW (11’) [38]	47.01	Places-CNN features (14’) [21]	68.24
RBoW (12’) [31]	37.93	CNNaug-SVM (14’) [13]	69.00
Hybrid-Parts+GIST+SP (12') [34]	47.20	SRP+Encoding(16’) [41]	71.80
n-SOINN (13’) [36]	33.73	CNN+Sparse-Coding(17’) [42]	87.22
Ours: MIT67-CNN features [1] + SVM	52.61		

[1] CNN features were from unsupervised learning, while others’ were from supervised learning.

5. Discussion

With experiments on cluttered indoor scene data, it can be found that visual representations of indoor scenes can be learned even without large scale well-labeled data. In this paper, a graph-based pairs mining approach has been adopted to obtain semantic relationships between unlabeled data. Then, a Siamese ConvNet was employed to learn indoor scene representations with mined pairs as supervisory information. Besides, more challenge indoor scene circumstances, which combined *MIT67* and indoor parts of *Places365*, have been taken into consideration and results on them have proved the effectiveness of proposed method.

For pairs mining, comprehensive experiments on positive and negative pairs were conducted respectively. By comparing results on different datasets with different parameters for *k*-NN graph, appropriate value of *k* and cycle length have been determined for indoor scene data, which are 3 and 4 respectively. The highest accuracy for positive mining is 93.49% and for negative mining is 98.06%. For representation learning, comparison of embedding results have been demonstrated to validate the learning effectiveness. At last, experiments of indoor scene recognition have been conducted to verify the learned representations further. Although, recognition results on MIT-67 achieve lower accuracy than other CNN feature based methods. The proposed method are based on unlabeled data setting and it can still achieve an accuracy with 52.61%, which is higher than most manual feature based methods in similar setting.

6. Conclusions

The main purpose of this paper is to investigate whether visual representations can be learned with a setting that there is no enough well-labeled indoor scene data. Considering application related to indoor circumstances and complexity of indoor scenes, learning effective indoor scene representations is a meaningful research area for future work. Although deep learning based methods have won great success in past years, most of them require massive manually annotating data to train the model, which is expensive and labor-intensive. Therefore, research on unsupervised representation learning for indoor scenes have great significance. Based on these, this paper proposed an unsupervised representation learning for indoor scenes to solve challenges both for indoor scene recognition and data annotations. Due to the ability of learning semantic relationships between unlabeled data, the proposed method has a great potential in automatic sample annotation and data cleaning for indoor scene. In future work, more flexible and end-to-end unsupervised learning methods can be considered for indoor scene representation learning.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: Statement for individual contributions of this paper is as follows. conceptualization, Mengyun Liu. and Ruizhi Chen; methodology, Mengyun Liu; software, Mengyun Liu and Haojun Ai.; validation, Haojun Ai and Yujin Chen; formal analysis, Mengyun Liu; investigation, Mengyun Liu; resources, Ruizhi Chen; data curation, Mengyun Liu; writing—original draft preparation, Mengyun Liu; writing—review and editing, Haojun Ai; visualization, Mengyun Liu; supervision, Deren Li; project administration, Ruizhi Chen; funding acquisition, Ruizhi Chen.

Funding: This research was funded by the National Key Research and Development Program of China (grant No. 2016YFB0502200 and 2016YFB0502201), NSFC (grant No. 91638203).

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

References

1. Oliva, A.; Torralba, A. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision* **2001**, *42*, 145-147. https://doi.org/10.1023/A:1011139631724
2. Kim, D.; Nevatia, R. Recognition and localization of generic objects for indoor navigation using functionality. *Image and Vision Computing* **1998**, *16*, 729–743. https://doi.org/10.1016/S0262-8856(98)00067-5
3. Epstein, Russell A. Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences* **2008**, *12*, 388-396. https://doi.org/10.1016/j.tics.2008.07.004
4. Kim, J.; Park, C; Kweon, I. S. Vision-based navigation with efficient scene recognition. *Intelligent Service Robotics* **2011**, *4*, 191–202. https://doi.org/10.1007/s11370-011-0091-x
5. Baddeley, B.; Graham, P.; Husbands, P.; Philippides, A. A Model of Ant Route Navigation Driven by Scene Familiarity. *PLoS Computational Biology* **2012**, 8(1): e1002336. https://doi.org/10.1371/journal.pcbi.1002336

6. Liu, M.; Chen, R.; Li, D.; Chen, Y.; Guo, G.; Cao, Z.; Pan, Y. Scene Recognition for Indoor Localization Using a Multi-Sensor Fusion Approach. *Sensors* **2017**, 17, 2847. https://doi.org/10.3390/s17122847
7. Chu, S.; Narayanan, C.; Kuo, J.C.; Mataric, M.J. Where am I? Scene recognition for mobile robots using audio features. In 2006 IEEE International Conference on Multimedia and Expo (ICME), Ontario, Toronto, 9-12 July, 2006; pp. 885-888.
8. Siagian, C.; Laurent I. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics* **2009**, *25*, 861-873. https://doi.org/10.1109/TRO.2009.2022424
9. Sünderhauf, N.; Dayoub F.; McMahon S.; Talbot B.; Schulz R.; Corke P.; Wyeth G.; Upcroft B.; Milford M. Place categorization and semantic mapping on a mobile robot. In 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16-21 May, 2016; pp. 5729-5736.
10. Li, L.J.; Fei-Fei L. What, where and who? Classifying events by scene and object recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 14-20 October, 2007; pp. 1-8.
11. Choi, W.; Khuram S.; Silvio S. Learning context for collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),Colorado Springs, CO, USA, 20-25 June, 2011; pp. 3273-3280.
12. Ikizler-Cinbis, N.; Stan S. Object, scene and actions: Combining multiple features for human action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Crete, Greece, 5-11 September, 2010; pp. 494-507.
13. Sharif R. A.; Azizpour H.; Sullivan J.; Carlsson S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Columbus, Ohio, 23, 28 June, 2014; pp. 806-813.
14. Yosinski, J.; Jeff C.; Yoshua B.; Hod L. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems (NIPS), Montréal, CANADA, 8-13 December, 2014; pp. 3320-3328.
15. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June, 2009; pp. 248–255.
16. Krizhevsky, A.; Ilya S.; Geoffrey E.H. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, Carson City, Nevada, USA, 3-8 December, 2012; pp. 1097-1105.
17. Simonyan, K. Andrew Z. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
18. 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 (CVPR), Paradise, Nevada, USA, 27-30 June, 2016; pp. 1097-1105.
19. 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* **2018**, *40*, 1452-1464. https://doi.org/10.1109/TPAMI.2017.2723009
20. Amazon Mechanical Turk. Available online: https://www.mturk.com/ (accessed on 20 February 2019).
21. Zhou, B.; Lapedriza A.; Xiao J.; Torralba A.; Olive A. Learning Deep Features for Scene Recognition using Places Database. In Advances in Neural Information Processing Systems (NIPS), Montréal, CANADA, 8-13 December, 2014; pp. 487-495.
22. Goodfellow, I.; Yoshua B.; Aaron C. Representation Learning. In *Deep learning*, 1st ed.; MIT press: Cambridge, Massachusetts, USA, 2016. Volume 1, pp. 517-544.
23. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a" siamese" time delay neural network. In Advances in Neural Information Processing Systems (NIPS), Denver, Colorado, USA, 21 November – 1 December, 1994; pp. 737-744.
24. Koch, G.; Richard Z.; Ruslan S. Siamese Neural Networks for One-shot Image Recognition. In Proceedings of the International Conference on Machine Learning (ICML) Deep Learning Workshop, Lille, France, 10-11 July, 2015; Vol. 2.
25. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A.; Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Paradise, Nevada, USA, 27-30 June, 2016; pp. 2921-2929.

26. Torralba, Murphy, Freeman and Rubin, Context-based Vision System for Place and Object Recognition. In Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV), Nice, France, 13-16 October, 2003; pp. 273-280.
27. Lu, L.; Toyama, K; Hager, G.D. A Two Level Approach for Scene Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20-26 June, 2005; pp. 688-695.
28. Van D.S.; Koen, T.G.; Cees S. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2010**, *32*, 1582-1596. https://doi.org/10.1109/TPAMI.2009.154
29. Li, L.J.; Su, H.; Fei-Fei, L.; Xing, E.P. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In Advances in Neural Information Processing Systems (NIPS), Vancouver, CANADA, 6-11 December, 2010; pp. 1378-1386.
30. Brown, M.; Süsstrunk, S. Multi-spectral SIFT for scene category recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20-25 June, 2011; pp. 177-184.
31. Parizi, S.N.; Oberlin, J.G.; Felzenszwalb, P.F. . In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, June 16-21, 2012; pp. 2775-2782.
32. Niu, Z.; Hua, G.; Gao, X.; Tian, Q. Context aware topic model for scene recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, June 16-21, 2012; pp. 2743-2750.
33. Kwitt, R.; Vasconcelos, N.; Rasiwasia, N. Scene Recognition on the Semantic Manifold. In Proceedings of the European Conference on Computer Vision (ECCV), Firenze, Italy, 7-13 October, 2012; pp. 359-372.
34. Zheng, Y.; Jiang, Y.G.; Xue, X. Learning hybrid part filters for scene recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Firenze, Italy, 7-13 October, 2012; pp. 172-185.
35. Quattoni A.; Torralba A. Recognizing Indoor Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, USA, 20-25 June, 2009; pp. 413-420.
36. Kawewong A.; Pimup R.; Hasegawa O. Incremental Learning Framework for Indoor Scene Recognition. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, USA, July 14-18, 2013; pp. 496-502.
37. Espinace, P.; Kollar, T.; Soto, A.; Roy, N. Indoor Scene Recognition through Object Detection. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Anchorage, Alaska, USA, 3-7 May, 2010; pp. 1406-1413.
38. Cakir, F.; Güdükbay, U.; Ulusoy, Ö. Nearest-Neighbor based Metric Functions for Indoor Scene Recognition. *Computer Vision and Image Understanding* **2011**, *115*, 1483–1492. https://doi:10.1016/j.cviu.2011.07.007
39. Elguebaly T.; Bouguila N. Indoor Scene Recognition with a Visual Attention-Driven Spatial Pooling Strategy. In Proceedings of the Canadian Conference on Computer and Robot Vision (CRV), Montreal, QC, Canada, 6-9 May, 2014; pp. 268-275.
40. Yu J.; Hong C.; Tao D.; Wang M. Semantic embedding for indoor scene recognition by weighted hypergraph learning. *Signal Processing* **2015**, *112*, 129-136. https://doi.org/10.1016/j.sigpro.2014.07.027
41. Khan S H.; Hayat M.; Bennamoun M.; Togneri R.; Sohel F.A. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing* **2016**, *25*, 3372-3383. https://doi.org/10.1109/TIP.2016.2567076
42. Nascimento G.; Laranjeira C.; Braz V.; Lacerda A.; Nascimento E.R. A Robust Indoor Scene Recognition Method based on Sparse Representation. In Proceedings of the Iberoamerican Congress on Pattern Recognition. Valparaíso, Chile, 7-10 November 2017; pp. 408-415.
43. Floyd–Warshall algorithm. Available online: https://en.wikipedia.org/wiki/Floyd%E2%80%93Warshall_algorithm (accessed on 20 February 2019).
44. Bengio Y.; Courville A.; Vincent P. Representation Learning: A Review and New perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2013**, *35*, 1798-1828. https://doi.org/10.1109/TPAMI.2013.50
45. Hinton G.E.; Osindero S.; Teh Y.W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* **2006**, *18*, 1527-1554.

46. Bengio Y.; Lamblin P.; Popovici D.; et al. Greedy layer-wise training of deep networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, 04-07 December, 2006; pp. 153-160.
47. Vincent P.; Larochelle H.; Bengio Y.; Manzagol P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki, Finland, 5-9 June, 2008; pp. 1096-1103.
48. Bosch A.; Zisserman A.; Muñoz X. Scene Classification using a Hybrid Generative/Discriminative Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2008**, *30*, 712-727. https://doi.org/10.1109/TPAMI.2007.70716
49. Srivastava N.; Salakhutdinov R.R. Multimodal Learning with Deep Boltzmann Machines. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, United States, December 3-6, 2012; pp. 2222-2230.
50. Le Q.V.; Ramzato M.A.; Monga R.; Devin M.; Chen K.; Corrado G.S.; Dean J.; Ng A.Y. Building High-level Features using Large Scale Unsupervised Learning. *arXiv* **2011**, arXiv:1112.6209.
51. Donahue J.; Jia Y.; Vinyals O.;Hoffman J.; Zhang N.; Tzeng E.; Darrell T. Decaf: A deep convolutional activation feature for generic visual recognition. In Proceedings of the International Conference on Machine Learning (ICML), Beijing, China, 21-26 June, 2014; pp. 647-655.
52. Wang X.; Gupta A. Unsupervised Learning of Visual Representations using Videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7-13 December, 2015; pp. 2794-2802.
53. Pathak D.; Krahenbuhl P.; Donahue J.; Darrell T.; Efros A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June, 2016; pp. 2536-2544.
54. Doersch C.; Gupta A.; Efros A.A. Unsupervised Visual Representation Learning by Context Prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7-13 December, 2015; pp. 1422-1430.
55. Li D.; Hung W.C.; Huang J.B.; Wang S.J.; Ahuja N.; Yang M.H. Unsupervised Visual Representation Learning by Graph-based Consistent Constraints. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, 11-14, October, 2016: 678-694.
56. Anselmi F.; Leibo J.Z.; Rosasco L.; Mutch J.; Tacchetti A.; Poggio T. Unsupervised Learning of Invariant Representations. *Theoretical Computer Science* **2016**, *633*, 112-121. https://doi.org/10.1016/j.tcs.2015.06.048
57. Cheriyadat A.M. Unsupervised Feature Learning for Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2014**, *52*, 439-451. https://doi.org/10.1109/TGRS.2013.2241444
58. Zhang F.; Du B.; Zhang L.P. Saliency-guided Unsupervised Feature Learning for Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2015**, *53*, 2175-2184. https://doi.org/10.1109/TGRS.2014.2357078
59. Hu F.; Xia G.S.; Wang Z.; Huang X.; Zhang L.P.; Sun H. Unsupervised Feature Learning via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2015**, *8*, 2015-2030. https://doi.org/10.1109/JSTARS.2015.2444405
60. Tao C.; Pan H.; Li Y.; Zou Z. Unsupervised Spectral–spatial Feature Learning with Stacked Sparse Autoencoder for Hyperspectral Imagery Classification. *IEEE Geoscience and Remote Sensing Letters* **2015**, *12*, 2438-2442. https://doi.org/10.1109/LGRS.2015.2482520
61. Yang J.; Jiang Y.G.; Hauptmann A.G.; Ngo C.W. Evaluating Bag-of-visual-words Representations in Scene Classification. In Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval (MIR), Augsburg, Bavaria, Germany, 24-29 September, 2007; pp. 197-206.
62. Simonyan K.; Parkhi, O.; Vedaldi A.; Zisserman A. Fisher Vector Faces in the Wild. In Proceedings of the British Machine Vision Conference (BMVC), Bristol, UK, 9-13 September, 2013; pp. 8.1-8.11.
63. Jégou H.; Douze M.; Schmid C.; Pérez P. Aggregating Local Descriptors into a Compact Image Representation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13-18 June, 2010; pp. 3304-3311.
64. Zheng L.; Yang Y.; Tian Q. SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2018**, 40, 1224-1244. https://doi.org/10.1109/TPAMI.2017.2709749
65. Vedaldi A.; Fulkerson B. VLFeat: An Open and Portable Library of Computer Vision Algorithms. 2008. http://www.vlfeat.org.

66. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the ACM International Conference on Multimedia (MM), Orlando, FL, USA, 3–7 November 2014; pp. 675-678.
67. Dijkstra's algorithm. Available online: https://en.wikipedia.org/wiki/Dijkstra%27s_algorithm (accessed on 20 February 2019).
68. Maaten, L.V.D.; Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9(Nov)*, 2579-2605.