Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.016287

ARTICLE

A Novel Feature Aggregation Approach for Image Retrieval Using Local and Global Features

Yuhua Li[1], Zhiqiang He[1,2], Junxia Ma[1,*], Zhifeng Zhang[1,*], Wangwei Zhang[1], Prasenjit Chatterjee[3] and Dragan Pamucar[4]

[1]Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou, 450001, China

[2]China Mobile Group Henan Company Limited, Xinxiang, 453000, China

[3]Department of Mechanical Engineering, MCKV Institute of Engineering, Howrah, 711204, India

[4]Department of Logistics, University of Defence in Belgrade, Belgrade, 11000, Serbia

[*]Corresponding Authors: Junxia Ma. Email: jxma@zzuli.edu.cn; Zhifeng Zhang. Email: zhangzhifeng@zzuli.edu.cn

Received: 23 February 2021 Accepted: 13 October 2021

ABSTRACT

The current deep convolution features based on retrieval methods cannot fully use the characteristics of the salient image regions. Also, they cannot effectively suppress the background noises, so it is a challenging task to retrieve objects in cluttered scenarios. To solve the problem, we propose a new image retrieval method that employs a novel feature aggregation approach with an attention mechanism and utilizes a combination of local and global features. The method first extracts global and local features of the input image and then selects keypoints from local features by using the attention mechanism. After that, the feature aggregation mechanism aggregates the keypoints to a compact vector representation according to the scores evaluated by the attention mechanism. The core of the aggregation mechanism is to allow features with high scores to participate in residual operations of all cluster centers. Finally, we get the improved image representation by fusing aggregated feature descriptor and global feature of the input image. To effectively evaluate the proposed method, we have carried out a series of experiments on large-scale image datasets and compared them with other state-of-the-art methods. Experiments show that this method greatly improves the precision of image retrieval and computational efficiency.

KEYWORDS

Attention mechanism; image retrieval; descriptor aggregation; convolutional neural network

1 Introduction

Content-based image retrieval (CBIR) has been a spotlight in the field of computer vision. It represents the image as a vector by feature extraction algorithms. It uses the nearest neighbor search methods to find the image like the given query image, among which feature extraction algorithms play a key role in improving image retrieval performance. So, in order to extract image features with more discriminability and form effective image representation, a lot of research has been done on feature extraction algorithms. In recent ten years, it has experienced a

development process from extracting shallow layer features based on Scale-invariant feature transform (SIFT) [1], speeded up robust features (SURF) [2] algorithms and embedding coding method in combination with bag of words (BOW) [3,4], fisher vector (FV) [5] and vector of local aggregated descriptors (VLAD) [6] to extracting deep layer features based on the deep convolutional neural network.

In the early stage, the convolutional neural network (CNN) based image retrieval model regards the output of the fully-connected layer as image features [7]. However, Liang et al. [8] and Liu et al. [9] proved that the image features extracted from the last convolutional layer, namely the deep convolution features, were more significant than the fully-connected layer. Later, CNN based image retrieval models used deep convolution features and achieved good retrieval results, such as regional maximum activation of convolutions (R-MAC) [10], sparse bundle adjustment (SBA) [11].

Recent studies show that features extracted by deep convolution neural networks pre-trained on large scale datasets, such as ImageNet, is better than traditional features in the image retrieval task [12]. The features output by the convolutional layer in the deep networks can be regarded as the local representation of the image and can display more details of the images, achieving a higher retrieval accuracy than the features output by the fully-connected layer. The accuracy was improved from 87.63% to 90.04% and 91.35% [13]. Therefore, the current mainstream algorithms are to aggregate the output features of the convolutional layer to form a local representation of the image.

Global features are appropriate if the user is interested in the whole image, rather than in the foreground. But the inability to distinguish between foreground and background is an inherent disadvantage of the global feature. Especially if the objects are affected by occlusion and so on, the global feature is likely to be invalid [14]. However, local features are stable and distinguished. In this way, under the condition that the objects are not completely occluded, some local features still exist stably to represent the objects (or even the images), which is convenient for the subsequent processing.

We can look at the Fig. 1, the left column is the whole image, the middle column is some corner points (local features), and the right column is the line segment except for the corner points. When we look at the middle column, we can more sensitively imagine them as the original objects in the left column. If we replace the whole image with these steadily occurring points, we can greatly reduce the large amount of information that the image originally carries, and at the same time suppress the influence of background noise. On the other hand, when the object is disturbed, even if some redundant information (such as parts with gentle color changes and lines) is occluded, we can still restore important information on the feature points that have not been occluded.

The deep local features have stronger discriminability than the classic hand-crafted features. However, some studies [15] have shown that the dimension of deep convolutional feature is high in the image retrieval model. For example, the dimension of the second fully-connected layer in visual geometry group network (VGG16) is 4096 and last fully-connected layer is 1000. ResNet's output is 1000 dimensions. In the traditional method, the dimension of feature points extracted by SIFT is 128, SURF is 64 and oriented fast and rotated brief (ORB) is 32 [16]. So, some scholars try to reduce the dimension of the features with embedding method. Whereas the traditional embedding is more suitable for the classic SIFT features than the deep convolutional features. In addition, the research [17] showed that compared with the common coding aggregation methods, the democratic

aggregation of deep convolutional features can obtain better performance. Yu et al. [18] did an experiment to destroy information in a random area of an image to see if the output of the network depended on that region. It was found that the area that the network could actually cover could reach about a third of the whole picture, far from the size of the receptive field. In order to solve the problem of how to utilize all the image context information, ParseNet proposes a method of combining global information with local information. Therefore, this paper focuses on the feature aggregation method of deep convolutional features, to overcome the existing problems of CBIR methods based on global and local features.

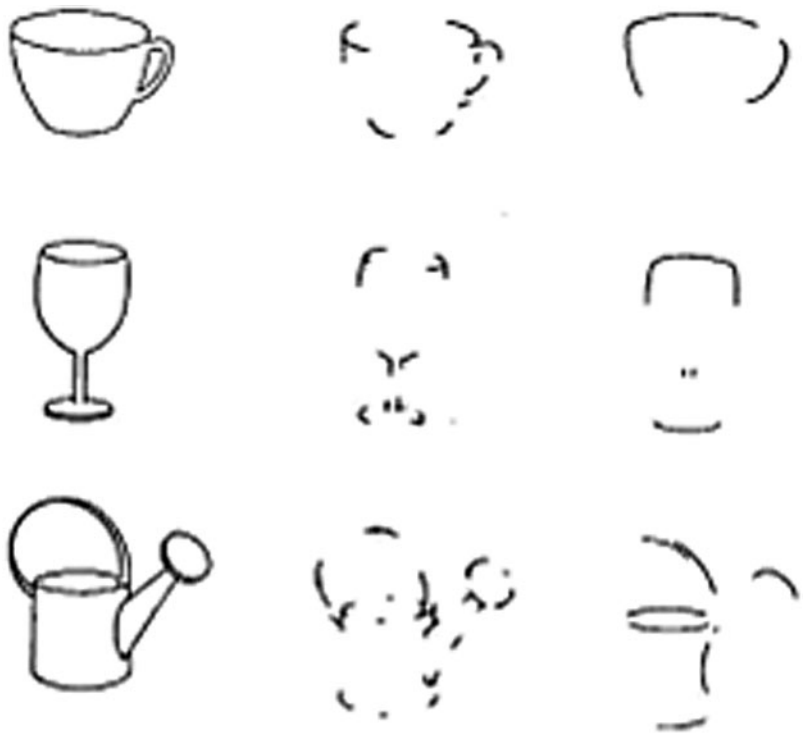

Figure 1: Illustration of the local features. When we look at the middle column, we can more sensitively imagine them as the original objects in the left column

For the study of feature aggregation mechanism, many papers proposed to use VLAD or Fisher vector to describe regions of images, and then aggregate multiple regions into image representation. Liu et al. [9] leverage the grid structure from [11], approximately 20 regions are selected per image, to pool pretrained CNN features into compact representations. Subsequently, the feature aggregation mechanism based on region gradually becomes the mainstream. R-VLAD and ASMK are the most representative algorithms [13]. All the methods contain two steps: regional search and regional aggregation. But the regional search method has the disadvantage of overlapping regions (i.e., there may be the same feature information in different regions) and can result in higher memory and complexity costs in the regional aggregation process [15].

In this work, we propose a new large-scale image retrieval method, which is based on aggregated local feature descriptors and global feature descriptor, utilizing the advantages of the combination of the two descriptors. Our first contribution is that we use the ResNet model to

extract the local features from the output of conv4x convolutional block, and use fully-connected layer output as the global feature of the image. Also employ attention mechanism to select keypoints from local features. Second, we propose a new feature aggregation technology with attention mechanism, which make each part of a feature vector has different contribution in the image matching process. In the method, the attention function can generate feature scores using very little extra computation.

The paper is organized as follows. In the next section, we briefly review the most relevant work. In Section 3, we describe our proposed method in detail. Section 4, we present experimental results on ROxf dataset, RPar dataset and Google Landmark dataset to explore the effectiveness of the proposed algorithm. Section 5 is our overall summary of this paper.

2 Related Work

The initial application of CNN in image retrieval was mainly to extract the output of fully-connected layer as the image feature vector [19]. Deep Image Retrieval (DIR) [20] is a global descriptor that achieves the state-of-the-art performance in several existing datasets. DIR feature descriptors are 2,048 dimensional and multi-resolution descriptors are used in all cases. Braux-Zin et al. [21] proposed a lighter deep convolutional neural network model (Lighten VGGNet), which used multi-task classification method to retrieve the images, so that the features extracted from the network during retrieval process have the ability to represent more refined category attributes, thus improve the accuracy of image retrieval. Aggregated Selective Match Kernel (ASMK) [22] is a recent global descriptor that obtains high performance in existing datasets. The CNN based on VGG16 extracts 512 dimensional global descriptor.

However, only extracting the whole connection layer cannot maintain the spatial structure, and this feature represents more global information and loses the local feature information of the image, resulting in a low mean Average Precision (mAP) of the retrieval. Therefore, the fusion of multiple image features has become a popular research method for image retrieval. Regional Maximum Activation of Convolutions (R-MAC) used a sliding window to obtain multiple local features for each feature graph, and considered the local features to be distinguished. Then, the local features obtained were summed and aggregated into full feature vectors. Kumar et al. [23] assumed that the most discriminative regions are located at the center of the image, and a space gaussian weighting matrix is designed to weight the deep convolutional features. Finally, max-pooling was adopted to aggregate the image features into compact vectors.

The feature aggregation of convolutional layer is divided into two types. One of them is encoding aggregation [24]. In this kind of methods, the deep features of convolutional layer are regarded as local features similar to SIFT, and finally these features are aggregated to get the representation of the input image. The other is direct aggregation, in which the feature maps of the convolutional layer are directly summed and aggregated or weighted to form an image representation. Wei et al. [25] aggregated the deep convolution feature of the image in the dataset into the feature vector, calculated the variance according to the channel, and selected the feature maps of some channels with the largest variance to carry out weighted aggregation of the deep features.

Based on the key feature points of deep convolutional neural network, Azizpour et al. [26] proposed a two-channel neural network model, which not only extracts the features of the fully-connected layer, but also takes samples under the convolution feature layer in the front as local features, and then combines two kinds of features. Hu et al. [27] designed the current largest labeling clothing image dataset DeepFashion. He divided the convolution of the last layer of

VGG16 into three branches. One branch performed feature point visibility prediction and position regression, and the other two branches extracted local features of the image, which is helpful to deal with the problem of clothing deformation and occlusion, and global features respectively. CONGAS [28] is a 40D hand-engineered local feature, which has been widely used for instance-level image matching and retrieval. This feature descriptor is extracted by collecting Gabor wavelet responses at the detected keypoint scale and orientation, and is known to have very similar performance and characteristic to other gradient-based local descriptors like SIFT.

Our work is related to the recent image search [29–31], which employ a new model as deep local and global features (DELG), combining global feature with local features into an image representation. The model performance based on the deep local features and global feature is better than that of other models. Because the local features can express images of different semantic classes, and the global feature can capture similarities across different poses of the instances. Similar to our method, Hou et al. [32] proposed an unsupervised semantic-based aggregation of deep features. However, different from the approach introduced [33–35], in our method, we use deep local features aggregation based on attention mechanism and global feature. Moreover, we aggregate the deep local features by using the new aggregation technology, in which each local feature has a different contribution according to the scores evaluated by attention mechanism.

3 Method

We propose method to improve image retrieval performance by using aggregated local features with attention mechanism [36–38] combined with global feature. Since attention mechanism recently demonstrated better performance on the large image retrieval benchmark, we adapted it to fit our system. We describe the whole process of the proposed method, mainly including features extraction, dimensionality reduction, descriptors aggregation, and image retrieval. Firstly, we introduce the method of feature extraction and the attention mechanism utilized in this process. And then explain dimensionality reduction for extracted features in the process. Thirdly, summarize the formulation of the feature descriptor aggregation. At last present the image retrieval procedure. The key workflow of the proposed image retrieval method is depicted in Fig. 2. And Fig. 3 is architecture of our proposed model for feature extracting and aggregating.

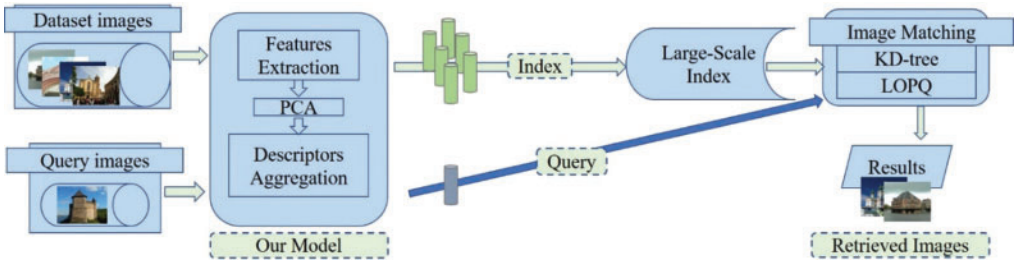

Figure 2: Key workflow of image retrieval with the proposed descriptor aggregation mechanism. The image coming from the dataset is input into the model to obtain the image feature descriptors, and then the composite feature descriptors are indexed and stored into Large-Scale Index

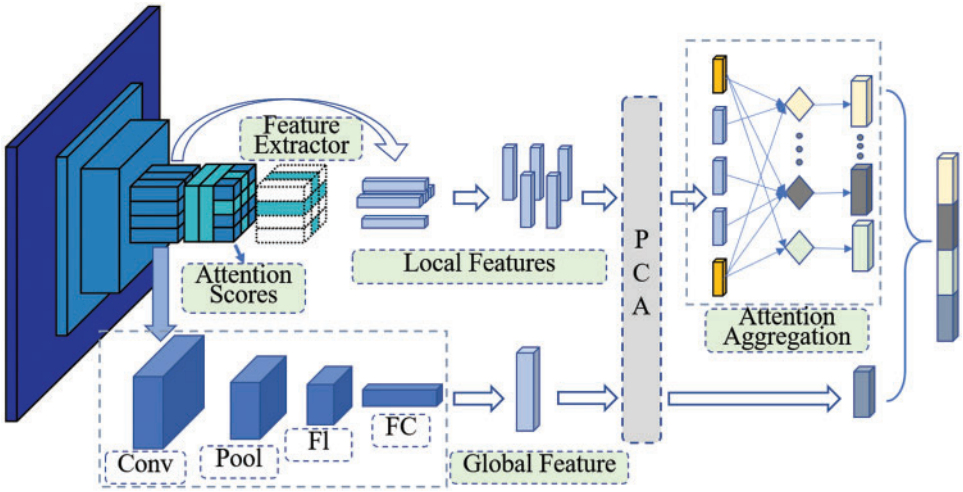

Figure 3: The network architecture of the model proposed in this paper. The global feature descriptor output from the fully connected layer and the aggregated local feature descriptors based on the attention scores are fused to form the final image presentation

3.1 Attention-Based Feature Extraction

Attention mechanism is widely used in computer vision [29,33]. The core of the attention mechanism is to make the neural network pay more attention to the feature information related to the image, which can speed up the processing time of the image feature, and then improve the computing efficiency and the expressing ability of the feature.

In this paper, the attention mechanism is used to learn the relationship between output channels, and then uses the score function to convolve the resulting feature graph to calculate the corresponding score. A vector with the same dimension as the number of channels is obtained, and the vector is scored as a feature and added to the corresponding local feature. In the attention network, firstly, the input feature maps are globally pooled, and then are two fully connected layers. The number of neurons in the first fully connected layer is C, and the number of neurons in the second fully connected layer is $D = 16 \times C$. Then, another Sigmod layer outputs $1 \times 1 \times D$ to get the corresponding score of each feature.

However, some shortcomings of this method are found in experiments. Although feature extraction and score function can be trained by backward propagation at the same time, the result of training is a weak model. Therefore, we use a two-step training strategy. First, we train the feature extraction ability of the model using datasets, and then train the score function on the model.

3.1.1 Local and Global Feature Extraction

The feature extraction process can be divided into two steps. The first step is feature generation, both dense local features and global features are extracted from images by improved ResNet-50 [39,40], which is trained on google landmark datasets and fine-tune for enhancing the discriminability. We extract the local features from the output of conv4x convolutional block, and use fully-connected layer output as the global feature of the image. And we trained the network with standard cross-entropy loss for images classification. The second step is the keypoint selection for local feature descriptors based on attention mechanism. Instead of using dense

features directly, attention mechanisms are used to select the most relevant features in lots of local features. The output of the attention mechanism is the weighted sum of the convolution features extracted by the network.

To handle the scale of input images, we construct an image pyramid. The obtained feature maps are regarded as a dense grid of local descriptors. Features are localized based on their receptive field of conv4 convolutional block. The receptive field size of the original image scale is 291×291 [41]. Through the image pyramid, features describing different size image regions can be obtained. Firstly, the input images are rescaled to 300×300. Then the images are randomly clipped to 224×224 size for training.

3.1.2 Attention Mechanism

Since a considerable part of the features extracted directly are not helpful to our retrieval task and will bring bad effects to the retrieval, the keypoint selection is particularly important for the retrieval system. We recommend training an attention mechanism to explicitly measure the correlation scores of local features. For training the function, the features are set into a weighted sum, where the weights are predicted by the attention network. The attention network is illustrated in Fig. 3.

The attention mechanism learns which features are relevant for each class and ranks them with relevance accordingly. The output y of the attention mechanism is the weighted sum of the convolutional features extracted by the network. A scoring function $\alpha(f_i; \Theta >)$ should be learned per local feature vector f_i, where $\Theta >$ representing the parameters of the function α. i = 1, …, n the i-th feature vector, and $f_i \in \mathrm{R}^{\mathrm{d}}$, with d denoting the size of the feature vectors. Using cross-entropy as the loss, the output of the network is given by

$$y = W\left(\sum_{n}(1 + a(f_i;\theta)) * f_i\right) \tag{1}$$

where y is the weighted sum of the local features extracted by the attention mechanism. $\mathrm{W} \in \mathrm{R}^{\mathrm{M}\times\mathrm{d}}$ represents the weights of the final fully-connected layer of the CNN trained to predict classes. α is the scoring function. We use cross entropy loss for training, which is given by

$$L = -y^* \cdot \log\left(\frac{\mathrm{e}^{y}}{{f_i}^{T} \cdot \mathrm{e}^{y}}\right) \tag{2}$$

where y* is ground-truth in one-hot representation and f_i is one feature vector. ${f_i}^{T}$ is the transpose of f_i. The parameters in the score function α (·) are trained by backpropagation, where the gradient is given by

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial y} \sum_{n} wf_i \frac{\partial a_i}{\partial \theta} \tag{3}$$

where the backpropagation of the output score $\alpha_i \equiv \alpha(f_i; \Theta >)$ with respect to $\Theta >$ is same as the standard multi-layer perceptron.

In order to prevent score function learning negative weighting, we make α $(\cdot)$ >= 0. And a 2-layers CNN with a softplus activation at the top is designed for the score function. The size of convolutional filters is 1×1, which work well in practice. The trained attention mechanism can be used to assess the relevance of features extracted by our model. This is intuitive because the attention model should be able to generate valid scores for different local feature descriptors. As a result of the training, local feature descriptors implicitly learn more relevant representations of landmark.

3.2 Dimensionality Reduction

We reduce the dimensionality of selected features to obtain improved retrieval accuracy, as common practice [42,43]. Firstly, the selected features were normalized by L_2 normalization, and the dimension of the features was reduced to d by principal component analysis (PCA), which balanced compactness and discriminability. Finally, the features are L_2 normalization again. The effect of PCA based dimensionality reduction technique is shown in Fig. 3.

We observe an interesting trend that directly applying PCA to the feature vectors provides better performance than reducing the dimensionality of the aggregation descriptors. The experimental part has detailed comparison data. This likely reason for this behavior is the removal of irrelevant components present in the original feature vectors. The contributions of these components cannot be completely removed from the descriptor after aggregation.

3.3 Descriptor Aggregation

Iscen et al. [44] proposed an image retrieval algorithm through triangulation embedding (TE) and democratic aggregation (DA). Their method modified some key steps of local feature aggregation and achieves obvious better results. At present, the method is the most effective and valuable among the traditional image retrieval algorithms. Referring to the optimization strategy adopted by their method, we propose an improved method.

In the TE+DA method, all local features participate in the calculation completely fairly. In other words, all the features have the same contribution when it comes to image matching. However, TE+DA is not the most ideal choice for the approximate and completely fair calculation strategy. Hence, in the process of feature extraction, our proposed method not only provides the deep local feature vectors, but also uses the attention mechanism to provide the score information of each local features.

We utilize attention scores to indicate the importance of the features in the image. The main idea of our algorithm proposed in this paper is to distribute the features with high scores to all clustering centers and then make them participate in the residual operation of all clustering centers. The features with low score value are only assigned to the nearest clustering center, and only participates in the residual operation of the nearest clustering center. Moreover, this distribution can have many transformation forms and be adjusted flexibly. Finally, the descriptor of local features aggregation is fused with the global descriptor by weighting to obtain the image representation.

3.3.1 Triangulation Embedding

The purpose of geometric embedding is to make the distance between similar features infinitely close to zero in the image matching process. The vectors obtained in the embedding step should not consider the magnitude of the vector modulus, but only focus on the direction of the vectors. This process can also be regarded as the normalization of residual vectors. Different from other embedding methods, the geometric embedding strategy only uses the orientation information of the feature for feature matching.

The local features after keypoint selection of an image are denoted as $X = \{x_1, x_2, x_3, \cdots x_m\}$. After k-means training, q clustering centers were obtained, which is $C = \{c_1, c_2, c_3, \ldots\ldots c_q\}$, $c_j \in R^d$. The calculation process of geometric embedding is as the formula (4)

$$r_j(x) = \left\{ \frac{x - c_j}{|x - c_j|} \right\}, j = 1 \cdots k \tag{4}$$

where c_j represents the *j-th* clustering center and $r_j(x)$ is the residual vector obtained by geometric embedding.

The detail process of triangulation embedding is shown in Fig. 4. Fig. 4a represents the local features of the image, and it is assumed that there are four clustering centers. Taking any feature of the image as an example, Fig. 4b represents the residual between the current feature and all clustering centers. The size of residual vector obtained at this time varies according to the distance. In the process of image matching, the size of residual affects the matching between features. Fig. 4c represents the normalization operation of the residual, which is usually performed by L_2 normalization. In this way, the residual vector obtained does not consider the magnitude of the vector module, but only considers the direction information.

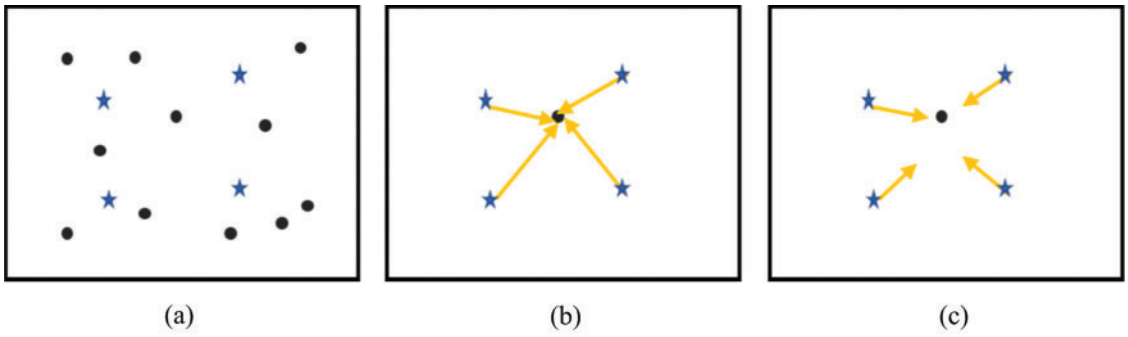

Figure 4: The strategy of triangulation. (a) Represents the local features of the image, and it is assumed that there are four clustering centers. (b) Represents the residual between the current feature and all clustering centers. (c) Represents the normalization operation of the residual, which is usually performed by L_2 normalization

Then, we can get q residual vectors by geometric embedding $r_j(x)$. You spliced these residuals to get the residuals matrix $R(x) = |r_1(x)^T, r_2(x)^T, \ldots\ldots r_q(x)^T|^T$. Dimension of local features is d, then the dimension of $R(x)$ is $D = |q| * d$.

3.3.2 Attention Aggregation

The attention aggregation operation is sum-pooling usually made after the feature embedding. When judging the similarity of two descriptors, it is usually used as a kernel matrix [45] like formulas (5), (6)

$$K(X, Y) = \sum_{x \in X} \sum_{y \in Y} k(x, y) = \psi(X)^T \psi(Y) \tag{5}$$

$$\psi_s(X) = \sum_{x \in X} \emptyset(x) \tag{6}$$

where the output K is kernel matrix. $\emptyset : R^d \to R^D$, R^d represents a space with low-dimensional vector and R^D represents a higher-dimensional vector. And $X = \{x_1, \ldots, x_m\}$, $Y = \{y_1, \ldots, y_m\}$ represent the descriptors of two different images. And k(x, y) is the matching kernel, which represents the inner product of the image features after embedding. Then $k(\mathrm{x}, \mathrm{y}) = \langle \varphi(x) | \varphi(y) \rangle$. The expression of the kernel matrix is shown as formula (7)

$$K(X, Y) = \sum_{x \in X} \sum_{y \in Y} \varphi(x)^T \varphi(y) \tag{7}$$

where the image feature after embedding is $\varphi(y)$ and $\varphi(x)^T$ is the transpose of $\varphi(y)$.

The purpose of the attention aggregation is to make the contribution of each feature different in the image matching. Therefore, the realization of the attention aggregation is making each feature vector multiply the attention score of the vector. The kernel matrix after the aggregation is shown in formula (8)

$$K(X, Y) = \sum_{x \in X} \sum_{y \in Y} \lambda_X(x) \lambda_Y(y) k(x, y) \tag{8}$$

where $\lambda_X(x) = \alpha(f_i; \Theta >)$, $x \in X$. $\lambda_Y(y) = \alpha(f_j; \Theta), y \in Y$. k(x, y) is the matching kernel. And $\lambda_X(\cdot)$ and $\lambda_Y(\cdot)$ are always greater than 0.

The improved aggregation algorithm using local features based on attention mechanism is pseudocode as follows.

Algorithm 1: The improved aggregation algorithm using local features based on attention mechanism

Input: 1. Input image: I;
2. The local feature set of an image and the attention score of each feature, A total of m;
3. K-means algorithm training clustering center, a total of q;

Output: A aggregated descriptor based on local features of the attention mechanism;

1. *Fea_s* = *zeros_matrix*($d * q$), m); % The matrix *Fea_s* represents the descriptors of all images.
2. For each image I.
3. idx = *yeal_*(C, X); % Assign the feature to the closest cluster center.
4. $r_i = \| x_i - c_j \|$, $j = 1 \ldots q$; % Any $x_i \in X$, Calculate the distances between x_i and q different clustering centers.
5. $R = [r_1, r_2, \ldots r_m]$; % The set of all distances.

(Continued)

Algorithm 1 (Continued)

For $ii = 1 \ldots m$ % m is the number of local feature set of an image.
  If The attention score of the current feature >= The set attention score threshold:
    $YS(:, ii) = r_i$; % Features do residual arithmetic with all clustering centers.
  else the attention score of the current feature < The set attention score threshold:
    Calculating the residual between the feature and the cluster center, which is closest to the feature.
Then get Ys(jj, ii); % $jj \in ((idx(ii) - 1) * d + 1, idx(ii) * d)$. $idx(ii)$ is the clustering center closest to local feature ii.
  End
End
$YS = n(YS)$; % L2 normalized residual vectors that only retained the orientation information.
$Y_{saliency} = sum(YS, 2)$; % Add the residual vectors.
$Fea\_s(:, I) = Y_{saliency}$;
END
Return Fea_s; % Fea_s is output. It is aggregated descriptor based on attention local features.

The proposed method improves the aggregation of local features by the attention scores of local features, and determines the calculation strategy of local features with the clustering centers through the preset threshold. After setting the threshold value for attention score, if the attention score of a local feature exceeds the threshold value, it indicates that the feature comes from the relatively core region of the image and is well distinguishable. At this time, it is necessary to calculate the residual between the local features and all clustering centers. Then, we will get the aggregated local feature descriptor based on attention mechanism by summing up the residual vectors. Different from TE+DA algorithm, the contribution of local features in image matching is not absolutely fair, but varies according to the attention score of local features.

At last, the global feature output from the fully-connected layer and the aggregated local feature descriptors based on the attention scores are fused together to obtain the image presentation.

3.4 Image Retrieval Algorithm

We focus on descriptor aggregation and landmark image retrieval by using the proposed methods. Before we can retrieve the image retrieval, we must process all the image data in the landmark datasets to build a feature index library. Algorithm 2 is the overall algorithm structure of the method we proposed in this paper.

Algorithm 2: Image retrieval algorithm

Input: Query image and dataset images;
Output: Top T ranked images list from dataset according to image similarity to query image;
1: Extract global and local feature descriptors respectively;
2: Select keypoints from local features according to the values from score function $\alpha(\cdot)$;
3: Reduce dimension of the features after selected to d;
4: Aggregate local descriptors and global descriptor to produce image representation;
5: Compute a set of similarities between query image and dataset images based on image representation;
6: Rank and select the top T images in dataset according the collection of similarities;
7: **Return** the ranked images list.

4 Experiments

In this section, we give the datasets and evaluation metrics used for evaluation purpose, the experimental environment and results, and evaluation metrics. All the experiments were performed on Ubuntu18.04 operating system with 3.4 GHz Intel(R) Core(TM) i7-4930 K CPU, 32 GB main memory and GTX 2080 GPU. We first discuss the performance of our features descriptors compared to existing global and local feature descriptors in the datasets. Second, feature descriptor aggregation is used to evaluate the quality of our proposed model, which is trained on the new dataset. Finally, the image retrieval system is enhanced by using different dimensionality reduction methods (descriptor reduction before/after aggregation). The training dataset we used is the Google Landmark Dataset.

4.1 Datasets

Our experiment used three datasets. There are Google Landmark dataset (GLD), Revisited Oxford (ROxf), and Revisited Paris (RPar). Among them, ROxfand RPar are updated versions based on the Oxford and Paris datasets.

Google Landmark Dataset: Google-Landmarks is the largest dataset of artificial and natural Landmarks currently. The dataset contains 2 million images describing 30,000 unique landmarks around the world, which is 30 times larger than the average dataset. There is only one corresponding landmark label for each image in the Google Landmark image, and the dataset is shown in Fig. 5.

Figure 5: Some landmarks in the Google Landmark Dataset

4.2 Evaluations Metrics

One of famous evaluation metrics for image retrieval systems is mean Average Precision (mAP), which is obtained by ranking the images in descending order of relevance for each query and calculating the AP for each query. However, this assessment method is not representative for datasets with intrusive queries because it is important to determine whether each image is relevant to the query. For performance evaluation, we used Precision-Recall curve, while taking

into account all the query images. PR curve is a curve made with precision and recall as variables, in which recall is the abscissa and precision is the ordinate.

$$Precision = TP/(TP + FP) \tag{9}$$

$$Recall = TP/(TP + FN) \tag{10}$$

where TP (True Positive) represents the correct classification of positive examples as positive examples; FN (False Negative) means to misclassify a positive example as a negative one; FP (false positive) means that negative cases are misclassified as positive ones.

4.3 Experimental Results of Proposed Method with Different Mechanisms

According to the proposed method in the third part, the aggregation strategy of local features can be changed by setting the significance threshold. In the significantly improved aggregation algorithm, the threshold value determines whether the features are allocated to all clustering centers or only to the clustering centers closest to the current features. The setting of the threshold value depends on the maximum value of the significance of all local features in the image, and the threshold can be obtained by multiplying the maximum value by the threshold parameter.

In order to select the most appropriate threshold, an image retrieval experiment was conducted in different clustering centers and the threshold with the best effect was selected. In order to observe the experimental effect of the improved algorithm more comprehensively, we extract up 1,000 local features from each image, T = 60, and 16, 32, 64, 128 clustering centers were used in this section for the experiment. Fig. 6 shows the results of our experiment.

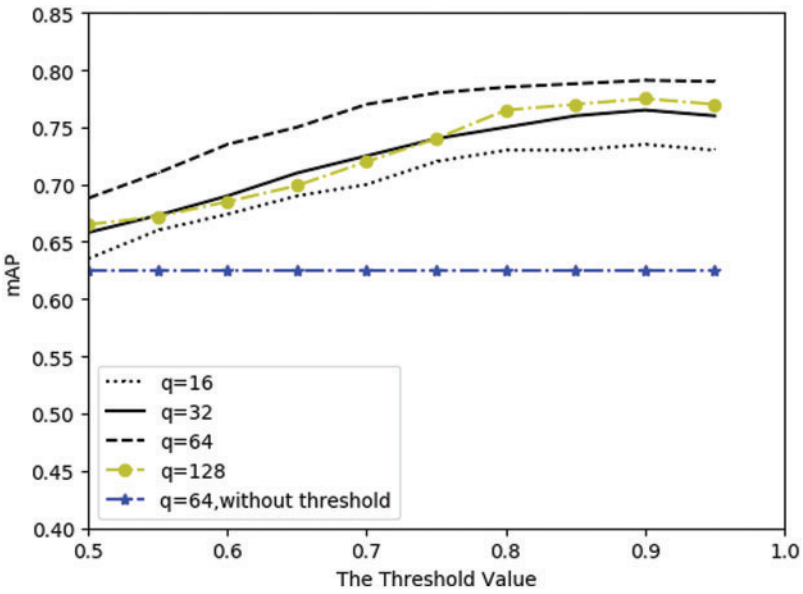

Figure 6: mAP values of different attention score thresholds in the Google Landmark Dataset. The mAP value of our method remained stable after the threshold parameter was 0.88

Under 64 clustering centers, the mAP value of the algorithm in this paper is 0.791. As can be seen in the Fig. 6, the mAP values of the algorithm are improved with different parameters. Compared with the traditional method, the mAP value remained stable after the threshold

parameter was 0.88, and the mAP value was 0.791. In order to achieve a stable experimental effect, the threshold parameter is set to 0.9 in the subsequent experiments in this paper.

In addition, the presented system also has an important parameter, the size of original local feature vectors d, that affect the final features dimensionality. We further tested the effects of this parameter on the retrieval accuracy and reduced dimension of the aggregated feature vectors to 1024, 256, 128, 64, 32 and 16. To test the robustness of the model, we repeated the retrieval experiment on ROxf and RPar datasets and show the results in Tables 1, 2.

Table 1: Mean retrieval precision using lower-dimensional descriptors. PCA is applied to local features and global features

	Datasets			
	ROxf		RPar	
Setup	LOC-ATT+ GLO+AGG	LOC-ATT+ AGG	LOC-ATT+ GLO+AGG	LOC-ATT+AGG
q = 64, d = 1024	0.856	0.944	0.864	0.854
q = 64, d = 256	0.921	0.923	0.912	0.903
q = 64, d = 128	0.937	0.949	0.923	0.925
q = 64, d = 64	**0.967**	0.951	**0.952**	**0.947**
q = 64, d = 32	0.942	**0.957**	0.931	0.939
q = 64, d = 16	0.877	0.862	0.869	0.873
q = 32, d = 128	0.933	0.919	0.890	0.819
q = 32, d = 64	0.955	**0.946**	**0.952**	**0.937**
q = 32, d = 32	**0.965**	0.939	0.943	0.931
q = 32, d = 16	0.943	0.928	0.937	0.931

Table 2: Mean retrieval precision with aggregated descriptors using lower-dimensional descriptors. PCA is applied to reduce the size of the aggregated descriptors

	Datasets			
	ROxf		RPar	
Aggregated descriptor size	LOC-ATT+ GLO+AGG	LOC-ATT+ AGG	LOC-ATT+ GLO+AGG	LOC-ATT+AGG
2048	0.921	0.883	0.902	0.894
1024	0.897	0.879	0.883	0.890
512	0.905	0.887	0.891	0.883
256	0.914	0.916	0.891	0.897
128	0.929	0.848	0.920	0.901
64	**0.940**	0.889	**0.925**	**0.919**
32	0.937	**0.922**	0.919	0.907
16	0.926	0.904	0.911	0.894

In the process of reducing the features dimensionality, we observed two interesting phenomena. First, the descriptors with dimensionality reduction using PCA have better performance than those without dimensionality reduction. We also found that the retrieval accuracy of using dimensionality reduction before descriptor aggregation (the number 0.965 in Table 1) was higher than using dimensionality reduction after descriptor aggregation (the number 0.940 in Table 2). The reason for this phenomenon is that some features with unimportant information before aggregation can be deleted, but using dimensionality reduction after descriptor aggregation cannot obtain that effect. In further experiments, we found that the complete descriptor was not the best choice. Using 32 or 64 dimensional descriptor have achieved the best results.

Dimensions of features above 1024 seems to contain a large amount of irrelevant data. It can be concluded from those Tables that the benefits of having eigenvectors of sufficient size is significant. By using smaller descriptors, we achieved better retrieval performance, increasing the retrieval efficiency on ROxf and RPar datasets by 11.1% and 8.8% respectively by aggregated local features with attention mechanism combined with global feature. At the same time, our method improved the accuracy by 1.13% and 0.93% in ROxf and RPar datasets, respectively.

To analyze the benefit of feature descriptor aggregation and global feature for image retrieval, we compare our full model (LOC-ATT+AGG+GLO) with its variations: LOC-ATT and LOC-ATT+AGG. LOC-ATT means that extracted local features based on attention mechanism but without features aggregation and without global features; LOC-ATT+AGG denotes that the model extracted local features based on attention mechanism with features aggregation but without global feature; LOC-ATT+AGG+GLO means aggregated local features with attention mechanism combined with global feature. Fig. 7 shows that the precision-recall curves of comparison experiment on large-scale Google Landmark dataset. The experiment result indicates that features aggregation and global feature in our approach are critical to performance improvements.

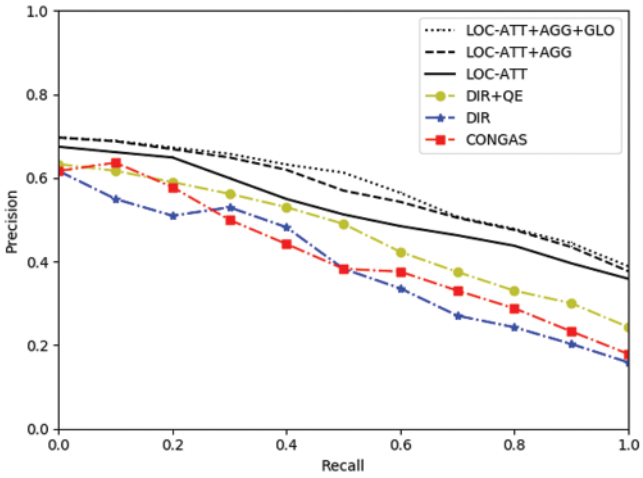

Figure 7: Precision-recall curve for the image retrieval experiment. As can be seen, our method can improve the performance of the proposed LOC-ATT model effectively

In particular, the use of feature aggregation is more important than the global feature. Our method and CONGAS are almost as complex as deep image retrieval (DIR) in terms of memory requirements. Our method and CONGAS adopt the same feature dimension and the maximum

feature number of each image. They require about 12 GB of memory. The directory descriptor requires 16 KB per image, which adds up to about 12 GB to index the entire dataset.

4.4 Experimental Results Compared with other Methods

In this experiment, we trained the model using the training set of the Google Landmark dataset, and then tested the experiment on the Google Landmark test set and calculated the mAP values for each model. Models for comparison include R-VLAD, ASMK, DIR, and CONGAS. We set q (=64) clustering centers for aggregating features, and set the attention score threshold is 0.9.

We further analyze the mAP results for each model counted on the test set of Google Landmark dataset, as shown in Table 3. Under the Google Landmark dataset, the mAP values of our model were 0.038, 0.039, 0.053, and 0.047 higher than the other four models, respectively, and R-VLAD performed best of 0.053 of the four network models. The mAP values of DIR and CONGAS are 0.521 and 0.527, respectively, which corresponds to the performance results shown by the P-R curve (Fig. 7). However, none of these four networks performed better than our model (0.574). Through experimental data analysis, it can be found that the model presented in this paper can better suppress background noise interference.

Table 3: The mAP value of each model on the Google Landmark dataset

Models	Google Landmark (mAP)
Ours	0.574
R-VLAD	0.536
ASMK	0.535
DIR	0.527
CONGAS	0.521

In order to explain more intuitively our model can improve the accuracy of image retrieval under the condition of noise influence such as partial occlusion, we randomly select several retrieval pictures from the experimental results. Fig. 8 is a comparison output of the model in this paper with other public models, R-VLAD, ASMK, DIR, and CONGAS. (a) in Fig. 8 is the input image, and each model retrieves the image most similar to the input image in the Google Landmark test set. (b) is the most similar output of our model. (c) (d) (e) and (f) are the outputs of the other four models. Fig. 8 consists of three rows of images that represent the results of each model under three partial occlusion scenarios in the Google Landmark test set. The foreground in the first line of images is cedar, and behind the cedar is the building. The second row is the building occluded by trees, but compared with the image of the first row, the building of the second row image is disturbed by the sky. The third line is problem of low precision detection of small-scale objects. The experimental results in Fig. 8 show that the AP values of our model are 0.684, 0.633, and 0.662, respectively, which are higher than each other.

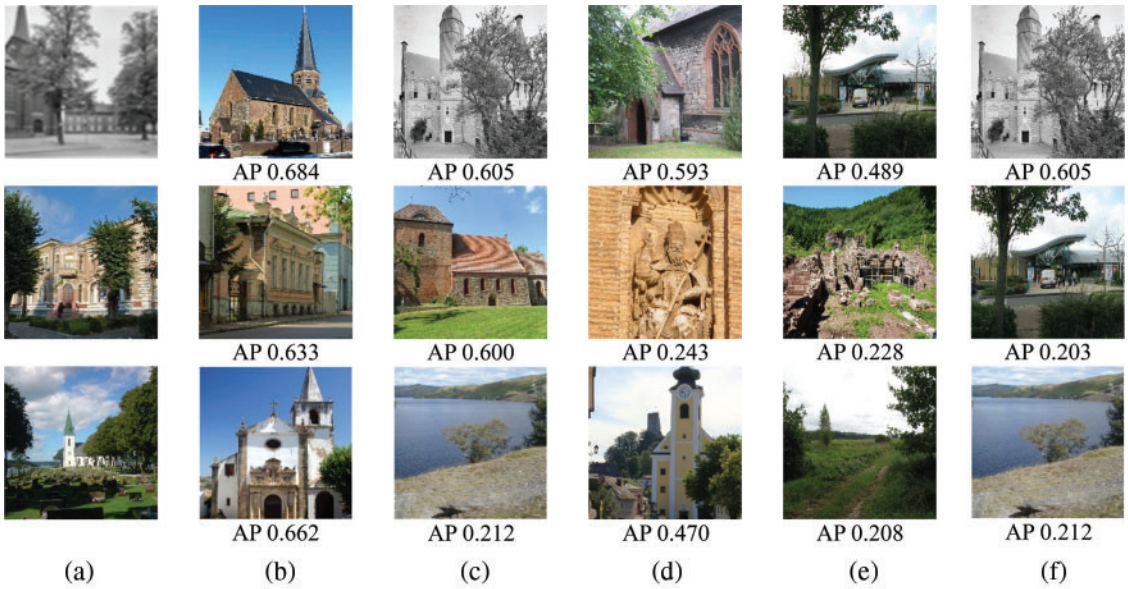

Figure 8: Retrieval results for Ours, R-VLAD and ASMK in Google Landmark data. (a) is query image (input), (b) is top-1 output of LOC-ATT+AGG+GLO, (c) (d) (e) (f) are top-1 output of other method. As can be seen from the specific retrieval examples, even for buildings from different angles, our retrieval results are better than R-VLAD and ASMK (a) Query (b) Ours (c) R-VLAD (d) ASMK (e) DIR (f) CONGAS

As we can see from the output of the first row, except for our model, all of other models are greatly affected by the foreground (cedar) because the results of all three model outputs contain trees. The image output by our model excludes the semantic information of the foreground, which further proves the superiority of our model. In the experimental results of the second line, due to the lack of prominent semantic information of the building target, the last three models have been seriously misjudged and the images unrelated to the input images have been retrieved. And, in the third row, other models have also appeared the wrong output. Through these three sets of experimental outputs, it can be found that our model can output more correct results under a variety of noise influence conditions, which proves that our model has the effect of suppressing background noise and good robustness.

Table 4 and Fig. 9 show the results of comparing different aggregation methods a on the ROxf and RPar datasets. We compared our method against other region selection aggregation methods: Regional Vector of Aggregate Locally Descriptor (R-VLAD) and Aggregated Selective Match Kernel (ASMK).

For the R-VLAD, the retrieval accuracy of the R-VLAD improved to 0.426 on the use of the ROxf dataset. When the detection area of the method in the R-VLAD method is set to 6, the retrieval accuracy of the model will fall sharply. This also proves that the retrieval accuracy of the R-VLAD is invert with the detection area under certain conditions. Comparing our proposed technique against with R-VLAD, our method obtains the best results. In the Table 3, our method improves the retrieval accuracy by 0.147 over the R-VLAD method.

Table 4: Comparison of proposed technique against R-VLAD and ASMK on ROxf and RPar datasets. Both R-VLAD and ASMK are regions-based aggregation methods, they get the best results when the number of regions is 3 or 6 on ROxf and Rpar datasets. Note that the method LOC-ATT-AGG-GLO is not based on the detected regions, so the region setting does not affect the mAP

Method	ROxf (mAP)		RPar (mAP)	
	3 regions	6 regions	3 regions	6 regions
LOC-ATT+AGG+GLO	0.574	0.574	0.559	0.559
R-VLAD	0.426	0.353	0.409	0.306
ASMK	0.351	0.476	0.314	0.441

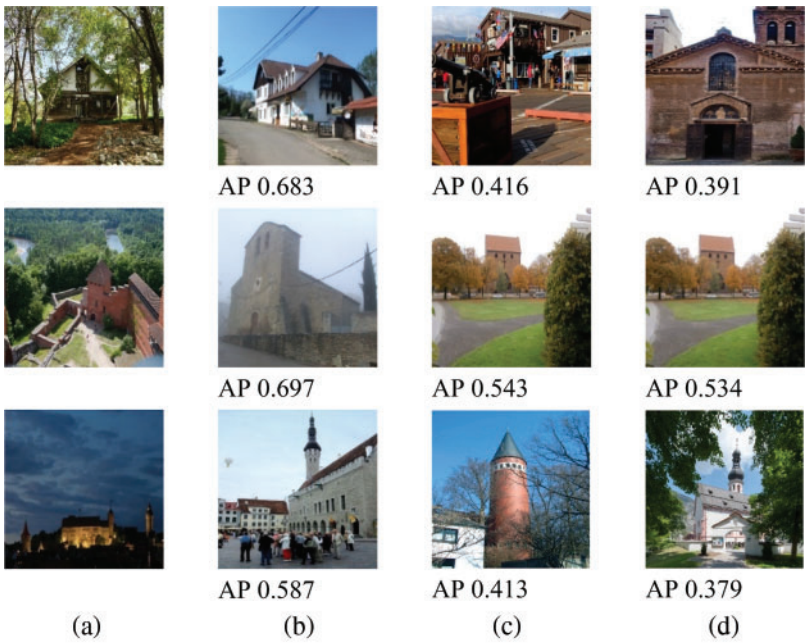

Figure 9: Retrieval results for Ours, R-VLAD and ASMK. (a) is query image (input), (b) is top-1output of LOC-ATT+AGG+GLO, (c) is top-1 output of R-VLAD, (d) top-1 image of ASMK. As can be seen from the specific retrieval examples, even for buildings from different angles, our retrieval results are better than R-VLAD and ASMK (a) Query (b) Ours (c) R-VLAD (d) ASMK

For ASMK, at first, there is only a slight improvement in our method. The mAP is 0.314 on ROxf dataset with 3 detected regions. When 6 regions per image were used for ASMK, the mAP increase to 0.476 and 0.441 on RPar and ROxf, respectively. By contrast, due to our attentional mechanism combined with global feature, our algorithm still leads ASMK algorithm by 0.117, which use different number of areas. The curves in the left of Fig. 10 are comparisons of ASMK, R-VLAD and LOC-ATT +AGG+GLO methods. As can be seen, our method outperforms the other two feature aggregation methods, among which the ASMK method performs the worst.

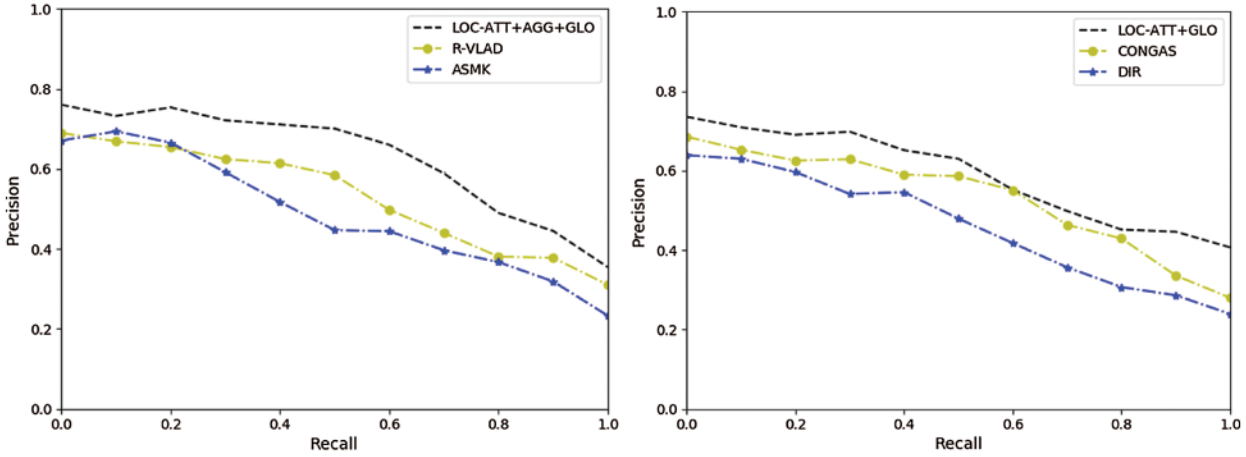

Figure 10: Precision-recall curve of our and the compared methods. As can be seen, our method outperforms the other two feature aggregation methods

We also compared other global feature methods with the proposed method on the ROxf and RPar datasets. The results of comparative experiments are in Table 4 and Fig. 9. The Precision-Pecall Curves of right side in Fig. 10 are the comparisons of DIR, CONGAS and LOC-ATT+GLO. Since DIR and CONGAS use global features, we compared these two methods with LOC-ATT+GLO in this experiment.

Our method *vs.* Deep Image Retrieval (DIR): Our method is superior to DIR as shown in Fig. 11. We used the released source code to implement the version. For retrieval, a parallel exhaustive search is used to avoid the error penalty of nearest neighbor search for DIR. Our method, aggregated local features with attention mechanism combined with global feature, achieves a good match between discriminating descriptors in the image, which is of great help in finding the same target under different imaging conditions. In many cases, DIR matches similar images, but it does not distinguish between these specific targets, which often do not match the instances of interest. But our method achieves a good result. However, DIR is better than LOC-ATT, when the ground or vegetation on different landmarks is similar. So, when we add global feature technology, the model output at this point is better than DIR, also further demonstrating the importance of global feature technology.

Our method *vs.* CONGAS: Our method retrieves more relevant information, indicating that our method descriptor is more discriminative. In the experimental results, there is no example that CONGAS is superior to the retrieval results of our method. Table 5 shows the mAP of the query against the database. The acceptance field of CONGAS model is generally quite large. Therefore, some features in the CONGAS model are greatly affected by undifferentiated regions, such as an image with a large number of white clouds. In these cases, however, the features extracted by our method take into account more discriminative areas in image.

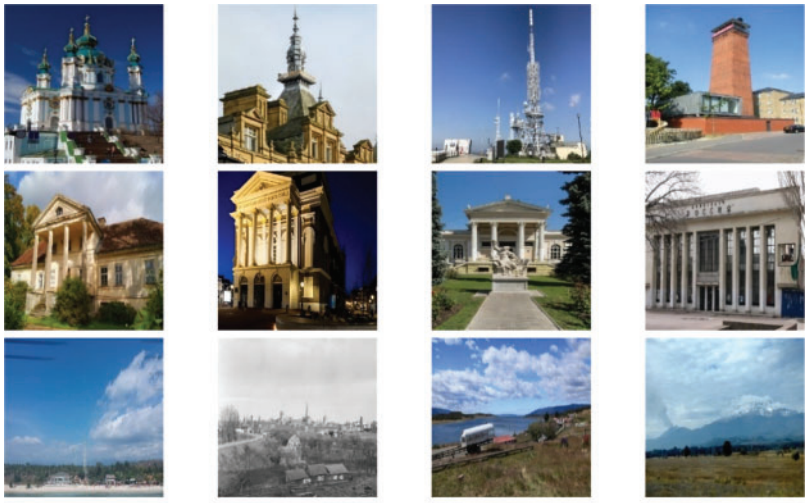

Figure 11: The image retrieval results. (a) Input image, (b) the output of our proposed method, (c) the output of CONGAS, (d) the output of DIR. In the experiment, we found that Our method retrieves more relevant information, indicating that our descriptor is more discriminative. However, the results of other methods only based on deep global feature are incorrect usually (a) Query (b) Ours (c) CONGAS (d) DIR

Table 5: The mAP of the different methods in this work

Method	ROxf dataset	RPar dataset
	mAP	mAP
LOC-ATT+GLO	0.446	0.549
DIR	0.374	0.409
CONGAS	0.437	0.528

4.5 Model Performance Analysis

Finally, we investigate the computational complexity of the proposed model. We first tested our proposed model to extract local and global descriptors from an image in 23 milliseconds. Under the condition of $T = 16$, k-means clustering of selected features takes about 7 min. Under the condition of $T = 64$, it takes about 31 min to perform k-means clustering on the selected features. Therefore, under the conditions where local features and codebook can be used. We also measure the amount of time of that consumed to generate an aggregation descriptor. After the measurement, we concluded that it would take about 30 milliseconds to aggregate an effective descriptor. Experiments show that constructing codebook and descriptor aggregation requires more computation time than feature extraction. In order to ensure the validity of the experimental data, we tested other methods in the same experimental environment.

The time spent on the whole image retrieval is determined by the size of dataset and the descriptor size. First, we have to traverse the entire dataset to extract and generate feature for each image. Second, the longer the descriptor is, the longer the computation time is required. We followed the procedure [46] to compare the computational complexity of different descriptor sizes under different database sizes. These results are shown in Table 6. In addition, we compare the

proposed method with several other methods. Experimental results show that better results can be obtained by using our descriptor. It can also be found that the execution time decreases with the reduction of feature dimension, but the correlation is nonlinear. For example, the dimension is reduced from 1,024 to 512, and the execution time is reduced by 3,448 ms. When the dimension is reduced from 128 to 64, the execution time is only reduced from 1,976 to 1,485 milliseconds. The application scenario of this paper is mainly about building image retrieval, although the proposed model can be applicable in other scenarios. We also use the more large-scale dataset and test set image of CIFAR for retrieval, and because the CIFAR dataset has relatively few pixels (32×32, the resolution of GLD is much higher than 32×32), the speed increases during the retrieval process.

Table 6: Time consumed in image retrieval (in millisecond) in different database and descriptor sizes

Dataset size (number of images)	R-VLAD	ASMK	DIR	Ours (d = 1024)	Ours (d = 512)	Ours (d = 128)	Ours (d = 64)
100	431	247	307	408	246	177	113
200	832	417	583	1064	720	489	249
500	1331	1150	1213	3789	2311	1192	795
1000	1691	2083	1769	7921	4473	1976	1485
2000	4592	5004	4741	14498	11946	7023	4527
5000	12267	14987	14098	46287	40998	30166	11983
10,000(CIFAR)	3298	4462	3754	8984	6528	4067	3319

5 Conclusions

In this paper, we presented an efficient method of feature descriptor aggregation with an attention mechanism combined with a global feature for image retrieval. We first trained the local feature extraction network and then the attention mechanism to select key local features. Both processes were accomplished by a single forward propagation way in the network. Also, we proposed a novel feature aggregation approach with the proposed method. Experiments showed that local features aggregation with attention mechanism combined with the global feature could effectively improve the image retrieval accuracy and computational efficiency. It is noted that the proposed method is not only applicable to landmark image search but can also be applied to image retrieval tasks in the general domain if the search candidate images carry abundant, legible local features.

Funding Statement: This research is jointly supported by the National Natural Science Foundation of China (62072414, U1504608, 61975187), and the Foundation and Cutting-Edge Technologies Research Program of Henan Province (212102210540, 192102210294, 212102210280).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Jian, M., Yin, Y., Dong, J., Lam, K. M. (2018). Content-based image retrieval via a hierarchical-local-feature extraction scheme. *Multimedia Tools and Applications, 77(21),* 29099–29117. DOI 10.1007/s11042-018-6122-2.
2. Radenovi, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O. (2018). Revisiting Oxford and Paris: Large-scale image retrieval benchmarking, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, Utah, IEEE.
3. Abioui, H., Idarrou, A., Bouzit, A., Mammass, D. (2018). Review: Automatic image annotation for semantic image retrieval. *International Conference on Image and Signal Processing*. Springer, Cham.
4. Angkoon, P., Rami, N. K., Erik, S. (2018). Feature extraction and selection for myoelectric control based on wearable emg sensors. *Sensors, 18(5),* 1615–1626. DOI 10.3390/s18051615.
5. Shi, Q., Cheung, Y. M., Zhao, Q. (2018). Feature extraction for incomplete data via low-rank tensor decomposition with feature regularization. *IEEE Transactions on Neural Networks & Learning Systems, 30(6),* 1803–1817. DOI 10.1109/TNNLS.5962385.
6. Mei, S., Jiang, R., Ji, J., Sun, J., Zhang, Y. (2018). Invariant feature extraction for image classification via multi-channel convolutional neural network. *International Symposium on Intelligent Signal Processing and Communication Systems.* pp. 491–495. Okinawa prefecture, Japan, IEEE.
7. Krestinskaya, O., James, A. P. (2018). Feature extraction without learning in an analog spatial pooler memristive-cmos circuit design of hierarchical temporal memory. *Analog Integrated Circuits & Signal Processing, 95(3),* 457–465. DOI 10.1007/s10470-018-1161-1.
8. Gu, F. F., Fu, M. H., Liang, B. S., Li, K. M., Zhang, Q. (2018). Translational motion compensation and micro-Doppler feature extraction of space spinning targets. *IEEE Geoscience and Remote Sensing Letters, 15(10),* 1550–1554. DOI 10.1109/LGRS.2018.2849869.
9. Liu, Y., Chen, C., Yuan, Y., Tong, K. Y. (2019). *Selective feature aggregation network with area-boundary constraints for polyp segmentation,* pp. 302–310. Cham: Springer.
10. Chaudhury, S., Ozaki, H., Kimura, D., Vinayavekhin, P., Kidokoro, S. (2019). Unsupervised temporal feature aggregation for event detection in unstructured sports videos. *IEEE International Symposium on Multimedia*. pp. 9–97. Cosenza, Italy, IEEE.
11. Qin, F., Lin, S., Li, Y., Bly, R. A., Moe, K. S. et al. (2020). Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision. *IEEE Robotics and Automation Letters, 5(4),* 6639–6646. DOI 10.1109/LSP.2016.
12. Nam, S. H., Kim, W. H., Mun, S. M., Hou, J. U., Choi, S. et al. (2018). A sift features based blind watermarking for DIBR 3d images. *Multimedia Tools & Applications, 77(7),* 7811–7850. DOI 10.1007/s11042-017-4678-x.
13. Tareen, S. A. K., Saleem, Z. (2018). A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. *International Conference on Computing, Mathematics and Engineering Technologies*, Wuhan, China, IEEE.
14. Shi, X., Xing, F., Xu, K., Chen, P., Liang, Y. et al. (2020). Loss-based attention for interpreting image-level prediction of convolutional neural networks. *IEEE Transactions on Image Processing, 30,* 1662–1675. DOI 10.1109/TIP.83.
15. Zhang, H., Zhang, T., Pedrycz, W., Zhao, C., Miao, D. (2019). Improved adaptive image retrieval with the use of shadowed sets. *Pattern Recognition, 90,* 390–403. DOI 10.1016/j.patcog.2019.01.029.
16. Sedaghat, A., Mohammadi, N. (2018). Uniform competency-based local feature extraction for remote sensing images. *Journal of Photogrammetry & Remote Sensing, 135,* 142–157. DOI 10.1016/j.isprsjprs.2017.11.019.
17. Kumar, K., Chauvin, L., Toews, M., Colliot, O., Desrosiers, C. (2018). Multi-modal analysis of genetically-related subjects using SIFT descriptors in brain MRI. *Computational Diffusion MRI,* pp. 219–228. Springer, Cham.
18. Yu, L., Song, J., Ke, Z. (2018). Deep self-taught hashing for image retrieval. *IEEE Transactions on Cybernetics, pp(99),* 1–13. DOI 10.1109/TCYB.2018.2822781.

19. Ma, J., Yuan, Y. (2019). Dimension reduction of image deep feature using PCA. *Journal of Visual Communication and Image Representation, 63,* 102578. DOI 10.1016/j.jvcir.2019.102578.
20. Mukherjee, M., Meenpal, T. (2019). Kinship verification using compound local binary pattern and local feature discriminant analysis. *10th International Conference on Computing, Communication and Networking Technologies*, pp. 1–7. Antofagasta, Chile, IEEE.
21. Braux-Zin, J., Dupont, R., Bartoli, A. (2013). A general dense image matching framework combining direct and feature-based costs. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 185–192. Sydney, Australia, IEEE.
22. Han, Z., Lu, H., Liu, Z., Vong, C. M., Liu, Y. S. et al. (2019). 3D2SeqViews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation. *IEEE Transactions on Image Processing, 28(8),* pp.3986–3999. DOI 10.1109/TIP.83.
23. Kumar, Y., Ai, T., Li, W., Yang, M., Feng, Y. (2019). A polygon aggregation method with global feature preservation using superpixel segmentation. *Computers, Environment and Urban Systems, 75,* 117–131. DOI 10.1016/j.compenvurbsys.2019.01.009.
24. Li, Y., Lei, H., Lin, S. (2019). A new sketch-based 3D model retrieval method by using composite features. *Multimedia Tools and Applications, 77(2),* 2921–2944. DOI 10.1007/s11042-017-4446-y.
25. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q. (2017). Glad: Global-local-alignment descriptor for pedestrian retrieval. *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 420–428. Hong Kong, China.
26. Azizpour, H., Sharif Razavian, A., Sullivan, J., Maki, A., Carlsson, S. (2015). From generic to specific deep representations for visual recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36–45. Boston, Massachusetts, USA.
27. Hu, P., Peng, D., Guo, J., Zhen, L. (2018). Local feature based multi-view discriminant analysis. *Knowledge Based Systems, 149,* 34–46. DOI 10.1016/j.knosys.2018.02.008.
28. Mishkin, D., Radenovic, F., Matas, J. (2018). Repeatability is not enough: Learning affine regions via discriminability. *Proceedings of the European Conference on Computer Vision*, pp. 284–300. Munich, Germany.
29. He, Z., Li, Y., Deng, L., Li, P., Shi, X. et al. (2019). A new two-stage image retrieval algorithm with convolutional neural network. *Proceedings of the 2019 8th International Conference on Networks, Communication and Computing*, pp. 98–102.
30. Bala, A., Kaur, T. (2016). Local texton xor patterns: A new feature descriptor for content-based image retrieval. *Engineering Science and Technology, an International Journal, 19(1),* 101–112. DOI 10.1016/j.jestch.2015.06.008.
31. Cao, B., Zhao, J., Yang, P., Gu, Y., Muhammad, K. et al. (2019). Multiobjective 3-D topology optimization of next-generation wireless data center network. *IEEE Transactions on Industrial Informatics, 16(5),* 3597–3605. DOI 10.1109/TII.9424.
32. Hou, Y., Zhang, H., Zhou, S. (2017). Evaluation of object proposals and ConvNet features for landmark-based visual place recognition. *Journal of Intelligent and Robotic Systems, 92(1),* 1–16. DOI 10.1007/s10846-017-0735-y.
33. Schroff, F., Kalenichenko, D., Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823. Boston, Massachusetts, USA.
34. Lu, H., Nolte, L. P., Reyes, M. (2012). Interest points localization for brain image using landmark-annotated atlas. *International Journal of Imaging Systems and Technology, 22(2),* 145–152. DOI 10.1002/ima.22015.
35. Mikulik, A., Perdoch, M., Ondřej, C. (2013). Learning vocabularies over a fine quantization. *International Journal of Computer Vision, 103(1),* 163–175. DOI 10.1007/s11263-012-0600-1.
36. Mohedano, E., McGuinness, K., O'Connor, N. E., Salvador, A., Marques, F. et al. (2016). Bags of local convolutional features for scalable instance search. *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 327–331. New York, USA.

37. Perd'och, M., Chum, O., Matas, J. (2009). Efficient representation of local geometry for large scale object retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9–16. Florida, USA, IEEE.
38. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H. (2010). Large-scale image retrieval with compressed fisher vectors. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3384–3391. Alaska, USA, IEEE.
39. Philbin, J., Chum, O., Isard, M. (2007). Object retrieval with large vocabularies and fast spatial matching. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 18–23. Minneapolis, Minnesota, USA.
40. Gammeter, S., Bossard, L., Quack, T., Van Gool, L. (2009). I know what you did last summer: Object-level auto-annotation of holiday snaps. *IEEE 12th International Conference on Computer Vision*, pp. 614–621. IEEE.
41. Gordo, A., Almazan, J., Revaud, J., Larlus, D. (2017). End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision, 124(2),* 237–254. DOI 10.1007/s11263-017-1016-8.
42. Hadsell, R., Chopra, S., LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1735–1742. New York, USA, IEEE.
43. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Nevada, USA.
44. Iscen, A., Tolias, G., Avrithis, Y. (2016). Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2077–2086. Nevada, USA.
45. Zhang, J., Yu, J., Tao, D. (2018). Local deep-feature alignment for unsupervised dimension reduction. *IEEE Transactions on Image Processing, 27(5),* 2420–2432. DOI 10.1109/TIP.2018.2804218.
46. Yue-Hei Ng, J., Yang, F., Davis, L. S. (2015). Exploiting local features from deep networks for image retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 53–61. Boston, Massachusetts, USA.