Spatial Pyramid Attention Enhanced Visual Descriptors for Landmark Retrieval

Luepol Pipanmekaporn *, Suwatchai Kamonsantiroj, Chiabwoot Ratanavilisagul, and Sathit Prasomphan

Department of Computer and Information Sciences, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
E-mail: {suwatchai.k, chiabwoot.r, sathit.p}@sci.kmutnb.ac.th
*Correspondence: luepol.p@sci.kmutnb.ac.th (L.P.)

***Abstract*—Landmark retrieval, which aims to search for landmark images similar to a query photo within a massive image database, has received considerable attention for many years. Despite this, finding landmarks quickly and accurately still presents some unique challenges. To tackle these challenges, we present a deep learning model, called the Spatial-Pyramid Attention network (SPA). This network is an end-to-end convolutional network, incorporating a spatial-pyramid attention layer that encodes the input image, leveraging the spatial pyramid structure to highlight regional features based on their relative spatial distinctiveness. An image descriptor is then generated by aggregating these regional features. According to our experiments on benchmark datasets including Oxford5k, Paris6k, and Landmark-100, our proposed model, SPA, achieves mean Average Precision (mAP) accuracy of 85.3% with the Oxford dataset, 89.6% with the Paris dataset, and 80.4% in the Landmark-100 dataset, outperforming existing state-of-the-art deep image retrieval models.**

***Keywords*—deep image retrieval, convolution neural network, feature embedding**

I. Introduction

The process of Content-based Image Retrieval (CBIR) involves finding and retrieving similar images from a database using a given query image [1]. This topic has gained much attention in research communities and has numerous practical applications, such as visual product finding [2], detecting ancient symbols [3], and identifying individuals [4]. CBIR systems generally involve two phases: describing the image's content with an image descriptor and then evaluating the similarity among descriptors to retrieve relevant images for the query.

Landmark retrieval is a specific task of CBIR [5, 6], which focuses on retrieval of the landmark according to a given query image. Despite this, the retrieval task has some specific challenges. Firstly, many landmarks often share similar appearances, such as temple building and religious churches. The second issue involves dramatic variations of landmark images. Many images are presented in a variety of viewpoints, scales and illuminations or background clutter. Fig. 1 illustrates challenges of landmark images. Consequently, these problems make it more difficult than other CBIR tasks.

Image Viewpoints of Tomb of Napoleon Bonaparte

Dramatic illuminations of All Souls College

Figure 1. Variation in landmark images.

To build accurate image retrieval systems, a bunch of previous works have focused on conventional image retrieval methods such as extraction of low-level features in region of interest (ROI) [7], bag-of-visual words descriptors based on SIFT [8], VLAD [9] and Fisher Vector (FV) [10]. Some of these work have used Convolutional Neural Networks (CNNs) to gain the retrieval performance by learning hierarchical features for image representations [11, 12].

The prevalent usage of CNN features is that they have strong generalization and capture semantic relatedness among pixels. The most straightforward CNN-based method relies on derived activations of convolutional or fully-connected layers. For instance, a certain performance gain in the retrieval on benchmarks is achieved by using features from FC6 layer in AlexNet. Babenko [13] has demonstrated that max-pooled activations of AlexNet's Conv5 layer generate the best image descriptor for the retrieval task. Recently, some of the methods have focused on aggregate operations of feature maps in the last convolutional layer to generate a compact global

Manuscript received March 15, 2023; revised April 10, 2023; accepted May 8, 2023.

descriptor such as NetVLAD [14], Sum Pooling [15], Regional Maximum Activations of Convolution (R-MAC) [16] and Generalized Mean (GeM) [17]. Some of these works have focused on enhancement of deep features by training the CNN network using ranking losses [18], such as contrastive loss and triplet loss, which use the two or three identical CNNs sharing their weight. Nevertheless, the quality of learned features is still challenging in landmark retrieval due to the large variations of landmark images.

In this paper, a novel CNN-based framework to landmark retrieval is present. We propose a pooling layer, named Spatial-Pyramid Attention (SPA) that can be incorporated into any CNN networks to learn a compact discriminative representation for the retrieval task.

The main idea is to combine the following three trainable modules: spatial feature extraction, feature attention and feature aggregation together. First, multi-scale pyramid pooling [19] is adopted to capture hierarchical regional structure of images. This region-based method encodes feature maps in a base CNN to obtain regional features with different scales. After that, the extracted features are intensified by using attention mechanism to highlight their informative region. Then, the regional features are aggregated to produce a compact global descriptor. Finally, the network model is optimized by utilizing triplet ranking loss.

The main contributions of this research can be summarized as follows:

- We present a novel deep image retrieval framework that utilizes both the hierarchical pyramid structure of images and an attention mechanism to improve the descriptive power of image features in end-to-end manner. Our framework encodes images into the multi-scale spatial features, which are then enhanced and aggregated using a weighted sum strategy.
- We introduce an adaptive weighting scheme that takes into account the whole content of an image to generate more accurate attention scores for feature aggregation. Our experiments demonstrate that the adaptive attention strategy is more effective than conventional feature attention mechanism.
- We demonstrate that our framework can be easily integrated into any deep convolutional neural networks (i.e., ResNet50 and VGG16). We also evaluate our method on benchmark datasets.

II. Methodology

A. Network Architecture

As shown in Fig. 2, the proposed pyramid attention network basically consists of a base CNN and three additional modules: spatial-pyramid pooling, feature attention and sum pooling. In this study, we use ResNet50 architecture introduced by He [20] as the base CNN since it achieves state-of-the-art performance of image classification with the comparative number of network parameters. The spatial pyramid pooling is adopted to generate regional features from activations of the last convolutional layer in the base architecture. After that, the attention block is employed to weight each extracted feature based on its informative region. Finally, the attentive regional features are aggregated by using sum pooling layer to obtain a compact descriptor for image retrieval.

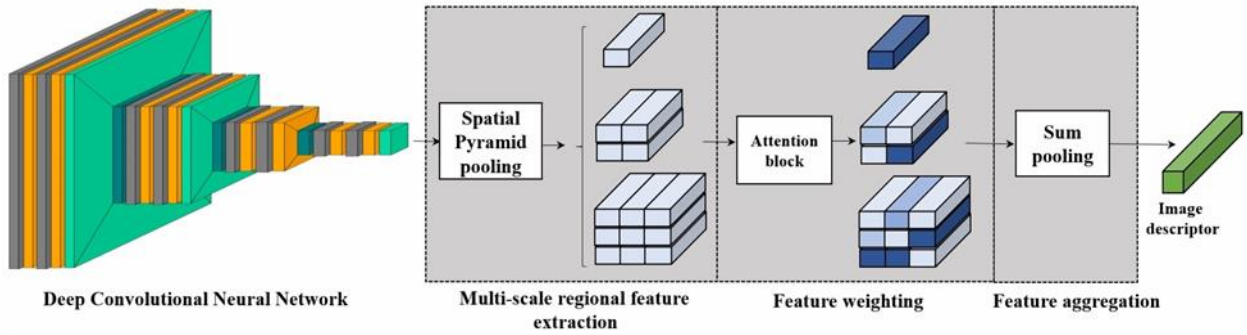

Figure 2. Illustration of the proposed framework.

B. Spatial Pyramid Encoding

Convolutional neural networks typically require a fixed-size input image (e.g., 224×224) for the fully-connected layer, which may limit the accuracy of image classification. Spatial Pyramid Pooling (SPP) was introduced as a solution to this issue, by adding it on the top of the last convolutional layer. In other words, SPP enables to generate fixed-length outputs from feature maps of images of any size. In recent studies, SPP has been shown to improve the generalization of models for tasks, including object detection and semantic segmentation. In this study, we leverage the pyramid pooling to aggregate multi-scale regions in a feature map. However, different from conventional pyramid pooling that applies max pooling on non-overlapping regions in the input map, we adopt overlapping max pooling that performs better in term of spatial invariance. Given the convolutional feature maps: $W \times H \times D$, where $W \times H$ is the size of input map and D is the number of channels, the pyramid pooling has a pooling window size in proposition to the size of feature map.

For a given scale n that generates the output size of $n \times n \times D$, we apply a pooling window size of $\left[\left[2 \times \left(\frac{W}{n+1}\right)\right], \left[2 \times \left(\frac{H}{n+1}\right)\right]\right]$ and the stride of $\left[\left[\frac{W}{n+1}\right], \left[\frac{H}{n+1}\right]\right]$ to enable pooling about 50% overlapping regions. Then, a regional feature set of feature map by scales is obtained as follows:

$$\mathcal{F} = \{f_r^s \mid s \in \{S_1, S_2, .., S_n\}, r = \{1, 2, \dots, N\}\} \quad (1)$$

where f_r^s is the r^{th} regional feature at scale s with a size of $1 \times 1 \times D$. There are n scales in total and N the total number of regions in scale s.

Once a regional feature set of image was obtained, we can embed the feature set to obtain a compact global descriptor: $1 \times 1 \times D$ as follows:

$$G = \sum_s f_r^s \quad (2)$$

C. Pyramid Attention

The attention mechanism was first introduced to improve translation accuracy in natural language processing [21]. The main role of the attention is to find salient words in the input text that needs to get attention. It has also gained popularity in deep neural networks as a powerful addition of computer vision tasks [21–23]. Inspired by these successful efforts, we incorporate the attention unit to improve the generalization of the network. The main reason is the fact that all the regional features may not equally describe the regions of interest. Given a landmark image, some regional features may describe the background or observed objects in the environment (e.g., trees and cars). When the feature set is aggregated to a global descriptor, these regional features can negatively affect the system performance. In this case, the attention unit helps the system earning benefits by assigning appropriate weights to these regional features according to their contributions. Specifically, the attention leverages a $1 \times 1 \times D$ convolutional layer on the regional features to obtain their attention scores. The attention score a_k of the k^{th} regional feature f_r^k is computed by the two operations as following:

$$a_k = \frac{\exp(e_k)}{\sum_j \exp(e_j)} \qquad e_k = q^T * f_r^k \quad (3)$$

where q denotes the D-dimensional vector of parameters and $*$ denotes the inner product operation. The sigmoid function is also applied to scale the corresponding regional feature for computing the attention score. The global descriptor G^* can be expressed as following:

$$G' = \sum_k a_k f_r^k \quad (4)$$

According to Eq. (4), the global descriptor is generated by the weight sum of the aggregated features. Inspired by Yan [22], the utilization of fixed weights for sum-aggregation might prove ineffective due to the impact of image variation. Instead, we look for an adaptive weighting scheme that enables the model to produce more reasonable scores for the feature aggregation by incorporating a content prior from the content of an image. To end this, we utilize the two-level attention. The first level attention generates the aggregated feature G' using the same scheme in Eqs. (3) and (4) with a D-dimensional vector q' as input. The second level attention then computes a D-dimensional vector q'' by using a linear transformation as following:

$$q'' = \tanh(W \circ G' + b) \quad (5)$$

where W and b are a transformation matrix and a bias vector respectively. The feature vector G'' generated by q'' will be the final aggregation results. The vector q' is randomly initialized in the first attention block; while the new vector q'' incorporates a content prior from the global image descriptor G'. By optimizing the training process, the model can adaptively learn the weights and form a global descriptor depending on the context of image.

D. Triplet Loss Training

The proposed network is trained by using triplet labels, a special case of pairwise labels, during the training process [24]. These labels consist of three images: (1) an anchor image, x^a, (2) a positive image, x^p, that has the same label as x^a and (3) a negative image, x^n, that has a different label from x^a. These images are grouped together to form the triple input: $\{x^a, x^p, x^n\}$ for training the network. The network is trained using a triplet loss function, which ensures that the anchor image is closer to the positive image and farther from the negative image at the same time. Given a triplet input: $\{x^a, x^p, x^n\}$, the loss is calculated as following:

$$\mathcal{L}(x^a, x^p, x^n) = \max\{0, d(x^a, x^p)\ d(x^a, x^n)\ m\} \quad (6)$$

where d is a distance metric and m is a margin that controls how far apart the positive and negative example should be. The goal is to minimize this loss, which will enable the network to produce discriminative descriptors. However, the task of creating triplets becomes very demanding in a large dataset. In this study, we utilize an online method to generate triplets.

As depicted in Fig. 3, a training batch consists of a set of images with a fixed batch size. The triplet inputs fed into the network are generated by using every image in the batch and then get the global descriptors. Afterwards, the network parameters are optimized using gradient descent. For more in-depth explanations regarding the online mining of training triplets, it refers to Ref [24].

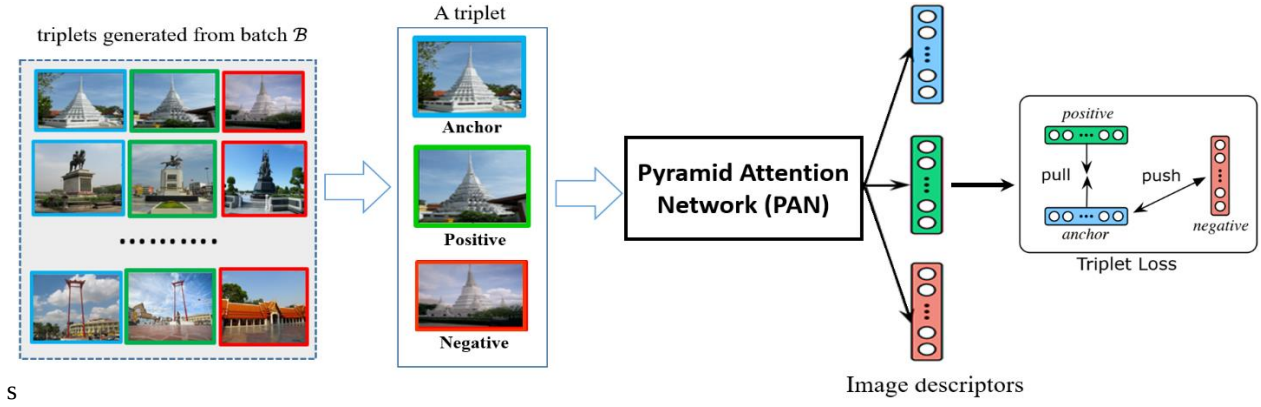

Figure 3. Illustration of triplet loss training in our model.

III. Experiments

In this section, we evaluate the performance of the proposed model. We will begin with introducing the datasets and the baseline methods. After that, we will demonstrate effectiveness of the proposed model.

A. Datasets

The proposed approach is evaluated on three different datasets.

- **Oxford5k Buildings Dataset**: The Oxford5k [25] contain 5,062 images of Oxford landmarks. These landmark images have been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 queries. This gives a set of 55 queries. These 55 query images are also used as the training set for the network training. The remaining is used as the image dataset for retrieval.
- **Paris6k Dataset [26, 27]**: This dataset consists of 6,412 images collected from Flickr by searching for particular Paris landmarks with 12 different landmarks. For each landmark, there are 5 query images. This gives a set of 60 queries, which are also used as the training set for the network training. The remaining is used as the dataset for retrieval as the same with the oxford dataset.
- **Landmark-100 Dataset**: we introduce a new dataset called Landmark-100, containing 20,946 photographs of 100 historical landmarks in Thailand, sourced from the website of the Department of Fine Arts (https://gis.finearts.go.th/fineart/), Thailand. Compared with the two standard datasets for landmark recognition, Landmarks-100 encompasses a broader range of landmark categories with greater variablity within each category. These characteristics make this task more challenging. Fig 4 demonstates the diversity of landmarks in this dataset.

Figure 4. Image samples from the Landmark-100 dataset.

For this experiment, we consider only the annotated region of interest of the query images in the Oxford and Paris datasets, while the entire query image is employed in the Landmark-100 dataset.

B. Evaluation Metrics

For performance evaluation of the proposed model, the goal is to retrieve images that are the same landmark to the query image and to retain as many similar images as possible. To achieve this, two evaluation metrics are used in this study: precision P@N and mean Average Precision (mAP).

The precision P@K measures the percentage of the queries that are correctly retrieved when given K potential positive candidates. The mean Average Precision (mAP) is calculated by finding the Average Precision (AP) for each query image and then taking the mean of these values as follows:

$$MAP = \frac{1}{Q}\sum_{q\in Q} AP(q) \tag{7}$$

where n denotes the number of returned images and Q is the number of query images. and $AP(q)$ denotes the average precision of a query image q, defined as follows:

$$AP(q) = \frac{1}{N}\sum_{i=1}^{n} \frac{i}{R_i} \times rel_i \tag{8}$$

where N is the number of related images in the database for a specific query, R_i is the rank of i-th returned image and $rel_i = 1$ if the image ranked in the i-th position is similar to the query image, otherwise it is 0.

C. Implementation Details

The ResNet-50 network [20] and the VGG-16 network [28] pre-trained on ImageNet dataset are adopted as our base networks. We extract local features by cropping the last convolutional layer of each network, with 512 feature maps for VGG-16 and 2048 for ResNet50. To formulate the global descriptor from feature maps, we append the proposed method, named SPA, and other comparative methods as the aggregation layer. The hyperparameters of SPA are also set. We use four grid scales (1×1, 2×2, 4×4 and 6×6) for the pyramid pooling layer to encode the feature maps into region-based features, resulting in a total of 57 regional features. For the triplet loss training, we use margin m = 0.3 with a batch size of 256 samples. Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) with the learning rate of 0.0005 is used for all the datasets. Table I demonstrates performance comparisions of our approach for different sizes of image descriptor.

As shown in Table I, it is obviously that after training, SPA with ResNet-50 architecture outperforms that of VGG-16 on all the datasets. Additionally, the model achievesthe best results in 512-D as well as 1024-D representations. For the reason of a compact descriptor, we use the SPA method that employs ResNet50 and a 512-D representation to compare with other aggregation methods.

TABLE I. MEAN AVERAGE PRECISION ON ALL THE DATASETS FOR DIFFERENT SIZES OF IMAGE DESCRIPTORS

Dataset	Base Network	Dimensions				
		64	128	256	512	1024
Paris6k	VGG16	69.9	78.7	80.6	**88.4**	87.9
	ResNet50	75.2	81.6	81.5	85.3	**86.4**
Oxford5k	VGG16	47.4	55.3	67.4	**78.7**	78.3
	ResNet50	68.5	73.1	77.4	**89.6**	89.3
Landmark-100	VGG16	49.8	61.4	70.0	78.5	**79.8**
	ResNet50	57.4	70.7	72.6	**80.4**	80.3

Note: Best results are highlighted in bold and second best results are underlined.

D. Comparison with State-of-the-Arts methods

In this section, we evaluate the proposed method with current CNN-based global descriptor methods for image retrieval on all the datasets. To ensure a fair comparison, all the methods are trained using the same protocol and followed by PCA-whitening to obtain a final 512-D descriptor. We first consider NetVLAD layer [14], a state-of-the-art trainable aggregation for local feature descriptors into a compact global representation. This aggregation method computes the difference between local descriptors and centroids, and aggregates these differences into a compact global descriptor. As mentioned in Arandjelovic [14], the NetVLAD layer is initialized a vocabulary size of 64 for K-Means clustering, and the soft assignment parameter α is set to 30. Sum Pooling of Convolutions (SPoC) introduced by Babenko [15] generates a global descriptor by summing convolutional features without feature embedding. The R-MAC descriptor introduced by Tolias [16] involves aggregating the maximum activations within a spatial grid through summation. Our SPA method differs from the R-MAC by adopting a different region choice and use attention mechanism to reinforce the regional features. Following the approach of Perronnin [10], the R-MAC grid scale is defined to 1×2, 2×3 and 3×4, resulting in a total of 20 regions. Generalized-Mean Pooling (GeM) layer [17] is also considered as the baseline. GeM pooling layer focuses on generalizing max and average poolings by introducing a trainable parameter. Radenović [17] has suggested that the pooling parameter can be either manually set or can be learned using backpropagation.

Lastly, our proposed method aims to be compared with BoW-CNN, as proposed by Mohedano [29], which attempts to aggregate local convolutional features using the Bag-of-Words model. In their work, a vocabulary size of 25,000 centroids is applied for all the datasets.

IV. RESULT AND DISCUSSION

Table II provides a summary of the comparison results of all the methods across the datasets. The results presented in the table clearly show the superiority of the proposed method (SPA) over the other methods. SPA has achieved mAP scores of 85.3%, 89.6%, and 80.4% on the Oxford5k, Paris6k and Landmark-100 datasets, respectively. Furthermore, SPA exhibits a consistent high-level performance in terms of top-N precisions. This

highlights the effectiveness of the spatial-pyramid attention layer, which can be trained to learn a compact and robust image descriptor for enhanced landmark retrieval. As seen in Table II, the second best mAP performance is achieved by the methods that extract and aggregate features from regions split in the image, including Bow-CNN and R-MAC descriptors, across all tested datasets. These results demonstrate the benefits of modeling spatial structures in image for feature embedding.

TABLE II. COMPARISON RESULTS WITH CNN-BASED GLOBAL DESCRIPTOR METHODS ON MAP AND P@N USING RESNET50

Method	Oxford5k				Paris6k				Landmarks-100			
	mAP	P@1	P@5	P@10	mAP	P@1	P@5	P@10	mAP	P@1	P@5	P@10
BoW-CNN	73.9	89.6	83.1	76.0	82.0	97.0	93.3	92.3	64.8	84.2	73.1	67.5
R-MAC	66.9	83.8	74.0	68.2	83.0	95.7	93.4	91.8	77.4	85.7	75.9	71.2
NetVLAD	71.6	87.2	79.2	69.3	79.7	94.3	93.1	90.6	75.5	81.0	70.4	62.2
SPoC	68.1	84.3	75.0	67.4	78.2	92.9	91.7	88.6	68.4	80.6	71.4	59.3
GeM	70.8	83.3	72.7	66.3	79.7	94.8	92.2	90.2	72.2	84.1	70.2	67.1
SPA*	**85.3**	**90.2**	**84.3**	**80.3**	**89.6**	**96.7**	**93.6**	**92.4**	**80.4**	**87.6**	**79.7**	**73.3**

Note: Best results are highlighted in bold and second best results are underlined.

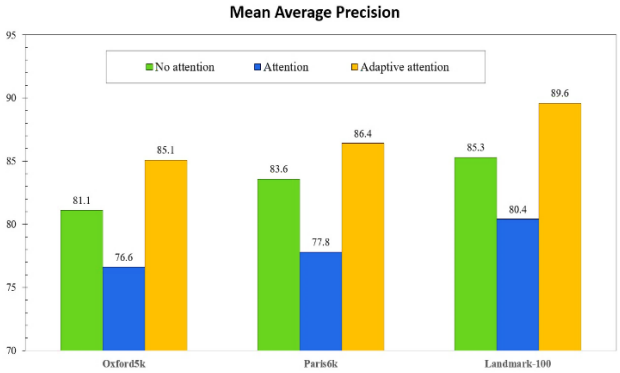

Figure 5. The performance (mAP) comparison for the impact of adaptive attention (%).

We also compare our proposed method with non-region based aggregation methods (i.e., NetVLAD, SPoC and GeM). Among such methods, NetVLAD, the state-of-the-art CNN-based method for image retreival, performs the best mAP performance with the highest mAP score of 79.7% on the Paris benchmark. It achieves this result using VLAD features. However, NetVLAD often requires a dense high-dimensional representation to achieve good results, making it computationally expensive to scale up to large datasets. Compared with NetVLAD, SPA representation is highly sparse, allowing for fast retrieval. Moreover, our proposed method achieves better results in terms of mAP and top-N precisions. The interpretation is that the attention block enhances the local CNN features whose regions of interest are described. Meanwhile it suppresses the confusing regional features captured by pyramid pooling.

Fig. 5 demonstrates the effect of the attention mechanism in our proposed method. We compare the mAP scores of SPA with no-attention, single attention and the two-level attention mentioned in Section II.C. The results shown in Fig. 5 demonstrate that the SPA without attention mechanism performs the worst across the datasets, while the highest mAP results in adopting the two-level attention. These results support the effectiveness of the two-level attention block that incorporates a content prior to refine regional features of image, leading to a more accurate global descriptor. Figs. 6–8 show the top-7 retrieval results of our proposed approach for the Landmark, Oxford and Paris datasets, where the query images (the left-hand side images) corresponds to the return images of the model's retrieval results. If the returned image belongs to the same landmark as the query image, it is enclosed in a green box; otherwise, it's in a red box.

Figure 6. Top seven retrieval results for the Landmarks-100 dataset. (top row) religious structure (middle row) old steel bridge and (bottom row) old fort.

Figure 7. Top seven retrieval results for the Oxford5k dataset.

Figure 8. Top seven retrieval results for the Paris6k.

V. Conclusion

In this paper, a novel deep image retrieval framework, named Spatial Pyramid-Attention (SPA), has been proposed. SPA is a well-designed layer that extracts the image pyramid structure and then generate a compact and discriminative global descriptor for visual landmark search. SPA takes advantages of encoding images into the multi-scale regional features and aggregating them using an adaptive attention strategy. In order to enhance the global descriptor, the network is trained using online triplet mining and triplet loss. The experimental results have demonstrated the effectiveness of the proposed framework, compared with existing deep image retrieval models on widely used benchmark datasets.

Our future work will consider to further improve the proposed framework in the following directions. First, given that we have landmark labels variable, we will work to improve the network capacity to achieve more accurate image retrieval by jointly training both triplet loss and classification loss. Second, we plan to integrate a query expansion strategy into the proposed approach to further improve retrieval performance.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Luepol Pipanmekaporn; data curation, Suwatchai Kamonsantiroj; investigation, Sathit Phasomphan; methodology, Luepol Pipanmekaporn and Chiabwoot Ratanavilisagul; software and writing original draft, Chiabwoot Ratanavilisagul; writing—review and editing, Suwatchai Kamonsantiroj. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

We would like to thank Information Technology Centre for Cultural Heritage, Department of Fine Arts, Thailand, for providing Landmarks-100 dataset, supporting, sharing insights and insightful discussions.

References

[1] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval: A review of recent trends," *Cogent Engineering*, vol. 8, no. 1, 1927469, 2021.

[2] S. Bell and B. Kavita, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–10, 2015.

[3] P. W. Kwan *et al.*, "Content-based image retrieval of cultural heritage symbols by interaction of visual perspectives," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 5, pp. 643–673, 2011.

[4] M. Ye *et al.*, "Person re-identification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[5] K. Ozaki and S. Yokoo, "Large-scale landmark retrieval/recognition under a noisy and diverse dataset," arXiv preprint, arXiv:1906.04087, 2019.

[6] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (CBIR)," *Neurocomputing*, vol. 452, pp. 675–689, 2021.

[7] E. R. Vimina and K. P. Jacob, "Content based image retrieval using low level features of automatically extracted regions of interest," *Journal of Image and Graphics*, vol. 1, no. 1, pp. 7–11, March 2013. doi: 10.12720/joig.1.1.7-11

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[10] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. the 2010 European Conference on Computer Vision*, 2010, pp. 143–156.

[11] W. Chen *et al.*, "Deep learning for instance retrieval: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2022.
[12] J. Lu, J. Hu, and J. Zhou, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76–84, 2017.
[13] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. the 12nd European Conference on Computer Vision*, 2014, pp. 584–599.
[14] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.
[15] A. Babenko and V. Lempitsky "Aggregating local deep features for image retrieval," in *Proc. the 2015 IEEE International Conference on Computer Vision*, 2015, pp. 1269–1277.
[16] G. Tolias, R. Sicre, H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," arXiv preprint, arXiv: 1511.05879, 2015.
[17] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.
[18] W.-K. Chen, *Linear Networks and Systems*, Belmont, CA: Wadsworth, 1993, pp. 123–135.
[19] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. the 14th European Conference on Computer Vision*, 2016, pp. 241–257.
[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[22] M.-H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.
[23] Y. Yan, B. Ni, and X. Yang, "Fine-grained recognition via attribute-guided attentive feature aggregation," in *Proc. the 25th ACM International Conference on Multimedia*, 2017, pp. 1032–1040.
[24] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.
[25] C. Wang, X. Zhang, and X. Lan, "How to train triplet networks with 100k identities?" in *Proc. the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1907–1915.
[26] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
[27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.
[29] E. Mohedano *et al.*, "Bags of local convolutional features for scalable instance search," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 327–331.