Human Interaction and Emerging Technologies (IHIET 2023), Vol. 70, 2023, 88–98
https://doi.org/10.54941/ahfe1002933

Visual Instance Retrieval for Cultural Heritage Artefacts using Feature Pyramid Network

Luepol Pipanmekaporn and Suwatchai Kamonsantiroj

Department of Computer and Information Science, King Mongkut's University of Technology North Bangkok, Bangsue, Bangkok 10800, Thailand

ABSTRACT

Digitized photographs are commonly employed by archaeologists to assist in uncovering ancient artefacts. However, locating a specific image within a vast collection remains a significant obstacle. The metadata associated with images is often sparse, marking keyword-based searches difficult. In this paper, we propose a new visual search method to improve retrieval performance by utilizing visual descriptors generated from a feature pyramid network. This network is a convolutional neural network (CNN) model that incorporates additional modules for feature extraction and enhancement. The first module encodes an image into regional features through spatial pyramid pooling, while the second module emphasizes distinctive spatial features. Additionally, we introduce a two-stage feature attention to enhance feature quality and a compact descriptor is then formed by aggregating these features for searching the image. We tested our proposed method on benchmark datasets and a public vast collection of Thailand's ancient artefacts. Results from our experiments show that the proposed method achieves 77.9% of mean average precision, which outperforms existing CNN-based visual descriptors.

Keywords: Image retrieval, Pyramid attention, Image representation, Convolutional neural network

INTRODUCTION

An archaeologist studies human cultures and societies by uncovering and examining various artifacts such as stone, pottery, metal, and wood. They use this information to gain a deeper understanding of human experience. The most task of the practitioners is to properly identify and determine the age and culture of the artifacts they uncover. To do this, they rely on their prior knowledge, expertise and preference for certain visual characteristics. This process often involves searching through several thousands of artifact images in archaeological archives.

To facilitate the particular task, image retrieval technology can assist by allowing the archaeologists for searching related images in response to a query image. Content-based image retrieval (CBIR) (Hameed et al. 2021), aiming to search for relevant images in an image collection, given a query image, has gained much attention as a search tool in many applications, such

as online produce searching, remote sensing, and landmark retrieval. Previous studies were made to apply CBIR in archeology (Díez-Pastor et al. 2018; Eramian et al. 2017; Kwan et al. 2011). However, these works rely on handcrafted feature extraction like SIFT (Lowe, D. G. 2004) and VLAD (Jégou et al. 2011) for image representation and utilizing machine learning for searching objects of interest. Recently, Convolutional Neural Networks (CNNs) (Rawat and Wang 2017) have recently been successfully applied to a variety of computer vision tasks. These networks, which are commonly used for image analysis, can automatically learn features from input images and use these features to accurately classify and infer useful information from the data. The benefits of CNN-based features are their strong generalization as well as capturing the semantic meaning of image pixels.

Recently, CNNs have demonstrated potential in archaeology such as recognizing pottery (Gualandi et al. 2021), dating ancient stones (Grove and Blinkhorn 2020) and identifying bone surfaces (Domínguez-Rodrigo et al. 2020). However, these studies tend to focus on specific scenarios, limiting their ability to handle the diversity of cultural heritage. The diverse culture of artefacts presents significant challenges for image retrieval. First, physically similar artefacts can come from different cultures, while distinct-looking artefacts can be from the same culture. Additionally, poor artefact conditions and image quality variations often pose challenges for feature extraction. Moreover, lack of training images due to the scarcity of specific artefacts can hinder CNNs. These challenges requires a specialized solution for retrieval in culture heritage.

In this paper, we propose a new CNN-based retrieval model for searching the diversity of archeological artefacts. The model uses a feature pyramid network that is designed to extract a compact and discriminative image descriptor. The proposed approach incorporate additional modules for feature extraction and aggregation. First, Spatial Pyramid Pooling (SPP) (He et al. 2015) is used to encode any size of image to obtain spatial information of artifacts. After that, an attention mechanism is applied to focus informative regions of the features. Then, these regional features are aggregated to produce a compact visual descriptor. To identify artefacts by types, periods and cultures, the CNN network is trained using online triplet mining and triplet loss. The effectiveness of the proposed approach is evaluated using a publicly available repository of digital artefact images from the Department of Fine Arts Thailand, which contains archeological items from thousands of years of Thai art history.

ARTEFACT DATASET

The dataset is publicly accessible on Suvannaphumi Cultural Information Center (CRMA) website: (https://research.crma.ac.th/). It consists of 240,933 images of 34,934 artefacts found in Thailand, covering the period from the 7^{th} century to the 21^{st} century. These artefacts were crafted from 14 different materials and are mainly categorized into 8 types, 4 periods (in AD centuries) and 10 culture backgrounds, provided by archaeologists. Table 1 shows these artefact attributes.

Table 1. Attributes of artefacts in the dataset.

Attribute	Value
Material	Wood; Fabric; Bone; Metal; Glass; Sculpture; Stone; Painting; Carving; Leather; Gold; Silver; Copper; Bronze
Culture backgrounds	Pre-history; Dhavaravati; Sriwijaya, Lopburi, Lan-Na; Sukhothai; Ayutthaya; Rattanakosin
Period	1st to 5th centuries AD; 6th to 10th centuries AD; 11st to 15th centuries AD; 16th to 21st centuries AD
Type	Jewellery; Architecture, Costumes; Coins and banknotes; Stone inscription,; Sculpture; Pottery; Tools

Figure 1: Samples of images by types and cultures (periods in century AD) in the dataset.

We group the artefacts into 3,360 classes based on their type, material, culture and period. In order to maintain a balanced dataset, we choose the 200 largest classes, encompassing a total of 40,912 images of 5,138 artefacts. After that we split the dataset into 80% for training (32,730 images) and 20% for validation (8,182 images). Figure 1 shows examples of artefact images categorized by their type and culture backgrounds.

NETWORK ARCHITECTURE

Figure 2 illustrates the proposed approach, named pyramid attention network. This network basically consists of a base CNN and two additional modules: spatial-pyramid pooling and attention mechanism. The spatial pyramid pooling is adopted to generate local features from activations of the

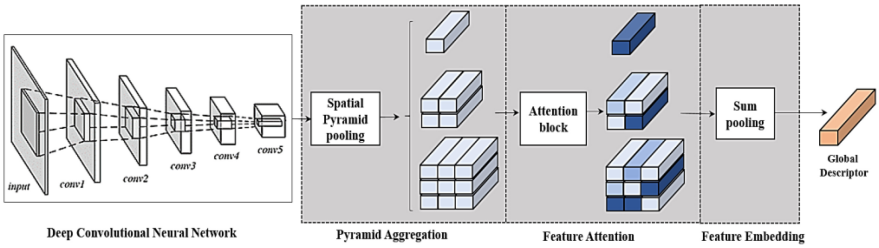

Figure 2: Illustration of the proposed pyramid attention network.

last convolutional layer in the base network. After that, the attention mechanism is employed to enhance the extracted features based on their informative region. Finally, these regional features are aggregated using sum pooling to obtain a global descriptor.

Multi-Scale Feature Extraction

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 in (He et al. 2015) 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 $\left[\left\lceil 2 \times (\frac{W}{n+1}) \right\rceil, \left\lceil 2 \times (\frac{H}{n+1}) \right\rceil\right]$ and the stride of $[\lceil W/(n+1) \rceil, \lceil H/(n+1) \rceil]$ to enable pooling about 50% overlapping regions. Then, a regional feature set of feature map by scales is obtained as follows:

$$\mathcal{F} = \left\{ f_r^s \mid s \in \{S_1, S_2, .., S_n\}, r = \{1, 2, \ldots, N\} \right. \tag{1}$$

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:

$$G = \sum_s f_r^s \tag{2}$$

Attention Mechanism

The attention mechanism was initially applied to enhance NLP translation accuracy. It locates crucial words in the input text that need attention. It has also gained popularity in computer vision tasks (Li et al. 2020). Inspired by these efforts, we incorporate the attention unit to better generalize the features extracted from the spatial-pyramid network. The main reason is the fact that not all the extracted features may describe regions of interest equally. For example, some regional features of an artefact image may describe the background or other objects, negatively impacting retrieval performance when aggregated into a global descriptor. 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$^{\text{th}}$ regional feature f_r^k is computed by the two operations as following:

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

where q denotes the D-dimensional vector of parameters and $*$ denotes the inner product operation. The sigmoid function is applied to scale the corresponding regional feature for computing the attention score. The global descriptor G^* can be expressed as $G' = \sum_k a_k f_r^k$. The global descriptor is generated by the weight sum of the aggregated features. However, applying the fixed weights for sum-aggregation may be ineffective due to the influence of image variation (Li et al. 2020). 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 (3) 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:

$$\mathrm{q}'' = \tanh\left(\mathrm{W}.G' + b\right) \tag{4}$$

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.

Online Triplet Mining

The network model is trained using triplet labels, a special case of pairwise labels, during the training process. 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

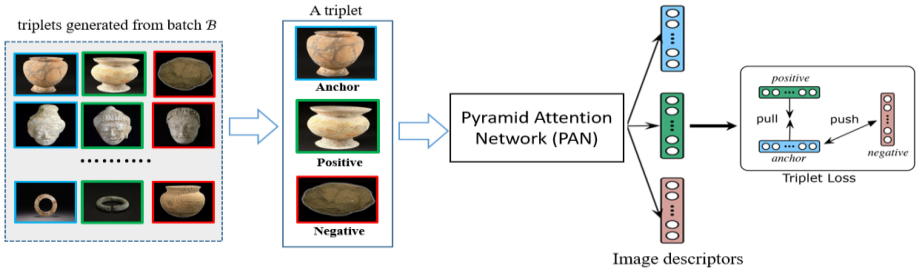

Figure 3: Online triplet mining and triplet loss training.

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:

$$\mathcal{L}\left(x^a, x^p, x^n\right) = \max\left\{0,\ d\left(x^a, x^p\right) - d\left(x^a, x^n\right) + m\right) \tag{5}$$

where d is a distance metric and m is a margin that controls how far apart the positive and negative example should be. The objective of minimizing the loss function is to enhance discriminative descriptors. However, generating triplets from all the training dataset is very challenging. To end this, we effectively mine the triplets in online manner (Wang et al. 2017). As depicted in Figure 3, a training batch consists of a set of images with a fixed size of batch. 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. As the limited pages, we refer readers to (Wang et al. 2017) for more information on online triplet mining and triplet loss training.

EXPERIMENTS

Experimental Setting

We use a pre-trained ResNet50 architecture as a base CNN. The last convolutional layer is cropped and we add the pyramid attention aggregation for fine tuning. To evaluate the proposed approach, we use the following state-of-the-art aggregation methods: NetVLAD (Arandjelovic et al. 2016), SPoC (Jégou et al. 2011), R-MAC (Gordo et al. 2016) and GeM (Radenović et al. 2018) as the baselines. For all these methods, we apply margin $m = 0.3$ with a batch size of 256 samples. We also use Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of 0.0005 for all the datasets. Furthermore, our proposed method employs four different scales of spatial grids ($1 \times 1,\ 2 \times 2,\ 3 \times 3$ and 4×4) in the pyramid pooling block to encode an image into regional features, resulting in a total of 30 regional features. For the R-MAC descriptor, we use three different scales of rigid grids ($1 \times 2,\ 2 \times 3,\ 3 \times 4$), which divide the image into around 20 regions. For the GeM descriptors, a power value of 3 is used as recommended in (Radenović et al. 2018) for general purposes. To ensure a fair competition, the input image size is fixed to be 224×224 pixels and 512

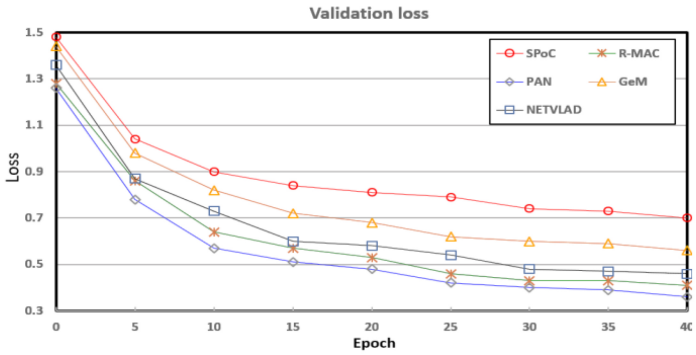

Figure 4: Losses for fine-tuning for all the methods.

dimensions of descriptor are used for all the methods. Figure 4 illustates the validation losses for all the methods on the dataset. As shown in this figure, these methods gradually decrease with similar trends.

Evaluation Metrics

For the evaluation of the dataset, the goal is to retrieve images that are as similar as possible to the query image and to retain as many similar images as possible. To achieve this, two evaluation metrics are used in this study: recall R@10 and mean average precision (mAP). The recall R@10 measures the percentage of correctly retrieved query images among the top 10 candidate images. 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.

Results

The experimental results are shown in Table 2. As can be seen from Table 2, the retrieval effect of the five models on the image dataset is relatively good. The lowest average retrieval accuracy mAP is 67.6% of sPoC and the highest is 77.9% of our proposed network (PAN). Therefore, it can be seen from the comparative experimental results that in the image retrieval experiment of the image dataset. The method PAN proposed in this paper highlights the features of similar images between the same class of artefact images through the attention mechanism and the retrieval effect is optimal with the average retrieval accuracy of 77.9% and a recall rate of 81.8%. The mAP of the other two mainstream methods NetVLAD and GeM descriptors are 71.6% and 69.2, respectively, which is lower than the retrieval accuracy of the proposed model. We also compare our proposed model with R-MAC taking the maximum value for each divided region of input map. By comparing the test results, we find that PAN is still more accurate than R-MAC for the dataset, with +6.42% of the average retrieval accuracy and +4.87% of the recall rate. The interpretation is that the attention enhances the local CNN features whose regions of interest are described. Meanwhile it suppresses the confusing regional features captured by pyramid pooling.

Table 2. Retrieval accuracy of different methods on the image dataset.

Model	mAP	R@10 (%)
NetVLAD	71.6	73.6
SPoC	67.6	70.1
R-MAC	73.2	78.0
GeM (*power* $p = 3$)	69.2	73.0
PAN	**77.9**	**81.8**

Table 3. Ablation experiment accuracy.

Model	R@1 (%)	R@5 (%)	R@10 (%)
PAN + No attention	47.2	67.4	72.6
PAN + Single attention	57.2	74.1	77.7
PAN +Two-level attention	**59.4**	**77.9**	**81.8**

To assess the impact on the accuracy of the network model, an attention mechanism ablation experiment was conducted. The results of this experiment are displayed in Table 3.

As shown in Table 3, the accuracy of the model without any attention mechanism (PAN + no attention) is between 47.2% and 72.6%, while the accuracy of the model using standard attention (PAN + single attention) is between 57.2% and 77.7%. This demonstrates that the attention mechanism has improved the retrieval accuracy of the feature pyramid network. Additionally, the retrieval accuracy of the model that uses the two-level attention mechanism (PAN + two-level attention) is higher than that of single attention one. These results indicate that incorporating the two-level attention mechanism proposed in this paper effectively enhances the image descriptor for searching visual artefacts.

Fig. 5 shows the top 7 most similar ranking results for 5 artifact images. The validation set query images are displayed in the left column, while the closest training set neighbours are to the right. If the neighbour is the same class as the query (same type, material, culture, and period), it's enclosed in a green box; otherwise, it's in a red box. In Figure 5, most of the retrieved images closely match the query images. For instance, the query image in row A that contains a *Ban-Chaing* culture (pre-historical) pottery jar results in all the retrieved images similar jars with the same characteristics. Row C exhibits a woven silk fabric from the *Rattanakosin* period (19th-21st century AD) in the query image, and mostly similar images in the results. However, there may be some false matches, such as in Row D, where the query image features a specific culture's ancient tobacco pipe, but the retrieved results display visually similar pipes from other cultures (red box).

CONCLUSION

In this paper, we propose a novel CNN-based approach for retrieval in culture heritage objects. In our approach, a new CNN architecture, named

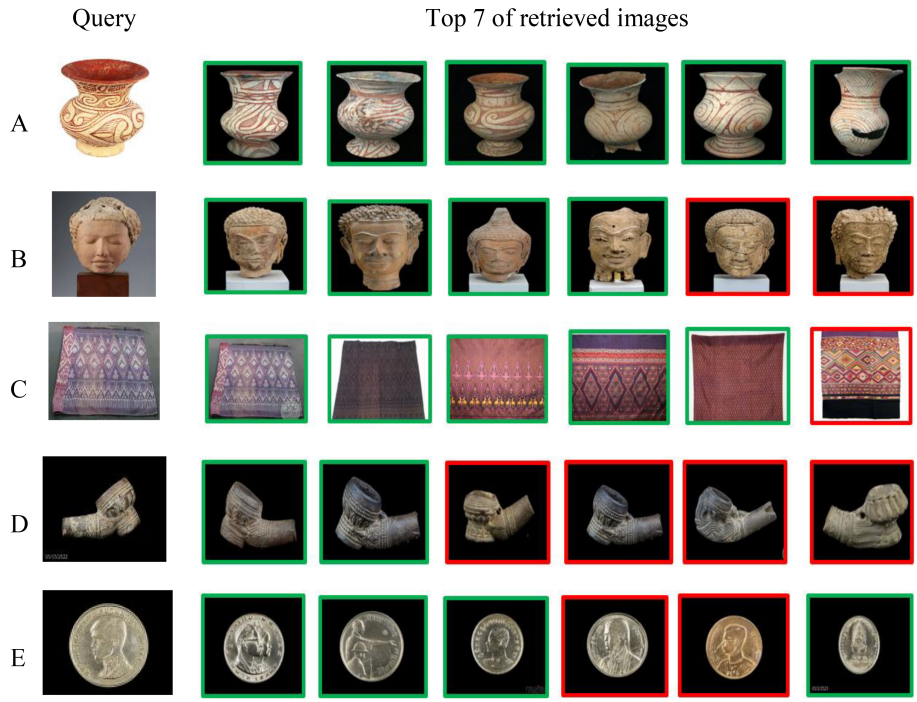

Figure 5: Retrieval results of top 7 culture heritage objects.

feature pyramid network, is well-designed to learn image features and effectively generate a compact visual descriptor by incorporating spatial pyramid pooling and attention mechanism. In order to capture discriminative descriptor, we train the network model with online triplet mining and triplet loss training. Experiment results demonstrated that the network model outperforms other CNN-based feature aggregation methods for image retrieval on standard measures.

ACKNOWLEDGMENT

This research was funded by Faculty of Applied Science, King Mongkut's University of Technology North Bangkok. Contact no. 6446102. We would like to thank IT Centre for Cultural Heritage, Department of Fine Arts, Thailand for providing the image dataset, supporting, sharing insights and insightful discussions.

REFERENCES

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307.

Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014). Neural codes for image retrieval. In *European conference on computer vision*. pp. 584–599.

Chen, W., Liu, Y., Wang, W., Bakker, E. M., Georgiou, T., Fieguth, P. and Lew, M. S. (2022). Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Derech, N., Tal, A., and Shimshoni, I. (2021). Solving archaeological puzzles. *Pattern Recognition*, *119*, 108065.

Díez-Pastor JF, Jorge-Villar SE and Arniaz-Gonzaléz Á et al. (2018) Machine learning algorithms applied to Raman spectra for the identification of variscite originating from the mining complex of Gavà. J Raman Spectrosc 51(9):1563–1574.

Domínguez-Rodrigo, M., Cifuentes-Alcobendas, G., Jiménez-García, B., Abellán, N., Pizarro-Monzo, M., Organista, E., and Baquedano, E. (2020). Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. Scientific Reports, 10(1), 1–11.

Eramian, M., Walia, E., Power, C., Cairns, P., and Lewis, A. (2017). Image-based search and retrieval for biface artefacts using features capturing archaeologically significant characteristics. Machine Vision and Applications, 28, 201–218.

Gominski, D., Poreba, M., Gouet-Brunet, V., and Chen, L. (2019). Challenging deep image descriptors for retrieval in heterogeneous iconographic collections. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents*.pp. 31-38.

Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2016). Deep image retrieval: Learning global representations for image search. In European conference on computer vision. pp. 241–257.

Grove M and Blinkhorn J (2020) Neural networks differentiate between Middle and Later Stone Age lithic assemblages in eastern Africa. PLoS ONE 15(8): e0237528.https://doi.org/10.1371/journal.pone.0237528

Gualandi, M. L., Gattiglia, G., and Anichini, F. (2021). An open system for collection and automatic recognition of pottery through neural network algorithms. Heritage, 4(1), 140–159.

Hameed, I. M., Abdulhussain, S. H., & Mahmmod, B. M. (2021). Content-based image retrieval: A review of recent trends. *Cogent Engineering*, *8*(1), 1927469.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9). PP. 1904–1916.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., and Schmid, C. (2011). Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, *34*(9), 1704–1716.

Kwan, P. W., Kameyama, K., Gao, J. and Toraichi, K. (2011). Content-based image retrieval of cultural heritage symbols by interaction of visual perspectives. International Journal of Pattern Recognition and Artificial Intelligence, 25(05), 643–673.

Li, W., Liu, K., Zhang, L., and Cheng, F. (2020). Object detection based on an adaptive attention mechanism. Scientific Reports, 10(1). pp. 1–13.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91–110.

Lu, J., Hu, J., and Zhou, J. (2017). Deep metric learning for visual understanding: An overview of recent advances. IEEE Signal Processing Magazine, 34(6). pp. 76–84.

Ma, K., Wang, B., Li, Y. and Zhang, J. (2022). Image retrieval for local architectural heritage recommendation based on deep hashing. Buildings, 12(6), 809.

Ozaki, K., Yokoo, S. (2019). Large-scale landmark retrieval/recognition under a noisy and diverse dataset. *arXiv preprint arXiv:1906.04087*.

Radenović, F., Tolias, G., and Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence, 41(7). pp. 1655–1668.

Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352-2449.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In 2007 IEEE conference on computer vision and pattern recognition, pp. 1-8.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In 2008 IEEE conference on computer vision and pattern recognition, pp. 1–8.

Wang, C., Zhang, X., & Lan, X. (2017). How to train triplet networks with 100k identities? In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1907–1915.