Neural Networks 164 (2023) 245–263

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Co-attention enabled content-based image retrieval

Zechao Hu, Adrian G. Bors *

Department of Computer Science, University of York, York YO10 5GH, UK

ARTICLE INFO

Article history:
Received 14 June 2022
Received in revised form 20 November 2022
Accepted 10 April 2023
Available online 23 April 2023

Keywords:
Content-based image retrieval
Co-attention
Clustering

ABSTRACT

Content-based image retrieval (CBIR) aims to provide the most similar images to a given query. Feature extraction plays an essential role in retrieval performance within a CBIR pipeline. Current CBIR studies would either uniformly extract feature information from the input image and use it directly or employ some trainable spatial weighting module which is then used for similarity comparison between pairs of query and candidate matching images. These spatial weighting modules are normally query non-sensitive and only based on the knowledge learned during the training stage. They may focus towards incorrect regions, especially when the target image is not salient or is surrounded by distractors. This paper proposes an efficient query sensitive co-attention[1] mechanism for large-scale CBIR tasks. In order to reduce the extra computation cost required by the query sensitivity to the co-attention mechanism, the proposed method employs clustering of the selected local features. Experimental results indicate that the co-attention maps can provide the best retrieval results on benchmark datasets under challenging situations, such as having completely different image acquisition conditions between the query and its match image.

1. Introduction

Due to the variability in the image content and extensive uncertainty in the data, selecting underlying image features has always been a challenging problem for the Content-Based Image Retrieval (CBIR) task. In earlier approaches, image features are described by hand-crafted descriptors based on low-level feature information (Bay, Tuytelaars, & Gool, 2006; Lowe, 2004; Manjunath & Ma, 1996; Papushoy & Bors, 2015; Park, Jin, & Wilson, 2002; Swain & Ballard, 1991). However, these approaches could not bridge the gap between the information carried by the low-level features and the high-level semantic meaning when considering hand-crafted descriptors. Significant progress was made following the success of deep Convolution Neural Networks (CNNs) on large-scale image classification tasks (Krizhevsky, Sutskever, & Hinton, 2012).

CNN-based image feature extraction for CBIR is categorized according to the features extracted by relying on global or local features. Global feature methods (Babenko & Lempitsky, 2015; Babenko, Slesarev, Chigorin, & Lempitsky, 2014; Radenović, Tolias, & Chum, 2018; Tolias, Sicre, & Jégou, 2016) extract a compact feature vector from each image following a single forward passing through the network. Local feature based CBIR output consists of a tensor, with each entry representing features from local image regions, followed by a separate aggregation method to build the final image representation (Arandjelovic, Gronat, Torii, Pajdla, & Sivic, 2016; Mohedano et al., 2016; Yue-Hei Ng, Yang, & Davis, 2015). In recent works, the local features are further used in spatial verification mechanisms for re-ranking (Cao, Araujo, & Sim, 2020; Noh, Araujo, Sim, Weyand, & Han, 2017).

Despite the successes of CNN-based methods for CBIR, existing spatial attention modules (Noh et al., 2017; Wu, Irie, Hiramatsu, & Kashino, 2018; Yang, Wang, Song, & Gao, 2019) are all query non-sensitive: for a given candidate image, they predict the region of interest purely based on the knowledge learned during the training, regardless of what the query content is about. These query non-sensitive spatial attention modules are very likely to focus on incorrect regions and ignore the object of interest when the target object is not salient or surrounded by distractors relevant to the training data. In Fig. 1, we show some examples in which the query-nonsensitive attention mechanism from the Weighted Generalized Mean (WGeM) pooling (Wu et al., 2018) fails. The Louvre Pyramid and Palace are both potential objects of interest in Fig. 1. When treating the Louvre Pyramid as a query item, it is always ignored by the attention module, while the adjacent Louvre Palace attracts more attention.

Ideally, the attention should be query sensitive, consistent with the current query content. In the examples shown in Fig. 2, when the Louvre Pyramid is treated as the query, this is then correctly highlighted in the resulting co-attention map and vice versa. This kind of query sensitive attention that dynamically

* Corresponding author.
E-mail address: adrian.bors@york.ac.uk (A.G. Bors).

[1] "Co-attention" in this paper refers to spatial attention conditioned on the query content.

https://doi.org/10.1016/j.neunet.2023.04.009

Fig. 1. Examples of query non-sensitive attention where WGeM fails.
Source: Images taken from Wu et al. (2018).

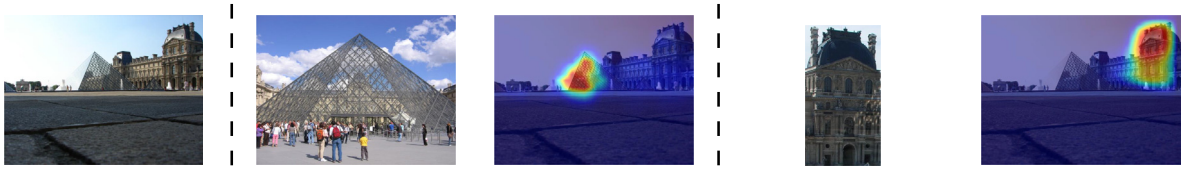


Fig. 2. Examples of query sensitive co-attention maps. (a): Candidate image. Co-attention maps in (c) and (e) are conditioned on the query image in (b) and (d) respectively.

changes with the actual query content is called co-attention. The intuition of applying co-attention is that we only focus on the visually similar regions to the query content when searching for an image. Moreover, in some other computer vision tasks (Hsieh, Lo, Chen, & Liu, 2019; Munjal, Amin, Tombari, & Galasso, 2019; Wang, Zhang, Bertinetto, Hu, & Torr, 2019), the query pattern was shown to be essential for feature extraction and object detection. However, by applying the co-attention to CBIR would require significant extra computation costs, as all potentially useful local features from each database image need to be cached. This cost could be unbearable and make co-attention impractical, especially for large-scale image retrieval.

In this research study, we propose an efficient co-attention method for CBIR, which does not need extra trainable layer optimization but is only used as a post-processing mechanism. In order to reduce the extra computation cost resulting from employing the co-attention mechanism, we consider local feature selection and clustering over candidate local features. Then the co-attention is calculated using the similarity between the query image global feature and the cluster centers of the candidate image local features. This approach dramatically reduces the computation costs while still generating good co-attention maps. The generated co-attention maps are then utilized to re-weight the feature tensor output by the final convolution layer, leading to much better retrieval results.

In summary, our contributions are: (1) a practical co-attention method for large-scale image retrieval; (2) we show that our method can generate good co-attention maps even for some hard image correspondence examples; (3) the retrieval performance is greatly improved with our co-attention method according to the experiments and reaches new state of the art performance on several benchmark datasets; (4) comprehensive ablation study experiments are provided to further prove the effectiveness of the proposed method.

The rest of the paper has the following content. Relevant research studies are reviewed in Section 2. The Generalized Mean pooling and how this is used in training is introduced in Section 3, while the efficiency of employing clustering on the co-attention results, in the context of CBIR, is discussed in Section 4. Further optimization and computation cost reduction are discussed in Section 5. Experimental results are provided and discussed in Section 6. Ablation studies and discussion are provided in Section 7. The conclusions are drawn in Section 8.

2. Related work

In this section, we review CNN-based CBIR works and applications of the co-attention mechanism in computer vision.

Global feature methods. The first deep CNN-based global feature method for content-based image retrieval can be tracked back to the Neural Code model (Babenko et al., 2014), where a pre-trained AlexNet, without the final classification layer, is used as the backbone network followed by a fully connected layer to map the convolutional feature map into a fixed size feature vector. The study from Razavian, Sullivan, Carlsson, and Maki (2016) further demonstrates that, after intensive training, CNN-based image representations can outperform conventional methods using hand-crafted features, and spatial pooling is more appropriate for object retrieval than using fully connected layers. After that, more spatial pooling based methods were proposed for CBIR, including sum-pooling (Babenko & Lempitsky, 2015), max-pooling (Tolias, Sicre, & Jégou, 2016) and generalized mean pooling (Radenović et al., 2018). Compared to the fully connected layer, spatial pooling is faster and requires lower computational resources without employing additional parameters. Moreover, spatial pooling extracts compact feature vectors without requiring any image transformation, while it is also not sensitive to the input image size. To make the global feature lay more emphasis on the designated target object, some spatial attention mechanisms have been implemented in the later works. For instance, in Gordo, Almazán, Revaud, and Larlus (2016, 2017), a Region Proposal Network (RPN) (Ren, He, Girshick, & Sun, 2016) is implemented to enhance the regional max-pooling for image retrieval. The whole model is end-to-end trainable and has a reasonable image retrieval performance. The Weighted Generalized Mean pooling (WGeM) (Wu et al., 2018) applies a trainable

spatial weighting module by adding an extra convolutional layer at the end of a CNN backbone structure. It can effectively localize objects of interest while ignoring redundant regions. However, the spatial weighting may fail when the target object is not discriminating or not matching the training data (Wu et al., 2018). The Second-Order Loss and Attention for image Retrieval (SOLAR) (Ng, Balntas, Tian, & Mikolajczyk, 2020) explores the co-relations between the features from location pairs from the CNN feature map using the second-order spatial information. Unlike the attention methods mentioned above, the SOLAR pipeline generates only one attention map applied on the CNN feature map for each location and a second-order attention map is created to indicate its connection to all other locations. SOLAR is trained on the Google Landmark Dataset (GLD) (Noh et al., 2017), so the model tends to treat all landmark relevant regions as regions of interest. For the irrelevant locations, such as those corresponding to common compact regions like grass or sky, the second-order attention is sparsely distributed over all landmark-like regions. Meanwhile, for locations on the landmark object, the second-order attention would highlight the most distinctive part of that landmark. The Deep Orthogonal Local and Global (DOLG) (Yang et al., 2021) proposes a more comprehensive global feature extraction pipeline, in which an Orthogonal Fusion module is implemented to complement the global feature vector with critical local feature information leading to the current state-of-the-art results for CBIR.

Local feature methods. The deep local feature methods could be further divided into two categories. The first category employs a separate feature aggregation method to generate a single compact feature vector from local descriptors. For example, Yue-Hei Ng et al. (2015) adapts the Vector of Locally Aggregated Descriptors (VLAD) algorithm (Jégou, Douze, Schmid, & Pérez, 2010) as an encoding method for aggregating local features of convolutions into a single feature vector for image retrieval. NetVLAD (Arandjelovic et al., 2016) embeds VLAD into the feature extraction pipeline as a generalized VLAD layer which is end-to-end trainable. The aggregating convolution kernels (ACK) (Wang et al., 2020) utilize convolution kernels to capture specific feature patterns and then combine top activated kernel outputs as the final image representation for CBIR. In order to make the local feature usage more selective, the Deep Local Feature (DELF) (Noh et al., 2017) utilizes a score function with two convolutional layers on top of the CNN backbone for relevant local feature selection. The DELF model has two stages for image retrieval. The first stage calculates a weighted sum of selected local features to get a global feature vector for the initial retrieval results. Then, the selected local features are utilized to perform the spatial verification for a second stage re-ranking and to yield the final retrieval result. After that, this two-stage CBIR framework were improved by other works. The DEep Local and Global features (DELG) model (Cao et al., 2020), based on DELF, unifies the training procedures of global and local features into a single pipeline and further improves the performance of this two-stage image retrieval framework. Detect-to-Retrieve (D2R) (Teichmann, Araujo, Zhu, & Sim, 2019) proposes the Regional Aggregated Selective Match Kernel (R-ASMK), which unifies the region of interest detection, regional local feature aggregation and the similarity measure into one pipeline. Instead of applying feature aggregation or geometry verification with local features, HOW (Tolias, Jenicek, & Chum, 2020) extracts deep local features using the Aggregated Selective Match Kernel (ASMK) (Tolias, Avrithis, & Jégou, 2016) to perform many-to-many local feature matching for CBIR task achieving better results with lower memory requirements than DELF (Noh et al., 2017).

Co-attention has drawn research interest from various computer vision tasks but was hardly considered for CBIR. For instance, the query-guided end-to-end person search network (QEEPS) (Munjal et al., 2019) proposes three query guided sub-networks: QSSE-Net, QRPN and QSimNet which embed the query information into the CNN feature channel after re-weighting, relevant region proposal, and similarity score prediction, respectively. The co-attention and co-excitation (CoAE) framework (Hsieh et al., 2019) utilizes the non-local operation (Wang, Girshick, Gupta, & He, 2018) to explore the correlated evidence revealed by the query-target pairs. The extended feature maps are then channel-wise re-weighted by the squeeze-and-co-excitation (SCE) technique. The Region Proposal Network (RPN) (Ren et al., 2016) selects relevant regions based on the extended target image feature map. RPN can predict relevant regions with respect to the query content even when images from the query class have not been seen during training. The SiamMask (Wang et al., 2019) uses depth-wise cross-correlation to generate response maps of the target image with respect to the query. Then the response map is fed into the convolution layers for pixel-wise classification in order to generate binary co-attention masks. The most relevant work to this paper is the Conditional Attention Network (CANet) proposed in Hu and Bors (2020). CANet considers the global feature vector of the query to each location of the candidate image's convolutional feature map. A set of Multi-Scale convolution blocks (Hu & Bors, 2020) is applied for feature fusion and co-attention map generation for CBIR. In addition, CANet is trained under the supervision of SuperPoint (DeTone, Malisiewicz, & Rabinovich, 2018), where the ground-truth co-attention label for training is automatically generated from the rSfM120k dataset (Radenović et al., 2018). Despite its positive impact on retrieval accuracy, CANet causes unaffordable computation costs, making it impractical for large-scale image retrieval. All these co-attention approaches involve different attention module structures requiring extra attention annotations for model training, while our co-attention method is based on the pre-trained image global descriptor and does not require any network structure modifications or parameter fine-tuning.

3. Spatial pooling and baseline model

The proposed co-attention mechanism does not involve any training and could be treated as a post-processing module for re-weighting on the feature tensor output by a pre-trained CNN-based spatial pooling CBIR model. Accordingly, in this section, we first discuss some insights about spatial pooling. Then, we introduce the structure and training details of the baseline model, which serves as the feature extractor for our co-attention method.

3.1. Spatial pooling

Given an input image $\mathbf{I}$ being processed by a CNN backbone, its output consists of a feature tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times D}$, where H, W, D represents feature map height, width and the number of channels from the last convolutional layer, respectively. Let us consider that the generalized mean pooling (GeM) layer maps the feature tensor $\mathbf{X} = [x_{l,d}] \in \mathbb{R}^{L\times D}$, where $l \in \{1, \dots, L\}$, $L = H \times W$, $d \in \{1, \dots, D\}$ is the channel index, into a compact feature vector $\mathbf{V} = [v_d] \in \mathbb{R}^D$ using :

$$v_d = \left(\frac{1}{L}\sum_{l=1}^{L} x_{l,d}^{p}\right)^{\frac{1}{p}}, \tag{1}$$

where p is a trainable power coefficient. Each feature vector element v_d represents the result of a mapping of the original

feature tensor $\mathbf{X}$. The ratio between each specific feature tensor element $x_{l,d}$ and the feature vector element v_d is expressed as:

$$\begin{aligned} r_{x_{l,d}} &= \frac{x_{l,d}}{v_d} \\ &= \frac{x_{l,d}}{\left(\frac{1}{L}\right)^{\frac{1}{p}} \left(\sum_{l'=1}^{L} x_{l',d}{}^p\right)^{\frac{1}{p}}} \\ &= L^{\frac{1}{p}} \left(\frac{x_{l,d}{}^p}{x_{1,d}{}^p + x_{2,d}{}^p + \cdots + x_{l,d}{}^p + \cdots + x_{L,d}{}^p} \right)^{\frac{1}{p}} \\ &= L^{\frac{1}{p}} \left(\frac{1}{(\frac{x_{1,d}}{x_{l,d}})^p + (\frac{x_{2,d}}{x_{l,d}})^p + \cdots + 1 + \cdots + (\frac{x_{L,d}}{x_{l,d}})^p} \right)^{\frac{1}{p}}. \end{aligned} \tag{2}$$

According to Eq. (2), when $p = 1$, v_d is the mean of each feature map element $x_{l,d}$ at channel d, and the pooling result equals the global average pooling (sum-pooling) (Babenko & Lempitsky, 2015). When $p \to \infty$, it has $r_{x_{max,d}} \to 1$ ($x_{\max,d} = \max_l x_{l,d}$), $v_d \to x_{\max,d}$, and the pooling gives similar result to the max-pooling (Tolias, Sicre, & Jégou, 2016). When $p \in (1, \infty)$ it is the so called Generalized Mean pooling (GeM) (Radenović et al., 2018), so we could treat sum-pooling and max-pooling as special cases of the GeM, and this explains why GeM outperforms the other two pooling methods. Through the power coefficient p, GeM is more selective than the simple global average pooling while considering additional feature information than the max-pooling.

Usually, the similarity measure between spatial pooling feature vectors is performed using cosine similarity or L2 distance (after being L2-normalized). Considering the query image $\mathbf{I}_q$ and a candidate image $\mathbf{I}_c$, their cosine similarity with spatial pooling feature vector is given by:

$$\begin{aligned} s_{q,c} &= (\eta(V_q)V_q)(\eta(V_c)V_c)^{\mathsf{T}} \\ &= \eta(V_q)\eta(V_c) \sum_{d=1}^{D} v_{q,d} v_{c,d} \\ &= \frac{\eta(V_q)\eta(V_c)}{(L_q L_c)^{\frac{1}{p}}} \sum_{d=1}^{D} \left(\sum_{l_q=1}^{L_q} \sum_{l_c=1}^{L_c} \left(x_{q,l_q,d} x_{c,l_c,d} \right)^p \right)^{\frac{1}{p}} \end{aligned} \tag{3}$$

where $\eta(V) = 1/\|V\|$ is a L2 normalization factor. According to Eq. (3), the cosine similarity between two global spatial pooling feature vectors can be interpreted as the sum of multiplications between the entries of the corresponding feature tensors. When the model is trained with either the contrastive loss (Chopra, Hadsell, & LeCun, 2005) or the triplet loss (Arandjelovic et al., 2016), that optimizes the cosine similarity between the global spatial pooling features of image pairs, we can identify the following situations in the context of CBIR: the content from background locations characterized by uniformly consistent information, such as sky, sand, grass, is usually shared among many images. Features from such backgrounds are not distinctive and could not be utilized to distinguish two distinct images or to find correspondences between two matching ones. Accordingly, the activation value across all channels, when considering such plain background locations, tends to be zero ($x_{l_{bg},d} \to 0$). So these locations make little contribution to the final similarity score. On the contrary, distinctive foreground feature locations across all channels tend to have large absolute values ($|x_{l_{fg},d}|$ is maximized), resulting in significant contributions to the final similarity score. Meanwhile, for certain foreground location pairs, which depict the matching objects or regions between $\mathbf{I}_q$ and $\mathbf{I}_c$, their feature representations are pushed closer together such that they yield large positive product values for the final similarity score. Conversely, feature representations for location pairs that depict unmatching objects are pushed away from each other, yielding minimal (negative) values for the final similarity score in Eq. (3).

Training with the cosine similarity loss between spatial pooling feature vectors from Eq. (3) provides useful hints to the CNN model. First, optimizing the global feature vector's cosine similarity between image pairs implicitly optimizes the local feature matching. Foreground locations have higher absolute feature activation values across all channels (having higher L2 and L1 norms), while the background locations have lower feature activation values.

3.2. Baseline model structure and training

The general framework that uses a deep CNN for feature tensor extraction followed by a global spatial pooling layer for compact global feature vector building has been widely used in CBIR works (Cao et al., 2020; Radenović et al., 2018; Wu et al., 2018; Yang et al., 2021). In this research study, we employ the ResNet (He, Zhang, Ren, & Sun, 2016) as the backbone network for the feature tensor extraction. The feature tensor output from the final convolution layer is pooled by a generalized mean pooling (GeM) layer, where we consider the power co-efficient $p = 3$ in Eq. (1), followed by a trainable fully connected layer for feature whitening.

The recent models DELG (Cao et al., 2020) and DOLG (Yang et al., 2021) models are trained on the Google Landmark version 2 dataset (GLDv2) (Weyand, Araujo, Cao, & Sim, 2020). For a fair comparison with these models, in our experimental section, we also train our baseline model on GLDv2. Following the approach in DELG (Cao et al., 2020), we also consider image-level class labels and ArcFace margin loss (Deng, Guo, Xue, & Zafeiriou, 2019) for the model training, defined by:

$$L(\widehat{\mathbf{V}}_g, \mathbf{y}) = -\log \frac{\exp(\gamma\ \mathrm{AF}(\widehat{\mathbf{V}}_g \widehat{\mathbf{w}}_i^T, y_i))}{\sum_{j=1}^{N_c} \exp(\gamma\ \mathrm{AF}(\widehat{\mathbf{V}}_g \widehat{\mathbf{w}}_j^T, y_j))}, \tag{4}$$

where $\widehat{\mathbf{V}}_g$ is the whitened, then L2 normalized global GeM feature vector from Eq. (1), for each input training image. $\widehat{\mathbf{w}}_i$ refers to the trainable L2 normalized classifier weights for class i from the ArcFace weight matrix $\mathcal{W} \in \mathbb{R}^{N_c \times D}$, N_c represents the number of classes in the training dataset. $\mathbf{y}$ is a one-hot class label vector and i is the index of ground-truth class of $\widehat{\mathbf{V}}_g$ ($y_i = 1$). γ is a trainable temperature parameter. $\mathrm{AF}(u, y)$ is the ArcFace-adjusted cosine similarity (Cao et al., 2020):

$$\mathrm{AF}(u, z) = \begin{cases} \cos(\arccos(u) + m), & \text{if } z = 1 \\ u, & \text{if } z = 0 \end{cases}, \tag{5}$$

where u is the cosine similarity, z indicates whether it is the ground-truth class and m is the ArcFace margin.

The ArcFace margin loss from Eq. (5) is also referred to as a "cosine classifier" (Cao et al., 2020). Within the ArcFace weight matrix $\mathcal{W}$, each row $\mathbf{w}_i$, $i \in \{1, 2, 3, \ldots, N_c\}$ can be treated as a proxy feature vector for class i. In other words, these proxy features approximate corresponding original class image features. Accordingly, the ArcFace loss aims to optimize the cosine similarity not between single image pairs but between query and proxies of image classes. Compared to the traditional image pair similarity loss (contrastive loss or triplet loss), this kind of proxy-based similarity loss does not need hard sample mining and converges faster than the simple similarity loss between specific image pairs (Movshovitz-Attias, Toshev, Leung, Ioffe, & Singh, 2017).

4. Enabling CBIR with co-attention

In the following, we consider using the convolution feature tensor output by the pre-trained CNN model for enabling the co-attention generation process. The baseline GeM model, trained as

described in Section 3.2, is used for feature extraction without considering any parameter fine-tuning or structure modification.

4.1. A naive way for co-attention feature generation

Let us consider a pair of images representing the query image $\mathbf{I}_q$ and the candidate image $\mathbf{I}_c$ from a given database. After feeding through the backbone CNN, these images yield the feature tensors $\mathbf{X}_q \in \mathbb{R}^{H_q \times W_q \times D}$ and $\mathbf{X}_c \in \mathbb{R}^{H_c \times W_c \times D}$ as the outputs. The former query tensor is transformed into a compact query feature vector $\mathbf{V}_q \in \mathbb{R}^D$ by the spatial pooling using Eq. (1). The feature tensor $\mathbf{X}_c$, resulting from the final convolutional layer, models the grid-structured representations according to the corresponding locations for the candidate image. The precision of the correspondence between each entry from the feature tensor and locations on the input image depends on the processing properties of the CNN backbone structure. For example, ResNet (He et al., 2016) contains five blocks, each down-sampling the input feature tensor by half. Each local feature from the output feature tensor $\mathbf{X}_c$ corresponds to a 32×32 ($2^5 = 32$) pixels region from the input image.

A naive and straightforward way to get the co-attention map $\mathbf{a}_{naive} = [a_{l_c}] \in \mathbb{R}^{H_c \times W_c}$ of the candidate image $\mathbf{I}_c$ with respect to the query image $\mathbf{I}_q$ is by simply calculating the cosine similarity between the global query feature vector $\mathbf{V_q}$ and the candidate feature tensor $\mathbf{X}_c$ from each location, as :

$$a_{l_c} = \widehat{\mathbf{V}}_q \cdot \widehat{\mathbf{x}}_{c,l_c}^{\mathsf{T}}, \tag{6}$$

where $\widehat{\mathbf{V}}_q$ represents the whitened (by the pre-trained fully connected layer) and L2 normalized query feature $\mathbf{V}_q$. $\widehat{\mathbf{x}}_{c,l_c} \in \mathbb{R}^D$ is a local feature vector at location l_c from the candidate image feature tensor $\mathbf{X}_c$ that has been whitened and then L2 normalized. We apply a Softmax operation on $a_{l_c} \in [-1, 1]$ to normalize their values into the range [0, 1]:

$$a'_{l_c} = \frac{\exp(a_{l_c})}{\sum_{i=1}^{K} \exp(a_i)}. \tag{7}$$

The visualization comparison of the results for the L2 norm attention and naive co-attention map is provided in Fig. 3. The L2 norm attention maps, shown in the third column of Fig. 3, are obtained by calculating the L2 norm of the feature tensor $\mathbf{X}_c$ at each location. The resulting attention map is then resized to the original image size and then overlapped onto the image as a heat-map for the sake of visualization. We can observe that L2 norm attention maps tend to highlight representative parts of all landmark buildings. The naive co-attention maps, shown in the fourth column of Fig. 3, are visualizations of the results provided by Eqs. (6) and (7). We can observe that simple cosine similarity between candidate local features and the query global feature can give some not-bad co-attention results. The first row from Fig. 3 shows an easy case of image retrieval, in which the target object is salient, while also being large in the candidate image without having any distractors around; both L2 norm attention and the naive co-attention indicate reasonable highlight regions. For the hard case from the second row of Fig. 3, the target object is not only small and remote but there are some similar architecture class buildings nearby, and the naive co-attention indicates the correct matching region, while the L2 norm highlights many irrelevant distractor objects and regions.

Despite the above discussion about the suitability of co-attention generated based on global-to-local feature matching, there are still two main problems with this naive co-attention implementation. First, each local feature from the candidate image feature tensor corresponds to a small region from the original input image. These localized features may not be comprehensive enough to represent the whole object or regions of interest, which may indicate wrong regions or even noisy undefined areas. Second but also the most critical problem with the method described above is its computation cost. Consider an input image $\mathbf{I}$ of size $h \times w$, after feeding through ResNet (He et al., 2016), the output feature tensor $\mathbf{X}$ is of size $\frac{h}{2^5} \times \frac{w}{2^5}$. For a high-resolution image of 1024×1024 pixels, the output candidate feature tensor size could be as large as 32×32. For each element of these local features, if we have a 4 Byte float number for representation, the total memory cost for each candidate image local features is $2048 \times (\frac{1024}{32})^2 \times 4$ Bytes $\approx$ 8 MB, where 2048 is the channel count for the feature output by ResNet network. If considering the multi-scale feature extraction (Radenović et al., 2018), the memory cost would increase exponentially. Pre-caching that many local features for a large image retrieval database becomes impractical.

In the following, we aim to make the co-attention mechanism efficient and practical to be used even for large-scale image retrieval tasks. As mentioned above, the most critical problem for co-attention is the memory and computation costs required when considering a large number of local features that could be extracted from a single image. To address this problem, we consider clustering in order to define and extract a smaller characteristic latent space. Such a representative latent space can uncover subspace data structures based on the feature self-expressiveness properties, which can be optimally used for retrieval.

4.2. Co-attention enabled through feature selection and clustering

Local feature selection and clustering. An intuitive way to reduce the extra cost in computation and memory requirements is to decrease the number of local features which have to be stored for each image. Not all local features from the feature tensor output by the backbone network are relevant for the CBIR task. For example, local features from the background are not relevant for the identification of the image content and should be discarded. As discussed in Section 3.1, the L2 norm of each entry from the CNN feature tensor reflects its importance. Accordingly, after feeding through the backbone network, L2 norm based feature selection is performed on the feature tensor $\mathbf{X}$ output by the final convolution layer. Consequently, we retain only the top N local features with the highest L2 norm, resulting in a set of local features $\mathbf{X}_N \in \mathbb{R}^{N \times D}$.

By considering ResNet as the backbone structure, each local feature could be treated as representing a small region of 32×32 pixels from the original image. These local features may correspond to a small region of the object of interest and lack high-level semantic meaning. Meanwhile, we want to reduce the number of candidate local features further. Thus, for the initially selected local features $\mathbf{X}_N$, we employ k-means clustering, grouping them into K clusters and extracting the centers of the clusters as characteristic feature vectors. Within each cluster, we perform generalized mean pooling to get the representative feature vector of the cluster followed by whitening with the fully connected layer, resulting in the local feature set $\mathbf{X}_K \in \mathbb{R}^{K \times D}$, $K \ll N$. Clustering the set of N features corresponds to grouping together image regions defined by similar features.

The original k-means clustering method randomly initializes cluster centers, resulting in large variations in the results, which is not desirable for a stable CBIR system. In order to address this drawback, we adapt a stable cluster initialization approach, namely k-means++, which was proposed in Arthur and Vassilvitskii (2007). Considering the local features $\mathbf{X}_N = [\mathbf{x}_n] \in \mathbb{R}^{N \times D}$, where $\mathbf{x}_n \in \mathbb{R}^D$ indicates the nth local feature from $\mathbf{X}_N$ as input, the cluster center initialization is conducted as :

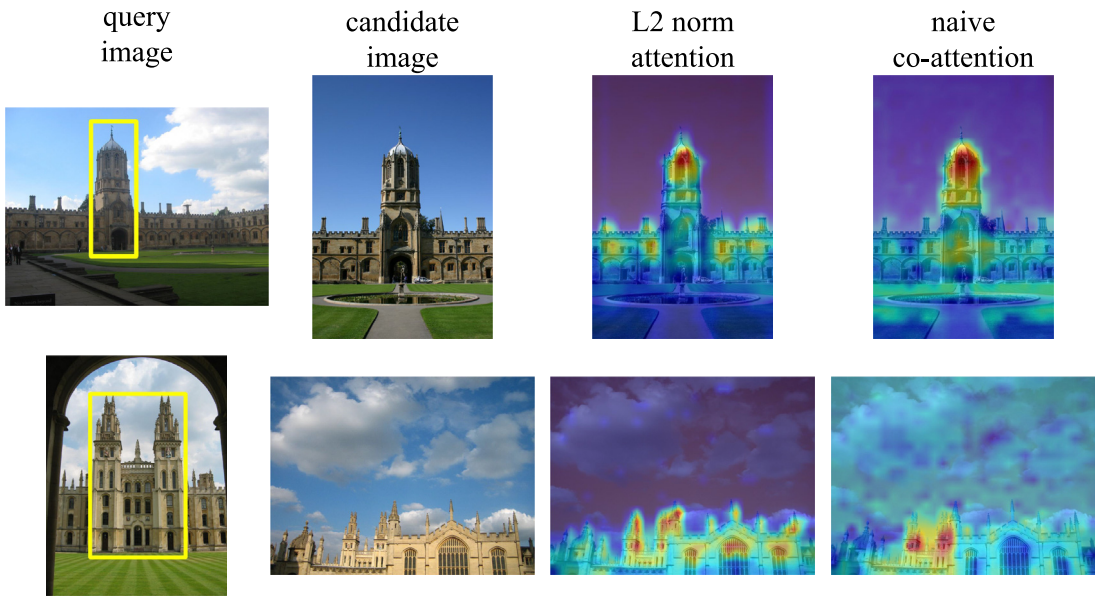


Fig. 3. Visualization comparison of L2 norm attention and naive co-attention. The first column shows the query image with a yellow bounding box outlining the target object, while the second column shows the candidate image. The third column shows the L2 norm attention while the fourth column represents the result of the naive co-attention evaluated as described in Section 4.1.

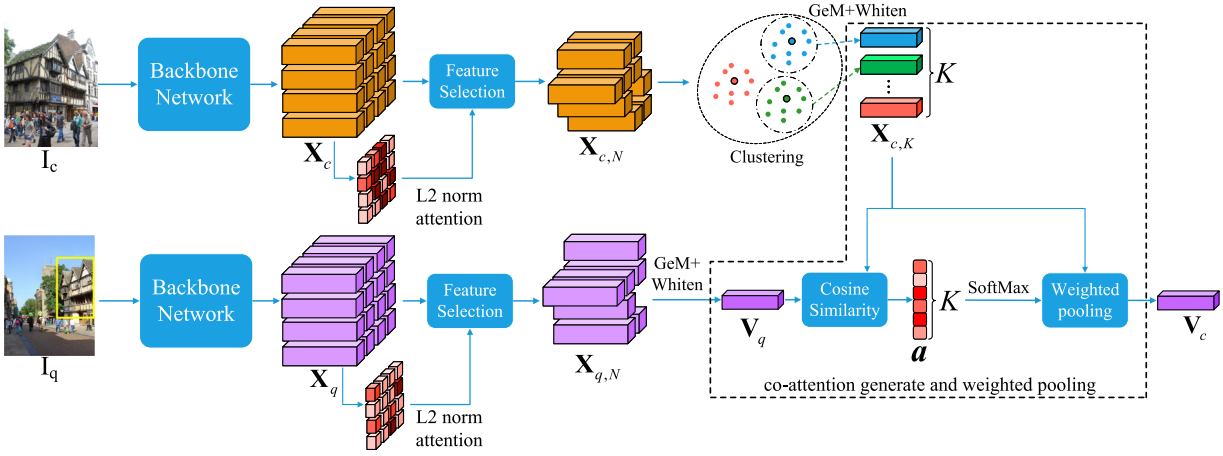


Fig. 4. Illustration of clustering-based co-attention generation and weighted feature extraction.

Step 1: Among $[\mathbf{x}_n]$, choose $\mathbf{x}_i$ ($i = \arg\max_n \|\mathbf{x}_n\|^2$) as the first cluster center.

Step 2: For each local feature $\mathbf{x}_n$ not chosen yet, compute $d(\mathbf{x}_n)$, the smallest distance between $\mathbf{x}_n$ and all centers that have already been chosen.

Step 3: Choose $\mathbf{x}_m$ ($m = \arg\max_n d(\mathbf{x}_n)$) as one of the new centers.

Step 4: Repeat Steps 2 and 3 until K centers are chosen.

The selected cluster centers are then used to initialize the standard k-means clustering and eventually get the clustered local feature set $\mathbf{X}_K$.

Co-attention generation with clustered local features. The pipeline for the co-attention generation and weighted feature extraction is illustrated in Fig. 4. Each time we consider the query $\mathbf{I}_q$ and the candidate image $\mathbf{I}_c$ pair, fed through the backbone network and following the local feature selection using the L2 norm, selected query local features $\mathbf{X}_{q,N}$ are directly GeM pooled and whitened to obtain the query global feature $\mathbf{V}_q$. Selected candidate local features $\mathbf{X}_{c,N}$ are clustered and then whitened, resulting in the clustered local feature set $\mathbf{X}_{c,K}$. Then, the co-attention weights $\mathbf{a} = [a_i] \in \mathbb{R}^K$ are the result of the cosine similarity between $\mathbf{V}_q$ and each local feature extracted from $\mathbf{X}_{c,K}$. As the feature weights are calculated by cosine similarity between query and candidate features, they range between $[-1, 1]$, which may not ensure a high contrast among the results. For better controlling the weight distribution, we normalize $\mathbf{a}$ into the range $[0, 1]$ using SoftMax function with a temperature parameter T defined by:

$$a_i' = \frac{\exp(a_i T)}{\sum_j^K \exp(a_j T)} \tag{8}$$

The final co-attention weighted candidate global feature vector $\mathbf{V}_c$ is defined by weighted sum pooling:

$$\mathbf{V}_c = \frac{1}{K} \sum_i^K a_i \mathbf{X}_{c,i}. \tag{9}$$

The final similarity measure $S_{q,c}$ between the query image $\mathbf{I}_q$ and the candidate $\mathbf{I}_c$ image is performed by evaluating the cosine

[2] $\|\cdot\|$ represents the L2 norm of the feature.

similarity between $\mathbf{V}_q$ and $\mathbf{V}_c$:

$$S_{q,c} = \cos(\mathbf{V}_q \cdot \mathbf{V}_c) . \tag{10}$$

5. Further computation cost optimization

In this section, we provide some further processing steps employed during the retrieval stage to ensure that the proposed co-attention is practical to be used for large-scale image retrieval.

5.1. Feature dimension reduction by PCA

Principal component analysis (PCA) has been used as a common method for feature dimension reduction. Unlike some other works that jointly perform dimension reduction and feature whitening by one fully connected layer (Tolias et al., 2020), we perform dimension reduction by using the Principal Component Analysis (PCA) as a post-processing step. There are two main reasons to use the PCA: first, we found that training with the original feature dimension, which is 2048 for ResNet, makes the model converge faster; second, it is more convenient and fair to compare the retrieval performance with different dimension settings as all experiments are based on the same pre-trained model. For the query image, PCA dimension reduction is applied on its whitened global feature vector $\mathbf{V}_q$. For the candidate image local features, PCA is applied on each whitened local feature from $\mathbf{X}_{c,K}$ before L2 normalization.

Given N_F sample features in D_F dimensions: $\mathbf{F} \in \mathbb{R}^{N_F \times D_F}$, let $\mathbf{m} \in \mathbb{R}^{1 \times D_F}$ denote the mean vector:

$$\mathbf{m} = \frac{1}{N_F} \sum_{i=1}^{N_F} \mathbf{F}_i . \tag{11}$$

Then the covariance matrix of standardized sample features $\mathbf{F}$ is:

$$\mathrm{Cov}_F = \frac{1}{N_F} (\mathbf{F} - \mathbf{m})^{\mathsf{T}} (\mathbf{F} - \mathbf{m}) . \tag{12}$$

Let $\mathbf{P} \in \mathbb{R}^{D_P \times D_F}$, denote the D_P largest eigenvalues of Cov_F, where $0 < D_P < D_F$. For a given feature vector $\mathbf{Y} \in \mathbb{R}^{N_Y \times D_F}$, the dimension reduced output feature $\mathbf{Y}' \in \mathbb{R}^{N_Y \times D_P}$ is calculated by:

$$\mathbf{Y}' = (\mathbf{Y} - \mathbf{m})\mathbf{P}^{\mathsf{T}}. \tag{13}$$

In our implementation, PCA components $\mathbf{m}_v \in \mathbb{R}^{1 \times D}$ and $\mathbf{P}_v \in \mathbb{R}^{D' \times D}$, where D' denotes the feature dimension after PCA dimension reduction, are learned from the whitened global GeM pooling feature vectors (without L2 normalization) of randomly selected images from the training dataset.

5.2. Filtering out evident non-matching images with the inverted file indexing

For image retrieval, especially on a large-scale candidate image database, we may not necessarily need to apply the co-attention mechanism for each candidate image. Actually, some candidate images are evidently not worth considering for the similarity measure with the query. We employ the inverted file indexing in order to reduce the number of candidate images to be considered when assessing the similarity with the query image. Similar techniques have been applied in other CBIR methods. For example, HOW (Tolias et al., 2020) only performs feature comparisons between the local features that share the same visual word. Similarly, after the feature dimension reduction using PCA, we use the local features from the feature tensor output by the final convolution layer to train the codebook. At the feature extraction stage, both query and candidate image local features $\mathbf{X}_{c,N}$ and $\mathbf{X}_{q,N}$, after dimension reduction and whitening, are clustered over the visual words from the codebook. We record the visual word indices to which each image is assigned. Then during the retrieval stage, for each query image, we only pick out candidate images that share at least one visual word with the query image to perform co-attention generation and assess their similarity. The other candidate images which are not selected are simply set to have zero similarity score with the query image.

The global pipeline of the proposed co-attention enabled CBIR framework when considering the inverted file indexing is provided in Fig. 5.

Codebook training. The inverted file indexing starts with the codebook training. As shown in Fig. 5(a), at the codebook training stage, each sample image is fed through the pre-trained backbone network followed by the L2 norm based feature selection, resulting in N_{cdb} local features. With N_s sample images from the training dataset, there would be $\mathrm{N}_s \times \mathrm{N}_{cdb}$ sample local features. To reduce the computation cost, PCA dimension reduction is applied, resulting in $\mathbf{m}_{cdb} \in \mathbb{R}^{1 \times D}$ and $\mathbf{P}_{cdb} \in \mathbb{R}^{D_{cdb} \times D}$, which are learned from these sample local features.[3] After the PCA dimension reduction, k-means clustering is applied to get the final K_{cdb} visual words, with their index represented by $\{v_1, v_2, \ldots, v_{cdb}\}$, as the codebook.

Feature caching. As shown in Fig. 5(b), each database image $\mathbf{I}_c$ is fed through the backbone network during the offline database image feature extraction and caching stage. After the feature selection, one processing branch performs k-means clustering over $\mathbf{X}_{c,N}$ followed by PCA dimension reduction with parameters $\mathbf{m}_v$ and $\mathbf{P}_v$, resulting in the dimension reduced clustered local features $\mathbf{X}_{c,K}{}' \in \mathbb{R}^{1 \times D'}$. Another processing branch performs PCA dimension reduction, with the parameters $\mathbf{m}_{cdb}$ and $\mathbf{P}_{cdb}$, followed by clustering over the codebook and assigning each local feature to the closest visual word. A dictionary is then used to record the database image ID for each visual word index. Each key of this dictionary is a visual word index corresponding to a set of database image IDs whose any one of the local features is assigned to this visual word. The dictionary is then updated for the entire database of candidate images.

Online retrieval. As shown in Fig. 5(c), at the online retrieval stage, the selected query image local features $\mathbf{X}_{q,N}$, after PCA dimension reduction with $\mathbf{m}_v$ and $\mathbf{P}_v$, are also clustered over the codebook. Then, based on the cached dictionary, we only pick out those database images that share at least one visual word with the query image for the following co-attention weighted feature extraction (as shown in Fig. 4) and similarity assessment. All other candidate images are treated as having zero similarity score to this query and are removed from the image search. For the inverted file indexing, the only extra thing needed to be cached is the visual word index dictionary and the codebook, so it would hardly require any extra memory.

6. Experiments

We initially discuss the experiment setup, including the hyper-parameter setting and implementation details. Then, we provide and analyze the co-attention and retrieval results for the proposed methodology, along with comparisons to the state of the art.

6.1. Experiment setup

Implementation details. For the training, we follow the methodology adopted in DOLG (Yang et al., 2021). Input images are data augmented by randomly cropping, changing image ratios

[3] Note that $\mathbf{m}_{cdb}$ and $\mathbf{P}_{cdb}$ are another set of PCA components used for the inverted file indexing only.

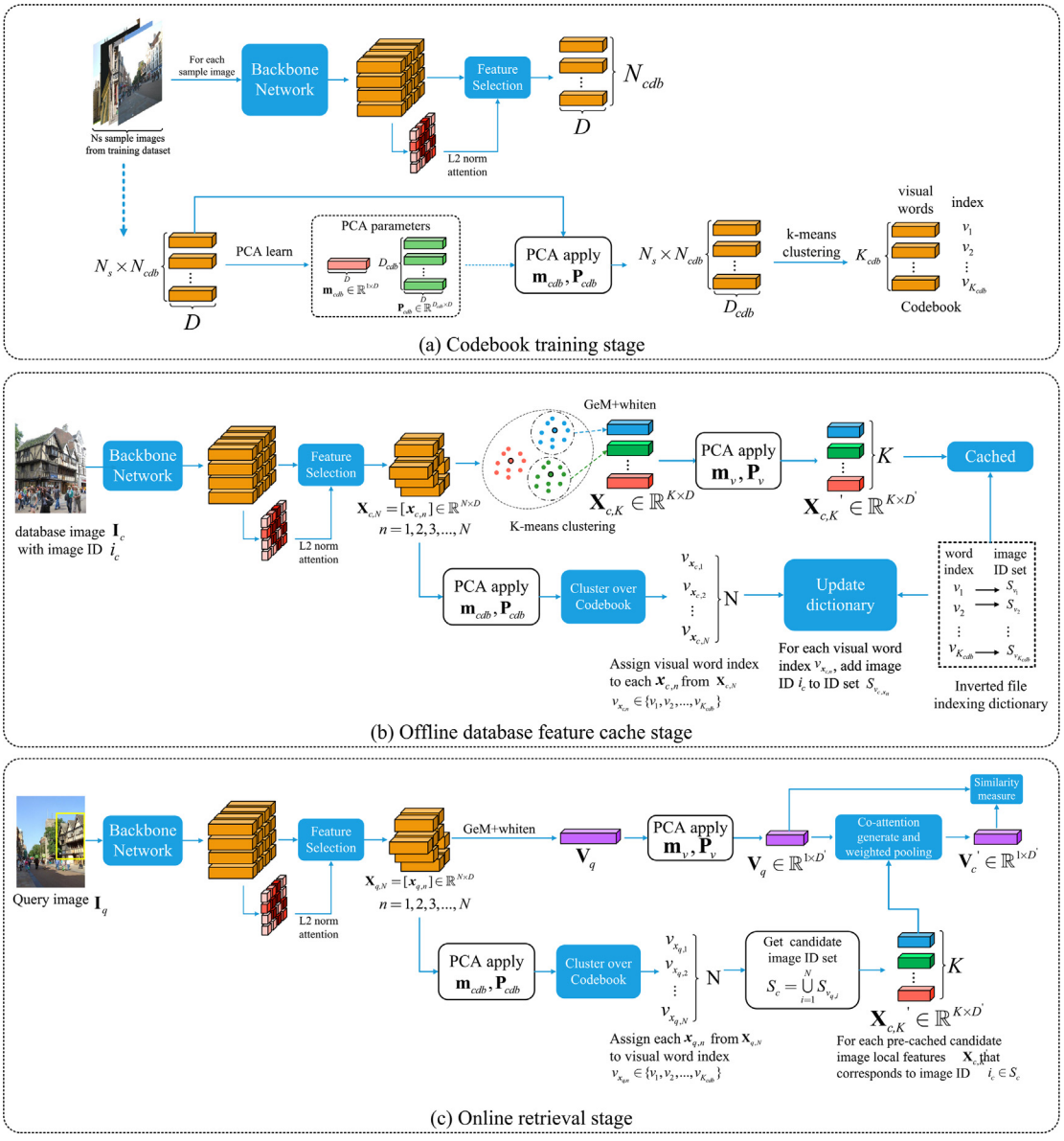


Fig. 5. Pipeline of the CBIR using inverted file indexing.

and then resizing them to 512 × 512 pixels. The batch size is set to 128. The model is optimized using the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.05 and a weight decay of 0.0001. A cosine learning rate decay strategy (Yang et al., 2021) is considered. We set $\gamma = 30$ and margin $m = 0.15$ for the ArcFace loss in Eqs. (4) and (5), while the power coefficient is $p = 3$ for the GeM pooling in Eq. (1). The training is conducted with 4 NVIDIA Tesla GPU and the model is trained for no more than 50 epochs.

For the co-attention mechanism described in Section 4.2, we set $N = 500$ for local feature selection, cluster count $K = 10$ for k-means clustering and $T = 10$ for the SoftMax temperature in Eq. (8).

For the dimension reduction of query image and candidate image features, the PCA components $\mathbf{m}_v$ and $\mathbf{P}_v$ are learned with whitened global GeM pooling feature vectors (without L2 normalization) of 50,000 random images from the training dataset. After whitening, the global query image feature vector $\mathbf{V}_q$ and clustered candidate image local features $\mathbf{X}_{c,K}$ are compressed using the PCA dimension reduction with parameters $\mathbf{m}_v$ and $\mathbf{P}_v$ to the dimension $D' = 512$.

For the inverted file indexing, we use $N_s = 60{,}000$ random images in a single original scale from the training dataset (GLDv2), with $N_{cdb} = 300$ local features being selected from each of them to train the codebook. The size (cluster count) of codebook $K_{cdb} = 65536$. For computation cost reduction, PCA parameters $\mathbf{m}_{cdb}$ and $\mathbf{P}_{cdb}$ are learned from these sample features and used to compress them to dimension $D_{cdb} = 128$.

Evaluation datasets. ROxf/RPar datasets (Radenovic, Iscen, Tolias, Avrithis, & Chum, 2018) have commonly been used for large-scale CBIR performance evaluation in recent years. The ground-truth matching images to each query image are divided into 3 categories, *Easy*, *Medium*, *Hard*, according to the level of difficulty in assessing the similarity of their image representation with the corresponding query. In addition, R1M (Radenovic et al., 2018) is an additional distractor set containing 1 million images, which is used in combination with ROxf and RPar. The retrieval results are reported by mean average precision (mAP) (Philbin, Chum, Isard, Sivic, & Zisserman, 2007).

Feature extraction with multi-scale scheme. The multi-scale feature extraction scheme has been widely applied in both global

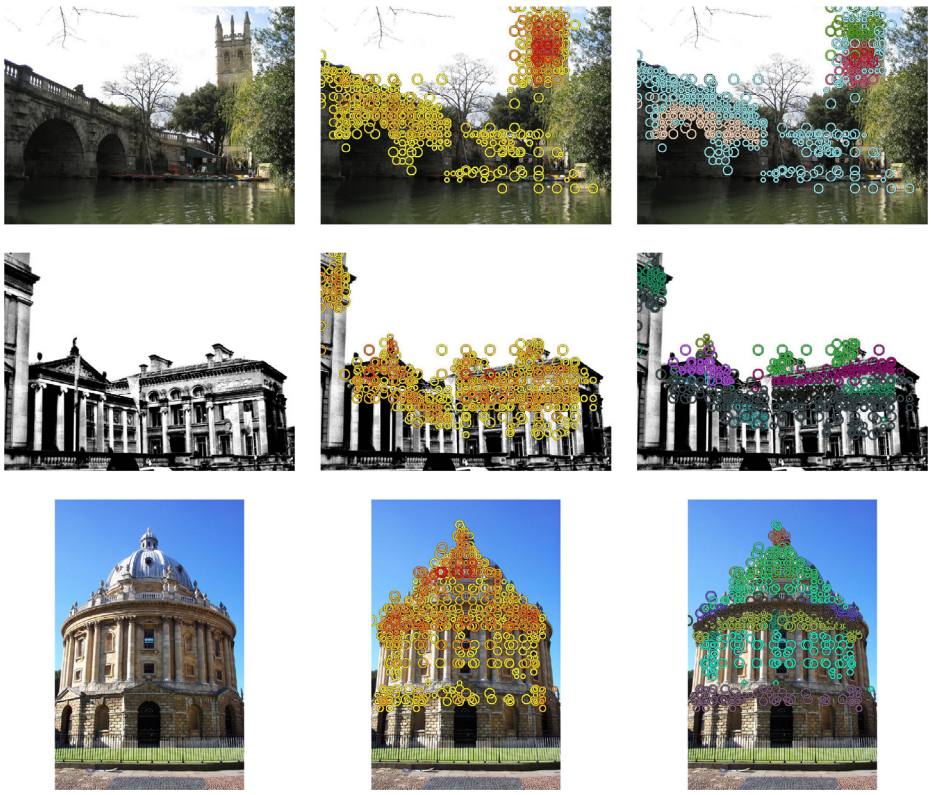


Fig. 6. Visualization of the feature selection and k-means clustering for the proposed co-attention mechanism. The first column represents original images, while the images from the second column show the selected local features marked with circles. The size of the radius in the circles indicates the scale of the image where they originate from. The color variation for circles, from yellow to red, indicates an increasing L2 norm attention score, with red being the highest score. The third column of images shows the result of k-means clustering over selected local features, where the local features assigned to the same cluster are marked by the same color. These examples consider $N = 500$ for feature selection and $K = 10$ for k-means clustering.

and local algorithms for CBIR. We implement our method considering 5 scales: $\left\{\frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}\right\}$. Local features extracted from different scales are merged together and jointly selected using the L2 norm.

6.2. Visualization of feature selection and clustering

Fig. 6 shows the visualization result of the L2 norm based feature selection and k-means clustering over selected local features identified from the images. Selected local features are mainly distributed in regions of the main landmark building displaying architectural details. The most representative parts of the building, like the towers on the top of the building in the top row of images from Fig. 6, have relatively higher attention scores. By applying k-means clustering we group the positions that are visually similar to each other while different parts of the building are assigned to different clusters. Those positions assigned to the same cluster, marked with the same color in the third column of Fig. 6, are considered to share the corresponding clustered local feature from $\mathbf{X}_{c,K} \in \mathbb{R}^{K\times D}$, as their representation. Clustering divides and groups the local image features into fewer but more comprehensive information description features.

6.3. Visualization of co-attention results

Examples of co-attention generation, considering the baseline GeM model trained on GLDv2 dataset, are shown in Fig. 7. The first and second columns show the query and target images, respectively. In the third column, we provide the co-attention generation results, where the local features grouped in the same cluster share the corresponding clustered feature vector as their representation. Co-attention scores for the locations that are not selected are set to zero. Co-attention scores of all local features from different input image scales are projected back to the corresponding locations on the original image and are accumulated to get the final co-attention map. The L2 norm attention of the baseline GeM model is also visualized in the fourth column of Fig. 7 for comparison. As discussed in Section 3.1, the L2 norm reflects the importance of each location with respect to how much it contributes to the final feature vector obtained by global pooling. In other words, the L2 norm is also a query non-sensitive attention that the spatial pooling model implicitly learns during the training.

In examples 1–4 from Fig. 7, some typical retrieval situations are shown in which the target object is not salient or there are similar distractors nearby in the image. The L2 norm attention tends to uniformly highlight all potential relevant regions as it has no access to the actual query information, and its action is only driven by the knowledge learned during training. As a consequence, the L2 norm attention could successfully discard background regions, but it has no idea of which foreground object to consider. In example 4, the L2 norm attention almost ignores the desired query house from the remote part of the scene while wrongly laying most emphasis on the tower building, which is more salient and appears as more significant. Example 5 shows another really hard example, in which the target building is not shown in its entirety, but it is only visible as a small part of the resized tower in the top-left corner of the target image. Moreover, there is a spire from the right side of the target building, which is very similar to the top part of the query object. In this case, the L2 norm mostly highlights the area around that spire, while the proposed co-attention mechanism focuses on the window and edge structure for the correct target object. Examples 6 and 7

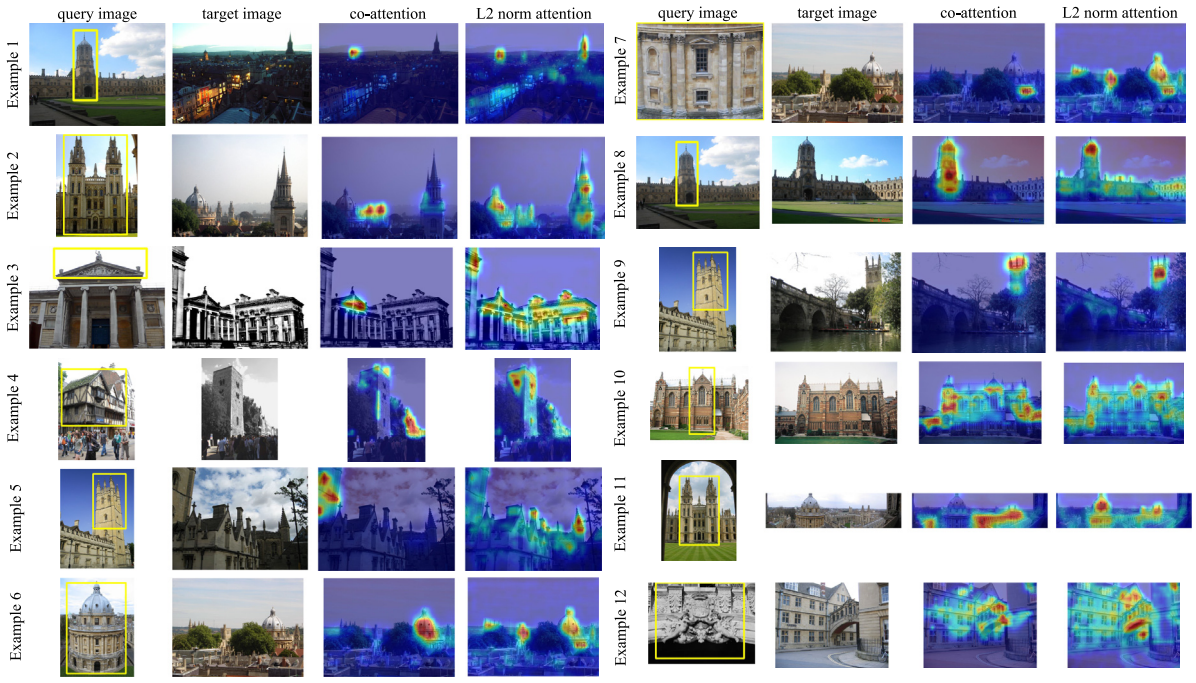


Fig. 7. Attention map visualizations for 12 cases. The first column shows the query image with a yellow bounding box outlining the target object, as provided in the ROxf/RPar dataset. The second column is the target image. The third column represents the co-attention map, while the final column provides the L2 norm attention of the Generalized Mean pooling (GeM).

show the co-attention with the same target image but different query content. In example 6, when considering the whole building as a query, the dome region is central to the co-attention generation. However, in example 7, when using only a window as query, the co-attention correctly focuses on the corresponding region on the target image, despite the dramatic change in the image acquisition conditions. These results indicate the high level of sensitivity of the proposed co-attention method to the query content.

For another set of cases, examples 8 and 9 from Fig. 7 show some easy situations where the target object is salient enough and not surrounded by hard distractors. In such cases, the co-attention mechanism and L2 norm both correctly highlight the target objects despite the challenges in the scene representations due to illumination as well as view perspective changes during image acquisition. Examples 10–12 from Fig. 7 show some cases when the proposed co-attention method fails or does not provide good enough results. Example 12 is one of the hardest cases in which the query content is not even a building but a small sculpture attached as one of the architectural elements on the skyway between two historic buildings. In this case, the co-attention does not highlight just the target region but also the surrounding regions.

6.4. Image retrieval results

Image retrieval results for the proposed method and comparisons with other methods are provided in Table 1. For a fair comparison, some of the recent state of the art (SOTA) works are re-implemented according to the setting from Section 6.1 and marked with "†". The results for DOLG are reported as those revised by the authors.[4] Group (A) from Table 1 shows the results for the local feature methods. R101$^-$-HOW (GLDv2)†[5] is re-implementation of HOW (Tolias et al., 2020) on GLDv2 dataset with ResNet101 backbone and ArcFace loss. Under this re-implementation, it does have a great improvement across all evaluation protocols, especially on ROxf *hard* set, as it reaches 71.3% mAP, up from 56.9% before. However, HOW has a weak performance on RPar+1M dataset with the *hard* evaluation protocol. Group (B) shows the results of the global feature methods. They give worse results than the local feature methods, like HOW (Tolias et al., 2020), on ROxf *hard* set, but they show better generalization ability in the case when considering the 1 million distractor set. The original DELG (Cao et al., 2020) was trained on GLDv2 with a small batch size of 32. R101-DELG† is its re-implemented version with ResNet101 as the backbone network, under the training setting from Section 6.1. It can be seen that the spatial verification gives limited improvement, especially when considering the 1 million distractor set. The bottom group (C) shows the results of the baseline model GeM†, as described in Section 3.2, and when it is combined with the proposed co-attention method (GeM†-CA). In other words, for the results of GeM† and GeM†+CA, they share the same exact GeM backbone network with the training setting from Section 3.2, the only difference is that GeM†+CA implements the co-attention method as described in Section 4.2 (as well as PCA dimension reduction and inverted file indexing from Section 5) to re-weight the candidate image feature tensor before the global GeM pooling. It can be observed that introducing the co-attention to the CBIR pipeline can greatly improve the retrieval performance. Especially on the *hard* set of ROxf (RPar), GeM†+CA reaches the best result 72.6% (85.6%). When considering the 1 million distractor set, the proposed co-attention method still gives the best retrieval results.

6.5. Qualitative retrieval results

Fig. 8 provides a qualitative comparison between the co-attention enabled GeM method "GeM†-CA" and the baseline retrained GeM model "GeM†", on the challenging ROxf dataset (Radenovic et al., 2018), considering the *Hard* evaluation protocol.

[4] https://github.com/feymanpriv/DOLG

[5] R101$^-$ represents the ResNet101 without the final convolution block. According to the study from Tolias et al. (2020), HOW gives better results when discarding the final block, and we follow this setting for our re-implementation.

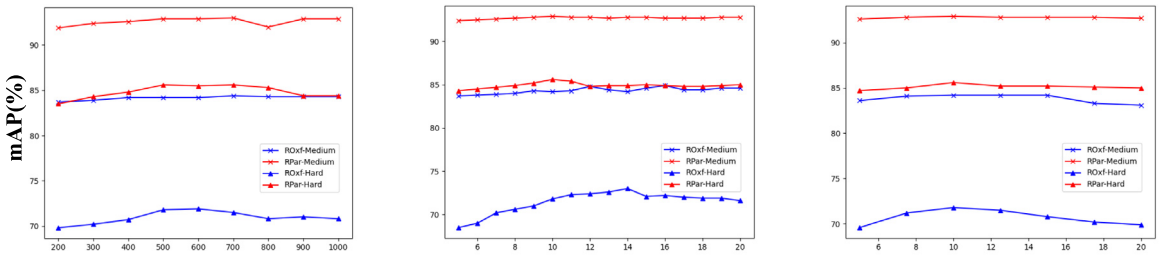
Table 1
Image retrieval results on ROxf/RPar datasets and their extended versions when adding the 1 million distractor set R1M, for the *Medium* and *Hard* evaluation protocols.

Method	*Medium (%)*				*Hard (%)*			
	ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1M
(A) Local feature								
DELF-ASMK*+SP (Teichmann et al., 2019)	67.8	53.8	76.9	57.3	43.1	31.2	55.4	26.4
DELF-D2R-R-ASMK*+SP (Teichmann et al., 2019)	76.0	64.0	80.2	59.7	52.4	38.1	58.6	29.4
R50⁻-HOW-MDA (Wu, Wang, Zhou, & Li, 2021)	82.0	68.7	83.3	64.7	62.2	45.3	66.2	38.9
R50⁻-HOW (Tolias et al., 2020)	79.4	65.8	81.6	61.8	56.9	38.9	62.4	33.7
R101⁻-HOW (GLDv2) †	83.9	77.9	87.9	76.4	71.3	52.8	76.0	56.4
(B) Global feature								
R101-R-MAC (Gordo et al., 2016)	60.9	39.3	78.9	54.8	32.4	12.5	59.4	28.0
R101-GeM (GLD) (Ng et al., 2020)	67.3	49.5	80.6	57.3	44.3	25.7	61.5	29.8
R101-DSM (Siméoni, Avrithis, & Chum, 2019)	65.3	47.6	77.4	52.8	39.2	23.2	56.2	25.0
R101-SOLAR (Ng et al., 2020)	69.9	53.5	81.6	59.2	47.9	29.9	64.5	33.4
R50-DELG (Cao et al., 2020)	73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4
R50-DELG + SP (Cao et al., 2020)	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7
R101-DELG (Cao et al., 2020)	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
R101-DELG + SP (Cao et al., 2020)	81.2	69.1	87.2	71.5	64.0	47.5	72.8	48.7
R101-DELG†	82.4	73.0	90.1	78.0	65.2	50.1	80.6	59.2
R101-DELG + SP†	84.1	75.9	91.0	79.2	68.8	53.6	83.0	62.3
R50-DOLG (Yang et al., 2021)[4]	81.2	71.4	90.1	79.0	62.6	47.3	79.2	59.8
R101-DOLG (Yang et al., 2021)[4]	82.3	73.6	90.9	80.4	64.9	51.6	81.7	62.9
R101-GLAM (Song, Han, & Avrithis, 2022)	78.6	68.0	88.5	73.5	60.2	43.5	76.8	53.1
(C) Proposed co-attention								
R50-GeM†	79.8	69.0	87.3	73.1	60.4	44.2	74.0	52.0
R50-GeM†-CA	83.8	75.3	91.5	77.2	67.8	52.4	82.7	56.8
R101-GeM†	83.0	72.8	90.2	77.6	65.5	49.8	80.7	59.1
R101-GeM†-CA	86.4	79.3	93.2	81.8	72.6	59.9	85.6	64.1

Groups (A) and (B) separately show the results of local and global feature methods, respectively.
Group (C) shows the results of the proposed co-attention method.
"†" indicates re-implemented model using the training details from Section 6.1.
"SP" refers to the spatial verification re-ranking (Noh et al., 2017).

Fig. 8. Top 5 retrieval results for GeM†-CA (with co-attention) and GeM† on images from the *hard* set of ROxf dataset (Radenovic et al., 2018). Co-attention maps are also provided underneath the retrievals provided by "GeM†-CA".

The query image is shown on the first column from the left side of each row with a yellow bounding box indicating the query region of interest. The top 5 retrieval results are shown with the green outline denoting correct retrieval results, while red markings denote incorrect results. The co-attention maps are shown below each row with the retrieved images, with a heatmap indicating the detection of targeted regions for each image. The proposed model outperforms the original GeM model, whose retrieved images are shown underneath. Especially in the top-left query example, the query region is not an intact building, but it only



(a) Ablation study on parameter N (b) Ablation study on parameter K (c) Ablation study on parameter T

Fig. 9. Ablation experiment results when varying the hyper-parameters.

shows a structure from its middle part, and GeM gives wrong results for 3 retrievals out of the top 5. Meanwhile, the proposed co-attention method correctly provides all top 5 retrievals for the specific target region.

7. Ablation experiment and discussion

In this section, we present the ablation experiment results for hyper-parameter settings and discuss the computation costs of the proposed method.

7.1. Impact of clustering parameters

Plots from Figs. 9(a), (b) and (c) show the impact of varying cluster hyper-parameters N, K, and the temperature T from Eq. (8), respectively, on the model retrieval performance. Generally, the proposed method is robust to changes in these hyper-parameters, the difference is mainly reflected on the ROxf *Hard* set. We can observe that using a small $N = 200$ cannot enable using enough local representative features, while a too large $N = 1000$ may pick out too many backgrounds or irrelevant local features, and also it would slow down the feature extraction procedure without bringing any obvious result improvement. Varying the number of clusters K, has implications not only on the performance but also on the computation cost. A smaller K could further reduce the computation cost, but it will arbitrarily fuse many local features into larger clusters reducing the co-attention benefits. A larger number of clusters K can further improve the retrieval performance as it leads to smaller clusters. However, it will require additional computation costs, and the improvement is very limited for $K > 16$.

For clearly showing the impact of parameter T on the co-attention generation, some co-attention maps, generated as described in Section 4.2, but considering different T values in Eq. (8) are shown in Fig. 10. We can observe, that for a small $T = 1$, the co-attention maps based on the clustered candidate image local features tend to cover more contextual regions of the target object. Nevertheless, there are still some unwanted regions. After considering a larger $T = 10$, the co-attention results become more focused on the target object.

7.2. The impact of PCA dimension reduction

The experiment results for the feature vector reduction using PCA are shown in Table 2. According to the results from Table 2, by increasing the feature dimension from the default setting of 512 to 1024 doubles the computation cost without bringing any significant improvement. On the contrary, when considering a feature dimension of 256 or even smaller will lead to significantly lower performance. In conclusion, a feature dimension of 512 is a good balance between the performance and computation cost.

Table 2
CBIR mAP results on ROxf and RPar datasets when varying the feature dimension.

FeatureDimension	*Medium* (%)		*Hard* (%)	
	ROxf	RPar	ROxf	RPar
128	84.1	91.4	68.3	82.5
256	86.0	93.0	71.3	84.4
512	86.4	93.2	72.6	85.6
1024	86.4	93.2	72.7	85.7

Table 3
Retrieval results on ROxf and RPar when considering different image scaling.

1	$\frac{1}{\sqrt{2}}$	$\sqrt{2}$	$\frac{1}{2\sqrt{2}}$	$\frac{1}{2}$	$\frac{1}{4}$	2	*Medium* (%)		*Hard* (%)	
							ROxf	RPar	ROxf	RPar
✔	–	–	–	–	–	–	83.3	89.4	66.3	79.2
✔	✔	✔	–	–	–	–	85.5	91.8	70.6	83.3
✔	✔	✔	✔	✔	–	–	86.4	93.2	72.6	85.6
✔	✔	✔	✔	✔	✔	✔	86.7	93.2	73.3	85.9

7.3. Impact of image scaling

Scaling can account for significant changes in image acquisition, such as under the perspective projection transformations. Retrieval results of the proposed method "GeM†+CA" when considering different image scaling are provided in Table 3. There are 3 different image scaling combinations implemented in the literature : $\left\{\frac{1}{\sqrt{2}}, 1, \sqrt{2}\right\}$ from Radenović et al. (2018), $\left\{\frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}\right\}$ from Yang et al. (2021), and $\left\{\frac{1}{4}, \frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2\right\}$ from Cao et al. (2020), Tolias et al. (2020). According to Table 3, when considering the combination of 5 scales gives the best result for the proposed co-attention method. Using 7 scales does not bring much improvement while increasing the computational cost for the feature extraction.

7.4. Impact of local feature clustering

Apart from the computation cost reduction brought by clustering, we also test implementing the co-attention without considering the clustering, which corresponds to the naïve co-attention case described in Section 4.1. According to the results from Table 4, the co-attention always improves the baseline GeM model's performance, even with the naïve co-attention implementation. Although the proposed clustering procedure forcibly merges many local features into a few groups, which makes it lose some local feature information, it actually provides a positive contribution to the final retrieval performance. This result again proves that considering clustering for co-attention not only relieves the extra computation cost caused by the query-sensitivity search but also further improves the retrieval results.

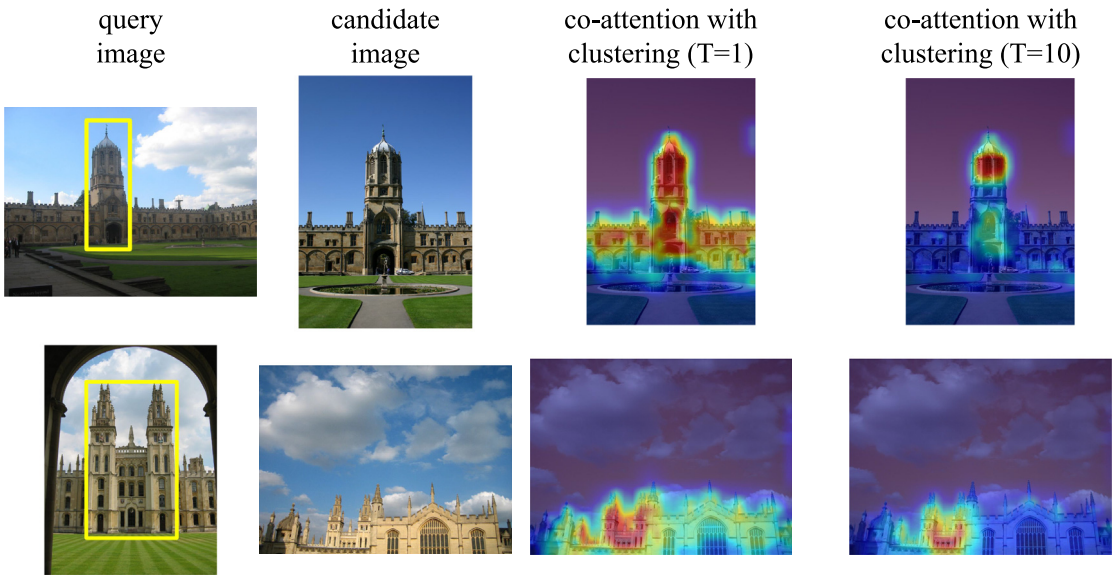


Fig. 10. Co-attention map generated with clustering as described in Section 4.2, when T is set to 1 and 10.

Table 4
CBIR mAP results on ROxf/RPar datasets with naïve co-attention.

Model	Co-attention	Clustering	*Medium* (%)		*Hard* (%)	
			ROxf	RPar	ROxf	RPar
R101-GeM†	✗	✗	83.0	90.2	65.5	80.7
R101-GeM†-CA (naive)	✔	✗	83.7	90.4	69.9	80.9
R101-GeM†-CA	✔	✔	86.4	93.2	72.6	85.6

Table 5
Retrieval results on ROxf and RPar with Spectral Clustering.

Clustering Method	*Medium* (%)		*Hard* (%)	
	ROxf	RPar	ROxf	RPar
Spectral	86.4	93.1	72.7	85.6
k-means++	86.4	93.2	72.6	85.6

Table 6
Retrieval results on ROxf and RPar datasets with Mean-Shift clustering and different bandwidth setting. "Selected features" indicate the average number of local features after clustering by the Mean-Shift.

Clustering method	Band width	Selected features	*Medium* (%)		*Hard* (%)	
			ROxf	RPar	ROxf	RPar
Mean-Shift	0.5	325	87.2	92.3	74.1	83.8
	1.0	76	85.4	91.6	71.7	82.6
	1.5	11	84.1	91.5	68.4	82.6

7.5. Clustering approach

In this ablation study, we provide further discussion about the clustering method selection. The classical *k*-means clustering works well only for convexly distributed data. In the following, we consider two completely different clustering algorithms: Spectral Clustering (Von Luxburg, 2007) and the Mean-Shift (Cheng, 1995). While the former relies on graph spectral analysis, the latter is based on non-parametric kernel-based data representation. The results are provided in Table 5 Spectral clustering is a graph-based clustering method which works well on certain non-convex distributed data. In the spectral clustering implementation, *k*-means is applied over the eigenvectors of the Laplacian of the graph and the cluster number is set to 10. According to the results from Table 5, *k*-means++ and spectral clustering actually give similar results. However, spectral clustering requires significantly more computational requirements and consequently has a higher time cost. The Mean-Shift does not require the manual setting of the cluster center count but requires setting a bandwidth parameter for the kernel (Bors & Nasios, 2009), which is assumed as Gaussian in our implementation. Although by carefully setting the bandwidth value, according to the results from Table 6, the Mean-Shift could give results close to those of the other two clustering methods from Table 5. However, the Mean-Shift eventually keeps more local features and requires more computation costs. From these results, we decided to use *k*-means++ for clustering the co-attention features.

7.6. The impact of re-ranking

The impact of re-ranking on the proposed co-attention enabled CBIR pipeline is explored in this subsection. We consider two different re-ranking methods : α-weighted query expansion (αQE) (Radenović et al., 2018) and diffusion (Iscen, Tolias, Avrithis, Furon, & Chum, 2017). αQE acts on feature vectors of top-ranked n images from the initial retrieval result by applying weighted average and re-normalization. The weight of the ith ranked image descriptor is defined by $(V_q{}^T V_i)^\alpha$ where V_q and V_i are the global feature vectors corresponding to the query image and the ith ranked image. The aggregated feature vector serves as a query descriptor for the second-round retrieval and produces the final retrieval result. Diffusion is another powerful re-ranking method and has been applied in CBIR works (Siméoni et al., 2019). Diffusion could be treated as an extension of the query expansion. Instead of only utilizing top n images as query expansion, based on the first round retrieval results, diffusion explores the nearest neighbors by building a connection graph with similarity scores between each pair of images from the whole database for re-ranking. The retrieval results of the baseline GeM (GeM†) and the proposed co-attention method (GeM†+CA) with these re-ranking methods are presented in Table 7. The proposed method GeM†+CA always gives better retrieval accuracy with or without re-ranking. Specifically, on ROxford hard set, even with re-ranking, GeM† is still outperformed by GeM†+CA without any re-ranking. From the visualization examples shown in Fig. 7, the reason why GeM is not working is not that the query information is not comprehensive enough, but because simply the query non-sensitive feature extraction manner will look at the wrong place in the image for feature extraction, especially when the target

Table 7
Retrieval results on ROxf and RPar datasets with re-ranking. For comparison, the re-ranking results for DELG (Cao et al., 2020) with spatial verification (SP) are also provided.

Method	*Medium (%)*				*Hard (%)*			
	ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1M
R101-GeM†	83.0	72.8	90.2	77.6	65.5	49.8	80.7	59.1
R101-GeM†+αQE	84.4	77.8	91.7	82.7	68.8	56.2	82.8	65.9
R101-GeM†+DF	85.6	79.9	91.9	84.3	69.4	60.1	85.3	69.3
R101-DELG†	82.4	73.0	90.1	78.0	65.2	50.1	80.6	59.2
R101-DELG + SP†	84.1	75.9	91.0	79.2	68.8	53.6	83.0	62.3
R101-GeM†+CA	86.4	79.3	93.2	81.8	72.6	59.9	85.6	64.1
R101-GeM†+CA+αQE	86.9	79.6	93.3	84.5	72.8	60.2	85.7	68.7
R101-GeM†+CA+DF	87.2	81.1	94.4	86.1	73.7	63.9	88.4	72.0

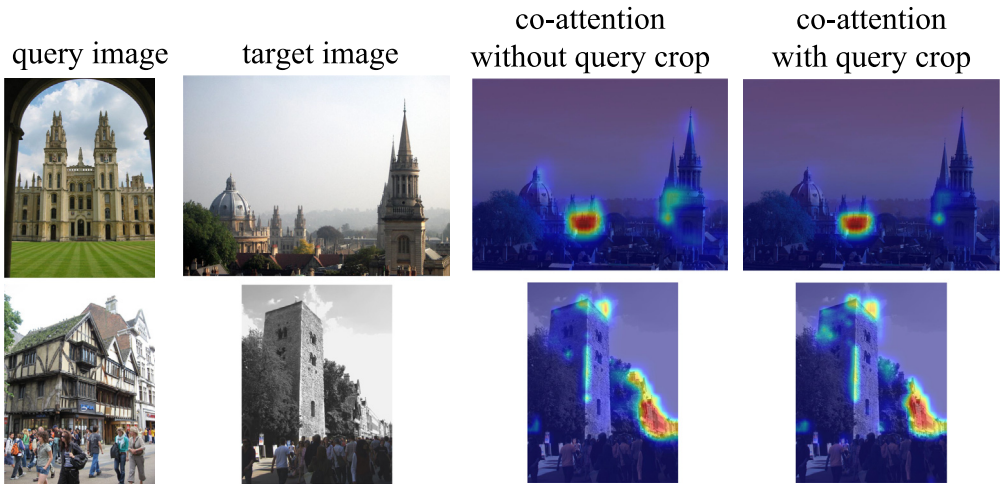


Fig. 11. Co-attention visualization without query crop.

object is not salient. This problem can only be solved by a proper query sensitive attention mechanism, which would force the model to look towards the regions that match the query content, namely the proposed co-attention mechanism.

7.7. Impact of the query clutter

As the standard evaluation protocol of ROxf/RPar dataset is to provide the bounding box for each query image, and by default, all existing research studies utilize the existing bounding boxes to crop the query image and only use the resulting image region as query input. One major concern for the proposed co-attention method is that of overfitting to ROxf/RPar dataset evaluation protocol or whether it is robust enough when the query image contains noise or is cluttered. Fig. 11 visualizes the co-attention map when not cropping the query image. When comparing the co-attention with and without the query crop, it can be observed that there is not much difference in the results, even though the query image in the second row contains a lot of background clutter. Following the discussion from Section 3.1, spatial pooling implicitly implements an L2 norm attention mechanism. Within the proposed co-attention method pipeline, the query image features are selected based on their L2 norms before global pooling. Thus it has a strong robustness to the background clutter that is irrelevant to training data (such as grass, sky, street, the presence of humans, and so on) from the query image.

We do not consider the situation when one query image contains more than one potential object of interest because the benchmark datasets ROxf/RPar (Radenovic et al., 2018) does not include such situations. On the other hand, the search purpose depends on the user; if the input query image contains multiple potential objects of interest, the user is supposed to specify which exact object (or region) needs to be considered for the search, as the CBIR system cannot guess the user intention. If the user would still like to use a whole image that contains multiple training data relevant objects (regions) as query input and uniformly retrieve image content that matches with the query, as the example shown in Fig. 12, then our co-attention will work similarly to the query-nonsensitive attention, uniformly highlighting all training data relevant regions. Table 8 provides the retrieval results of the baseline "GeM†", the proposed method "GeM†+CA" and the current state of the art work DOLG with/without the query image crop. It can be seen that with or without query crop, the co-attention method always improves over the baseline GeM model's results.

7.8. Robustness to the baseline model training

The previous retrieval results are all based on using the GeM model pre-trained on GLDv2 dataset with large data batch size and with the ArcFace margin loss from Eq. (4). In the following, we consider that the baseline model is trained with a much smaller dataset and a simpler loss function in order to test the efficiency of the adapted training approach. In order to test the robustness of baseline model training, we follow the practice from the original GeM pooling paper (Radenović et al., 2018). Another GeM baseline model is trained on rSfM-120k dataset (Radenović et al., 2018), which only contains around 90,000 images. The model is optimized with the simplest contrastive loss (Chopra et al., 2005). Following the setting from Radenović et al. (2018), the batch size is set to 5. Each batch contains 5 image tuples, with each tuple containing 1 query image, 1 positive matching image and 5 negative match images. The hard sample mining is also performed according to the description in Radenović et al. (2018),

(a) An image contains multiple training data relevant object as query

(b) An image contains multiple training data relevant object as candidate

(c) Result co-attention map

Fig. 12. Co-attention visualization when consider a query image that contains multiple training data relevant object.

Table 8
Retrieval results on ROxf and RPar datasets with/without considering the query crop.

Method	Query crop	*Medium (%)*				*Hard (%)*			
		ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1M
GeM†	✗	82.5	78.4	90.7	81.3	62.9	56.1	81.0	65.1
GeM†	✓	83.0	72.8	90.2	77.6	65.5	49.8	80.7	59.1
DOLG (Yang et al., 2021)[4]	✗	83.2	79.0	91.6	82.9	64.8	57.9	82.6	67.3
DOLG (Yang et al., 2021)[4]	✓	82.3	73.6	90.9	80.4	64.9	51.6	81.7	62.9
GeM†+CA	✗	85.5	81.8	93.6	83.9	69.2	61.4	85.8	67.7
GeM†+CA	✓	86.4	79.3	93.2	81.8	72.6	59.9	85.6	64.1

Table 9
Retrieval results on ROxf and RPar datasets when trained on rSfM dataset with contrastive loss.

Backbone	Co-attention	*Medium (%)*		*Hard (%)*	
		ROxf	RPar	ROxf	RPar
Res50	✗	62.6	75.4	39.9	53.1
Res50	✓	69.3	79.2	42.9	57.5
Res101	✗	66.8	78.9	41.8	55.2
Res101	✓	70.1	80.3	45.1	59.5

Table 10
Retrieval results on ROxf and RPar with different selection methods.

Method	*Medium (%)*		*Hard (%)*	
	ROxf	RPar	ROxf	RPar
Max	81.4	90.3	64.8	81.6
Mean	77.4	88.1	58.7	77.2
SoftMax	79.6	89.0	62.5	79.3
GeM†+CA	86.4	93.2	72.6	85.6

and the model is trained for 100 epochs. Then, the proposed co-attention method is applied with the trained GeM model and we provide the experimental results in Table 9. From these results, it can be seen that even with training on a smaller dataset and a simpler loss function, the proposed co-attention still brings positive effects for the retrieval performance. Of course, the improvement is not as impressive as that from above, as proposed in this paper, because when considering a smaller training dataset and a simpler loss function for the training, the feature tensor output by the backbone network is not as efficient and well-trained as when considering a larger dataset GLDv2 and the ArcFace loss for training.

7.9. Why not directly consider the similarity measure in a one-to-many manner?

In the proposed co-attention method, the similarity scores between the query image global feature $\mathbf{V}_q$ and candidate image clustered local features $\mathbf{X}_{c,K}$ are used as co-attention scores to re-weight $\mathbf{X}_{c,K}$ and then perform GeM pooling to get the final candidate image global feature $\mathbf{V}_c$. One concern about the proposed co-attention method could be the necessity of co-attention weighted pooling. Why not just evaluate the similarity measure with $\mathbf{V}_q$ and $\mathbf{X}_{c,K}$ in a one-to-many manner? In the following, three different selection methods are tested to evaluate the final image pair matching score from K local matching similarity scores between $\mathbf{V}_q$ and $\mathbf{X}_{c,K}$. In Table 10, "Max" means choosing the maximum from among K local matching similarity scores as the final image pair matching score while "Mean" means to calculate their average as the final result. "SoftMax" means applying SoftMax function over the K local match scores and then performing the weighted sum over the K local matching similarity scores. All these methods lead to much worse results than the co-attention pipeline proposed in Section 4.2. This is expected because each local feature from $\mathbf{X}_{c,K}$ represents a regional level descriptor which is not comprehensive enough as the representation built by the co-attention weighted global pooling.

7.10. Why not directly perform the similarity measure in a many-to-many manner?

Another concern about the proposed co-attention method is that: why not perform local feature clustering on the query image and then use the clustered query local features (after the PCA dimension reduction) $\mathbf{X}_{q,K} = [\mathbf{x}_{q,k}] \in \mathbb{R}^{K\times D'}$, where $\mathbf{x}_{q,k} \in \mathbb{R}^{1\times D'}$ represents the kth clustered local feature from $\mathbf{X}_{q,K}$ and $D' = 512$, to evaluate the similarity with the clustered candidate image local features $\mathbf{X}_{c,K} = [\mathbf{x}_{c,k}] \in \mathbb{R}^{K\times D'}$ in a many-to-many manner. To perform the many-to-many similarity with clustered local features from the query image and candidate image, as each clustered local feature could be treated as a representation of a corresponding region from the input image, in the following, we define a local patch matching strategy. Firstly, a similarity matrix $\mathbf{M} = \{[m_{i,j}] \in \mathbb{R}^{K\times K}\}$ is obtained by calculating the cosine similarity score between each pair of query local feature $\mathbf{x}_{q,i}$ and local feature candidates $\mathbf{x}_{c,j}$:

$$m_{i,j} = \mathbf{x}_{q,i} \cdot \mathbf{x}_{c,j}. \tag{14}$$

Table 11
Retrieval results comparison on ROxf and RPar with the baseline "GeM†", the proposed co-attention method "GeM†+CA" and the many-to-many local match "GeM†+Local Match".

Method	*Medium* (%)		*Hard* (%)	
	ROxf	RPar	ROxf	RPar
GeM†	83.0	90.2	65.5	80.7
GeM†+CA	86.4	93.2	72.6	85.6
GeM†+Local Match	86.4	92.3	71.5	84.1

Table 12
Retrieval results on ROxf and RPar when considering different feature selection approaches.

$s(\mathbf{b}_{q,i}, \mathbf{I}_c)$ define	$S(\mathbf{I}_q, \mathbf{I}_c)$ define	*Medium* (%)		*Hard* (%)	
		ROxf	RPar	ROxf	RPar
Max	Mean	86.4	92.3	71.5	84.1
Max	Max	78.2	89.0	60.1	76.4
Max	SoftMax	86.3	92.0	71.7	83.5
Mean	Max	55.9	78.1	41.4	64.6
Mean	Mean	77.3	87.9	59.9	76.3
Mean	SoftMax	77.2	87.7	59.5	76.0
SoftMax	Max	62.9	81.8	47.0	68.0
SoftMax	Mean	80.0	89.0	63.1	78.5
SoftMax	SoftMax	79.9	88.8	62.9	78.2
GeM†+CA		86.4	93.2	72.6	85.6

Accordingly, the ith row of the similarity matrix $\mathbf{M}$ stores the similarity score between $\mathbf{x}_{q,i}$ and each local feature from the candidate image feature set $\mathbf{X}_{c,K}$. In principle, the matrix $\mathbf{M}$ needs to be transformed into a single similarity score between local features $\{\mathbf{X}_{q,K}, \mathbf{X}_{c,K}\}$. Thus, the similarity score between a single query local feature $\mathbf{x}_{q,i}$ and the whole candidate image is defined by:

$$s(\mathbf{x}_{q,i}, \mathbf{I}_c) = \max_j m_{i,j}, \tag{15}$$

and eventually, the similarity between clustered local features $\mathbf{X}_{q,K}$ and $\mathbf{X}_{c,K}$, is given by :

$$S(\mathbf{X}_{q,K}, \mathbf{X}_{c,K}) = \frac{\sum_{i=1}^{K} s(\mathbf{x}_{q,i}, \mathbf{I}_c)}{K}. \tag{16}$$

The retrieval result comparison between the baseline GeM, the proposed co-attention method and the many-to-many match method is provided in Table 11. We can observe that by performing many-to-many local matching, as mentioned above, could also improve the baseline GeM model's performance, but the results are still worse than the proposed co-attention method "GeM†+CA".

Apart from the many-to-many definition as mentioned above from Razavian et al. (2016), we further consider different definitions of $s(\mathbf{x}_{q,i}, \mathbf{I}_c)$ and $S(\mathbf{X}_{q,K}, \mathbf{X}_{c,K})$, where by "Max", "Mean" and "SoftMax" represents using the maximum, average or applying SoftMax function overall values then applying the weighted sum, respectively. The corresponding retrieval results are provided in Table 12. Despite considering all these matching possibilities, the many-to-many local matching is always outperformed by the proposed co-attention weighted global feature "GeM†+CA".

7.11. Computation cost and memory requirements

Considering $K = 10$ clusters, feature dimension $D' = 512$, the memory cost to cache one candidate image is $10 \times 512 \times 4$ Bytes ≈ 0.02 MB and it takes around 21 GB to cache the whole ROxf/RPar database considering the 1 million distractor set. The feature extraction takes an average 240 ms to cache one candidate image's local features with 5 scales, including the time cost for the local feature clustering. This could be time consuming, especially for a large database, but it can be performed offline, and it is only done once. With pre-cached features and inverted file indexing, searching on ROxf/RPar with the 1 million distractor dataset for one query image takes around 530 ms with the help of acceleration by an NVIDIA Tesla GPU.

Detailed computation cost requirements and comparison with other models are provided in Table 13. The proposed method "GeM†+CA" requires a similar memory cost as DELG (Cao et al., 2020). When it comes to the retrieval time cost, "GeM†+CA" takes longer than others when considering a Tesla GPU, and is especially slower than GeM and DOLG, because they are simple global feature methods in which each image is only represented by a single global feature vector and the similarity measure is as simple as just calculating the cosine similarity with the global feature vector. However, the proposed method provides the best retrieval performance.

7.12. Additional discussion about the inverted file indexing

The local features selected by the L2 norm but without clustering are used for inverted file indexing implementation. One concern about this approach is that: why not apply inverted file indexing with the clustered local features $\mathbf{X}_{c,K}$. Intuitively speaking, within the co-attention enabled CBIR pipeline, the inverted file indexing is a very general coarse-level filter that tries to filter out candidate images that are unlikely to match the query. When applying the inverted file indexing, it is unwanted to accidentally filter out candidate images that actually are ground-truth matched with the query. Accordingly, the inverted file indexing is implemented with a large codebook size of 65536, as well as a large enough number of local features $N = 500$ from each image, so that any of the candidate image local features shares the same visual representation with any of the query local features. This candidate image will be selected for later co-attention generation and similarity evaluation. The features from the clustered local feature set $\mathbf{X}_{c,K}$ represent a compact and focused representation of relatively high-level semantic meaning, which are deemed as being highly distinguishable between images. According to the testing results, by applying inverted file indexing with clustered local features $\mathbf{X}_{c,K}$ will filter out all unwanted images in the database.

Moreover, the visual word codebook is important in both inverted file indexing and image representation building in HOW (Tolias et al., 2020), as ASMK method used by HOW to build local feature representations is based on using the visual word codebook. On the contrary, in the proposed co-attention enabled CBIR framework, the inverted file indexing only serves as a very general coarse-level filter to initially pick out necessary candidate images for later comparison. In other words, the performance of the proposed co-attention method and image representation building does not actually rely on it. The inverted file indexing module only helps to speed up the retrieval. Although the inverted file indexing setting appears to show that the filter condition is very loose, according to the experiment results, it speeds up the retrieval considerably. To be more specific, the inverted file index, as described in Section 5.2, significantly speeds up the retrieval by filtering out around 70% distractor images from the database. When considering the evaluation dataset ROxf/RPar with 1 million distractor images, each query image only needs to perform the similarity measure with around 300,000 database images. According to the results from Table 14, the proposed method "GeM†+CA" gives almost the same mAP results across ROxf/RPar datasets with or without the inverted file indexing (IVF), while it requires lower computational resources and time. The entire inverted file indexing module hyper-parameter setting, like the codebook size, which is set to 65536, is based on the setting from HOW (Tolias et al., 2020) and is not specifically optimized, given that it already results in a good retrieval speed.

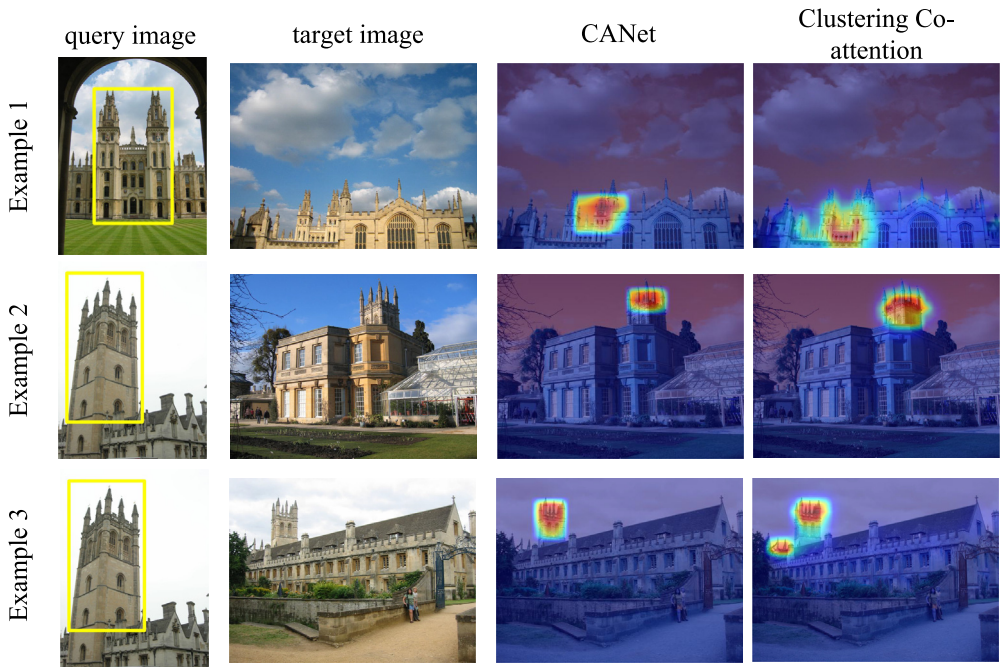


Fig. 13. Attention map visualization. The first column shows the query image with a yellow bounding box outlining the target object. The second column represents the target image. The third column shows the co-attention map generated by CANet, while the fourth column indicates the co-attention map generated by the proposed clustering-based co-attention method.

Table 13
Computation cost comparison.

Method	Device	Memory (GB) ROxf/RPar+1M	Retrieval time (ms) in average
HOW (Tolias et al., 2020)	CPU	14	750
GeM (Radenović et al., 2018)	Tesla GPU	8	250
DOLG (Yang et al., 2021)	Tesla GPU	2	220
DELG+SP (Cao et al., 2020)	Tesla GPU	22	383
GeM†+CA (ours)	Tesla GPU	21	530

Table 14
Retrieval results on ROxf and RPar datasets with/without inverted file indexing (IVF).

Method	IVF	*Medium (%)*				*Hard (%)*			
		ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1M
GeM†+CA	✗	86.4	79.3	93.1	81.8	72.7	59.9	85.7	64.1
GeM†+CA	✓	86.4	79.3	93.2	81.8	72.6	59.9	85.6	64.1

7.13. Comparison with the Conditional Attention Network (CANet)

In the following, we compare the proposed clustering-based co-attention method with CANet (Hu & Bors, 2020). CANet and the proposed co-attention method are both relying on post-processing feature re-weighting modules for pre-trained CNN. For a fair comparison, the GeM model described in Section 3.2 is used as the baseline model. The results on ROxf/RPar datasets are provided in Table 15. We can observe that both methods greatly boost the baseline model's performance, while the "CA (cluster)" gives the best results.

In Fig. 13, we compare the generated attention map results for the method proposed in this paper and for CANet. Both methods can generate rather good query sensitive attention maps. Due to the usage of the convolution layer based fusion module, the CANet tends to highlight a regular square area. Meanwhile, the proposed clustering-based co-attention is based on local feature clustering. As a result, highlighted regions could be irregularly shaped. To limit the computation costs, the cluster count is manually set to a small value ($K = 10$ in the experiments), and some neighbor locations belonging to different objects may inevitably be grouped together, leading to a slight lack of accuracy in the final attention map. In the example 3 from Fig. 13, the region of the target building along with some spire structures at the bottom-left side, which is quite similar to the top parts of the target building, are equally highlighted. On the contrary, in this case, the CANet gives a better attention map. However, as mentioned above, the CANet is globally outperformed, according to the results from Table 15, by the clustering-based co-attention method with respect to the global retrieval accuracy. In addition, CANet requires feeding the query image and candidate image in an online manner, leading to significant extra computation costs at the online retrieval stage. For comparison, CANet takes several hours to search on ROxf/RPar with the 1 million distractor dataset for one query image, while the proposed co-attention method only requires 530 ms, as mentioned in Section 7.11.

In summary, when comparing to the CBIR results by CANet (Hu & Bors, 2020), the clustering-based co-attention method proposed in this paper represents a globally improved query sensitive attention mechanism for CBIR. Although both methods are based on modeling the interaction between the query image global features and the candidate image local features, CANet utilizes stacks

Table 15
Retrieval results on ROxf and RPar datasets.

Method	*Medium (%)*				*Hard (%)*			
	ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1M
R101-GeM† (baseline)	83.0	72.8	90.2	77.6	65.5	49.8	80.7	59.1
R101-GeM†-CANet	84.3	74.5	91.0	78.7	68.9	51.4	82.0	60.8
R101-GeM†-CA (cluster)	86.4	79.3	93.2	81.8	72.6	59.9	85.6	64.1

"R101-GeM† (baseline)" indicate the baseline model as described in Section 3.2.
"R101-GeM†-CANet" represents the baseline model combined with CANet from Hu and Bors (2020).
"R101-GeM†-CA (cluster)" represents the baseline model combined with the clustering-based co-attention method proposed in this paper.

of trainable convolution layers for feature fusion and co-attention generation, which requires significant computation costs at the retrieval stage. Conversely, in the methodology proposed in this paper, the co-attention is intuitively generated by the cosine similarity between the query global feature and clustered candidate image local features. By adopting this approach, we significantly reduce the additional computation costs required by the query sensitivity while still generating high-quality co-attention maps under challenging situations.

8. Conclusion

In this paper, we enable large-scale content-based image retrieval with an efficient co-attention mechanism. The proposed co-attention method can be treated as a non-trainable-parameter module for a pre-trained spatial pooling model. It is intuitively based on the similarity score between the global feature vector of the query image and the clustered local features from the candidate image. The extra computation cost caused by the query sensitivity is addressed by employing local feature clustering while also considering the inverted file indexing to speed up the retrieval procedure. While being straightforward, the proposed co-attention method generates good co-attention maps even under some challenging conditions of image acquisition. By simply embedding our co-attention method with the pre-trained baseline GeM model, the retrieval performance is greatly improved and results in a new state of the art retrieval performance on benchmark datasets requiring comparable computation costs to those of other models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is publically available

Acknowledgments

Dr Adrian G. Bors, would like to thank for the partial support from the COUSIN Project through the Engineering and Physical Sciences Research Council (ESPSRC), U.K. (Grant Number: EP/V009591/1).

References

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & 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).
Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The advantages of careful seeding. In *Proceedings of the annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035).
Babenko, A., & Lempitsky, V. (2015). Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision* (pp. 1269–1277).
Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. In *Proceedings of the European conference on computer vision: Vol. LNCS 8689*, (pp. 584–599).
Bay, H., Tuytelaars, T., & Gool, L. V. (2006). SURF: Speeded up robust features. In *Proceedings of the European conference on computer vision: Vol. LNCS 3951*, (pp. 404–417).
Bors, A. G., & Nasios, N. (2009). Kernel bandwidth estimation for nonparametric modelling. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *39*(6), 1543–1555.
Cao, B., Araujo, A., & Sim, J. (2020). Unifying deep local and global features for image search. In *Proceedings of the European conference on computer vision: Vol. LNCS 12365*, (pp. 726–743).
Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(8), 790–799.
Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 539–546).
Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4690–4699).
DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 224–236).
Gordo, A., Almazán, J., Revaud, J., & Larlus, D. (2016). Deep image retrieval: Learning global representations for image search. In *Proceedings of the European conference on computer vision: Vol. LNCS 9910*, (pp. 241–257).
Gordo, A., Almazán, 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.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
Hsieh, T. -I., Lo, Y. -C., Chen, H. -T., & Liu, T. -L. (2019). One-shot object detection with co-attention and co-excitation. In *Advances in neural information processing systems* (pp. 2725–2734).
Hu, Z., & Bors, A. G. (2020). Conditional attention for content-based image retrieval. In *Proceedings of the British machine vision conference, paper 0356* (pp. 1–13).
Iscen, A., Tolias, G., Avrithis, Y., Furon, T., & Chum, O. (2017). Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2077–2086).
Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3304–3311).
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.
Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.
Manjunath, B. S., & Ma, W. -Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(8), 837–842.
Mohedano, E., McGuinness, K., O'Connor, N. E., Salvador, A., Marques, F., & Giró-i Nieto, X. (2016). Bags of local convolutional features for scalable instance search. In *Proceedings of the ACM international conference on multimedia retrieval* (pp. 327–331).
Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., & Singh, S. (2017). No fuss distance metric learning using proxies. In *Proceedings of the IEEE international conference on computer vision* (pp. 360–368).
Munjal, B., Amin, S., Tombari, F., & Galasso, F. (2019). Query-guided end-to-end person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 811–820).

Ng, T., Balntas, V., Tian, Y., & Mikolajczyk, K. (2020). SOLAR: Second-order loss and attention for image retrieval. In *Proceedings of the European conference on computer vision*: *Vol. LNCS 12370*, (pp. 253–270).

Noh, H., Araujo, A., Sim, J., Weyand, T., & Han, B. (2017). Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision* (pp. 3456–3465).

Papushoy, A., & Bors, A. G. (2015). Image retrieval based on query by saliency content. *Digital Signal Processing*, *36*(1), 156–173.

Park, M., Jin, J. S., & Wilson, L. S. (2002). Fast content-based image retrieval using quasi-Gabor filter and reduction of image feature dimension. In *Proceedings of the IEEE southwest symposium on image analysis and interpretation* (pp. 178–182).

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

Radenovic, F., Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2018). Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5706–5715).

Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(7), 1655–1668.

Razavian, A. S., Sullivan, J., Carlsson, S., & Maki, A. (2016). Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, *4*(3), 251–258.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Siméoni, O., Avrithis, Y., & Chum, O. (2019). Local features and visual words emerge in activations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11651–11660).

Song, C. H., Han, H. J., & Avrithis, Y. (2022). All the attention you need: Global-local, spatial-channel attention for image retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2754–2763).

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Teichmann, M., Araujo, A., Zhu, M., & Sim, J. (2019). Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5109–5118).

Tolias, G., Avrithis, Y., & Jégou, H. (2016). Image search with selective match kernels: Aggregation across single and multiple images. *International Journal of Computer Vision*, *116*(3), 247–261.

Tolias, G., Jenicek, T., & Chum, O. (2020). Learning and aggregating deep local descriptors for instance-level recognition. In *Proceedings of the European conference on computer vision*: *Vol. LNCS 12346*, (pp. 460–477).

Tolias, G., Sicre, R., & Jégou, H. (2016). Particular object retrieval with integral max-pooling of CNN activations. In *Proceedings of the international conference on learning representations* (pp. 1–12). arXiv preprint arXiv:1511.05879.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).

Wang, Q., Lai, J., Claesen, L., Yang, Z., Lei, L., & Liu, W. (2020). A novel feature representation: Aggregating convolution kernels for image retrieval. *Neural Networks*, *130*, 1–10.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. (2019). Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1328–1338).

Weyand, T., Araujo, A., Cao, B., & Sim, J. (2020). Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2575–2584).

Wu, X., Irie, G., Hiramatsu, K., & Kashino, K. (2018). Weighted generalized mean pooling for deep image retrieval. In *Proceedings of the IEEE international conference on image processing* (pp. 495–499).

Wu, H., Wang, M., Zhou, W., & Li, H. (2021). Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11416–11425).

Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., et al. (2021). DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11772–11781).

Yang, X., Wang, N., Song, B., & Gao, X. (2019). BoSR: A CNN-based aurora image retrieval method. *Neural Networks*, *116*, 188–197.

Yue-Hei Ng, J., Yang, F., & Davis, L. S. (2015). Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 685–701).