Proceedings Paper:
Hu, Zechao and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2020) Conditional Attention for Content-based Image Retrieval. In: British Machine Vision Conference (BMVC). , Manchester, UK .

Conditional Attention for Content-based Image Retrieval

Zechao Hu
zh955@york.ac.uk
Adrian G. Bors
adrian.bors@york.ac.uk

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

Abstract

Deep learning based feature extraction combined with visual attention mechanism is shown to provide good results in content-based image retrieval (CBIR). Ideally, CBIR should rely on regions which contain objects of interest that appear in the query image. However, most existing attention models just predict the most likely region of interest based on the knowledge learned from the training dataset regardless of the content in the query image. As a result, they may look towards contexts outside the object of interest, especially when there are multiple potential objects of interest in a given image. In this paper, we propose a conditional attention model which is sensitive to the input query image content and can generate more accurate attention maps. A key-point detection and description based method is proposed for training data generation. Consequently, our model does not require any additional attention label for training. The proposed attention model enables the spatial pooling feature extraction method (generalized mean pooling) improves image feature representation and leads to better image retrieval performance. The proposed framework is tested on a series of databases where it is shown to perform well in challenging situations.

1 Introduction

Content based image retrieval (CBIR) aims to find the most similar images to a given query image. Due to the variation in the image content, a simple comparison of the pixel representations could not provide an appropriate result. The main challenge in CBIR systems is the ambiguity in the high-level (semantic) concepts extracted from the low-level (pixels) features of the image. In earlier conventional CBIR systems, image features are normally described by a hand-crafted feature extractor, which is based on sets of low-level features, such as colour [29], texture [15], shape [3] or gradient [14], or by modelling the visual attention on top of the features' representation, [20]. However, by using low-level feature based methods we are not able to fill the gap between the low-level representations and the high-level semantic meaning [35]. To solve this problem, a series of convolution neural network (CNN) based methods are proposed to extract compact semantic-aware image representation for the CBIR task. The Neural Code model [2] fine-tunes a pre-trained AlexNet [11] on the relevant dataset and directly uses the fully connected layers' outputs as the feature vector for image retrieval. Other approaches include the hash codes [13], bag of visual words [16] and spatial pooling [25, 26, 30, 34].

All CNN-based methods used so far for CBIR rely on uniformly transforming the final convolution layer's output into a feature vector, lacking the ability to localize and focus on the region of interest (ROI). Such approaches are likely to be misled by irrelevant information, while missing the relevant information during the image retrieval. Hence, methods considering weighting mechanisms or visual attention are proposed to address these drawbacks. Bags of local convolutional features (BLCF) [17] combines saliency weighting over local CNN features by using element-wise multiplications for instance search. This design improves model's performance but suffers from the inconsistency between defining the the human attention and the actual matching regions between two images. The DEep Local Features (DELF) [19] employs an attentive local feature descriptor for image retrieval with a tightly coupled attention mechanism which can score and select most relevant local features for image matching. The weighted generalized mean pooling (wGeM) [33] applies a trainable spatial weighting mechanism over the activation of the last convolution layer to describe how important each activation's location is for image retrieval. However, the attention module in wGeM model actually does not attempt finding the image region with the highest probability of relevance to the query image but just predicts the most likely region of interest based on the knowledge learned from the training dataset. In some cases, the attention module would fail, looking for regions which are outside the objects of interest [33].

Overall, the main contributions of this paper are listed as follows:

- We propose a new conditional attention model for localizing the region of interest (ROI) from the candidate image that matches the content of the query image. The proposed attention model is sensitive to the input query content and it can be combined with existing feature extraction method to boosts original method's retrieval performance.
- We consider that repeating scene details in various images, represent important clues for that scene. We then use the pre-trained key-point detector SuperPoint [4] to find correspondences of matching image pairs which is used for generating training data for the proposed CBIR conditional attention model.

We show that our attention model can generate accurate attention maps for candidate images based on the content in the query image, even when there are actually multiple potential objects of interest in a given image. When combined with the generalized-mean (GeM) pooling from [25], our attention model can always improve the original feature extraction method's performance and lead to the state of art results in some evaluation datasets. The related work is outlined in Section 2 The proposed Conditional Attention Network and training data generation is described in Section 3 and how this is embedded into the CBIR pipeline is explained in Section 4. Experimental results are provided in Section 5 and the conclusions are drawn in Section 6.

2 Related work

Spatial pooling. Early CNN based image feature extraction for CBIR, like the Neural Code model [2], implements fully connected (FC) layers transforming the 3D feature tensor output of the last layer to a fixed length feature vector. Razavian *et al.*, [26] perform spatial pooling to get a compact feature vector from the 3D convolution feature tensor of a CNN. Later, different types of global pooling are proposed for the CBIR such as sum pooling [34], max pooling [30] and the generalized mean (GeM) pooling [25]. Compared with the FC layer based method, spatial pooling is not sensitive to the input image size. It can process images of any size, without any cropping or change in the aspect ratio. These global pooling

methods are then combined with a Region Proposal Network (RPN) selection mechanism [5] or employs an end-to-end trainable weights mechanism [33]. However, these attention mechanisms are not sensitive to query content and may look outside of target object.

Co-attention. There are already some query sensitive attention models proposed for different image recognition tasks. The query-guided end-to-end person search network (QEEPS) [18] implements a Query-guided Region Proposal Network (QRPN), leveraging query-ROI-Pooled features to emphasize discriminant patterns in the target image to produce relevant solutions. The co-attention and co-excitation (CoAE) framework [8] implements the non-local operation [32] to fuse the features from the target and query images, generating query relevant region proposal for one-shot object detection tasks. The SiamMask [31] uses depth-wise cross-correlation to fuse features from the query and search images. Then the response map is fed into convolution layers to generate pixel-wise binary masks for visual object tracking. These co-attention architectures involve different feature fusion methods and require extra bounding box annotation for a region proposal network training. Meanwhile, our conditional attention network uses layers of convolutions for feature fusion and it is trained using automatically generated data.

3 Conditional Attention Network and training data

In the following we describe the characteristics of the Conditional Attention network architecture and the generation of the training data using the attention maps.

3.1 Network architecture

We develop a conditional attention model which defines the region of interest (ROI) in candidate images under the condition of the content in the query image. The architecture of our Conditional Attention Network is shown in Fig. 1. The conditional attention map generation pipeline consists of three processing stages: visual encoding, feature fusion and the attention map generation.

Visual feature encoding. The proposed attention model takes a candidate image and a query region of interest (ROI) image as input. We consider a VGG16 network [28], without the fully connected layers and the last max-pooling layer, as the backbone network to extract the visual feature information from input images, as shown in the upper-left side of Fig. 1. Given the candidate image $\mathrm{I_c}$ of size $H_c \times W_c$, the output of the VGG16 network is a 3D tensor $\mathrm{X_c} \in \mathbb{R}^{512 \times \frac{H_c}{16} \times \frac{W_c}{16}}$, where 512 is the number of feature channels. The query ROI image $\mathrm{I_q}$ of size $\mathrm{H_q} \times \mathrm{W_q}$, is also fed into the same backbone network and outputs a 3D tensor $\mathrm{X_q} \in \mathbb{R}^{512 \times \frac{H_q}{16} \times \frac{W_q}{16}}$. In order to obtain the global feature vector of the query ROI, we implement the channel-wise global average pooling (GAP) to construct a compact feature vector $\mathrm{V_q}$ from the 3D feature tensor $\mathrm{X_q}$. Let X_q^n be the n-th channel of $\mathrm{X_q}$ where $n \in \{1,2,\ldots,512\}$, the GAP is given by:

$$\mathrm{V}_q = [{v_\mathrm{q}}^1 \ldots {v_\mathrm{q}}^n \ldots {v_\mathrm{q}}^{512}], \quad {v_\mathrm{q}}^n = \frac{1}{|\mathrm{X_q^n}|} \sum_{x \in \mathrm{X_q^n}} x \tag{1}$$

Feature fusion. To fuse a one-row query ROI feature vector with a 3D feature tensor of the candidate image, we firstly consider the ROI feature vector to compare with the information at each location of the 3D feature tensor from the candidate image. The ROI feature vector and the 3D feature tensor are separately location-wise L2-normalised before being concatenated. Then, the concatenated feature tensor is fed into the fusion module. Within

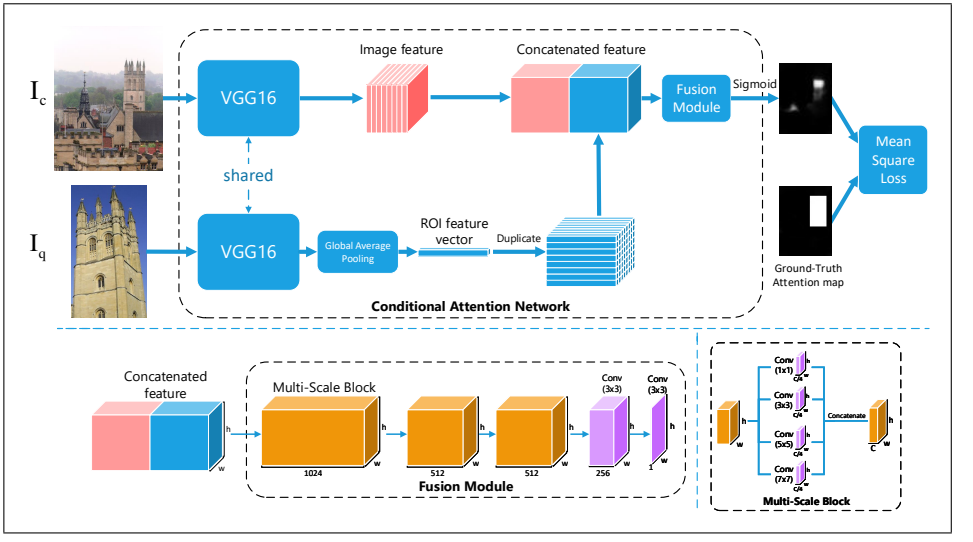

Figure 1: Architecture of the proposed Conditional Attention Network.

the fusion module, the concatenated feature tensor is first processed sequentially through three multi-scale convolution blocks, shown in orange at the bottom left part of Fig. 1. Each multi-scale convolution block consists of 4 convolution layers with different kernel sizes and with all outputs of the convolution layers concatenated as the block's final output, as illustrated in the bottom right part of Fig. 1. The output channel count of the first multi-scale block is 1024 while that of the other two is 512. After that, we use two 3×3 convolution layers for dimension reduction, shown in violet at the bottom of Fig. 1, and finally obtain a one-channel feature map.

Attention maps. After the fusion step, we use a sigmoid activation function to normalize each location value on the one-channel feature map to the range of (0,1) and generate the final attention map for the candidate image I_c under the condition of the query image content from I_q. The attention map models the likelihood that each location from I_c that matches with the I_q with the precision of $\left(\frac{H_c}{16}, \frac{W_c}{16}\right)$.

We train the network using a large number of image pairs, where each pair represents the context of the same scene with annotated matching ROIs and corresponding ground-truth attention maps, as explained in Section 3.2. Let us consider A be the generated attention map of the candidate image I_c conditioned by the information from the query's ROI I_q, while $\hat{A}$ is the ground-truth attention map. We then consider the mean square error (MSE) between the corresponding image regions as the loss function:

$$MSE(A, \hat{A}) = \frac{1}{|K|} \sum_{k \in K} (A_k - \hat{A}_k)^2, \tag{2}$$

where K represents all locations from the attention map and A_k is the attention map value at location $k \in K$.

3.2 Training data generation

In the following we assume that we have pairs or sequences of corresponding images, representing sections of the same scene, but which have been acquired at different times, under

different conditions and characterized by different image acquisition parameters. A good example of such data is the image tuple dataset from [23, 25] which contains a sizeable number of annotated matching image pairs. The image pairs, displaying parts of the same scene, can be used to find the corresponding regions. By finding the correspondences between the image pairs, we generate query ROIs and corresponding ground-truth attention maps, which serve as the training data for the Conditional Attention network.

In order to find the correspondences between identical regions from the given paired images we use an intermediate feature descriptor. The SuperPoint [4] network is able to extract local image descriptors and find key-point correspondences among matching images. To obtain robust key-points, for each matching image pair, both query and positive images are separately resized keeping the original image aspect ratio. Then we perform key-point matching at the resolution for each of the images. In our implementation, we consider 4 different resolutions with $\{128, 256, 362, 512\}$ for the long side, so we obtain 4×4 maps of key-point matches for each image pair. Matching key-points of the query and positive images at different scales are separately projected to the key-point map $\mathrm{M_Q}$ of size $H_{MQ} \times W_{MQ}$ and $\mathrm{M_P}$ of size $H_{MP} \times W_{MP}$ while keeping the original aspect ratio. Choosing the right size for the key-point map is important. If the key-point map size is too small, the precision of the generated ROI will be very low, while if the key-point map size is too large, then the key-points will be too sparse to localize and represent the appropriate ROIs in the images. During the training stage, we resize all positive image and query ROI to a maximum size of 362×362 while keeping the original image ratio. After processing by the fully convolutional VGG16 architecture and being down-sampled 4 times by 4 max-pooling layer, where each max-pooling layer will reduce the size of its input to half, the output generated attention map of our Conditional Attention Network has a maximum size of 22×22 ($\frac{362}{16} \approx 22$). The size of the ground-truth attention map is supposed to be equal to that of the generated attention map for the mean square error calculation. By taking all these aspects into consideration we set the long side of both key-point maps $\mathrm{M_Q}$ and $\mathrm{M_P}$ to be 22 while keeping the original image's aspect ratio. According to the empirical results, this setting can generate accurate ROIs and attention maps which would also meet the calculation requirements for the mean square error from equation (2).

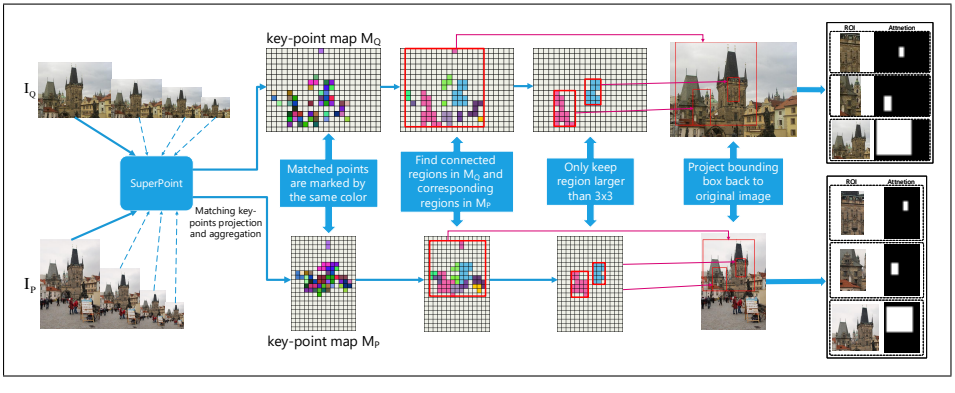

Figure 2: The pipeline for training data generation. The selected matching regions are projected back into the original images in order to define the ROI. The long side of all key-point maps and the final generated ground-truth attention map is 22 while preserving the original image ratio.

By using the matching key-points we consider two criteria for defining the regions of interest: 1. the region is defined within the top-left and bottom-right key-points; 2. the region is defined by connected key-point regions from the key-point map which is larger than 3×3 pixels. Then we label all locations within ROIs by 1 and with 0 otherwise in order to create the final binary ground-truth attention map. The pipeline for calculating the matching regions and defining ROIs, is shown in Fig. 2. Depending on the SuperPoint model output, each image pair can generate several matching pairs for the ROI and the corresponding ground-truth attention map. In other words, one positive image pair can derive several sets of (I_q, I_c, $\hat{A}$) for training the Conditional Attention network.

4 Embedding the conditional attention model into the CBIR pipeline

The proposed Conditional Attention model, described in the previous section represents a completely independent module which can be integrated into a deep learning CBIR model. Recently, spatial pooling was successfully used for feature extraction from images while the Generalized Mean pooling (GeM) [25], provides the state of the art performance on common image retrieval evaluation datasets. In the following we explain how to embed the proposed conditional attention map model into the original GeM feature extraction pipeline.

GeM feature extraction model contains two parts: a convolutional neural network (CNN) for the 3D feature map extraction and a generalized mean pooling layer to transform the 3D feature map into a compact feature vector. The dimension of the resulting vector represents the channel count of the 3D feature map. Assume that the 3D feature map of the candidate image I_c extracted by the GeM backbone network is F of size $C \times H_F \times W_F$ and the attention map of I_c under the condition of query ROI I_q generated by our attention network is A, which is resized to $1 \times H_F \times W_F$. We use the following equation from [12] to mask F with A:

$$F' = F \odot [(1-\theta)A \oplus \theta] \tag{3}$$

where $\odot$ and $\oplus$ denote element-wise multiplication and addition, respectively, while $\theta = 0.5$.

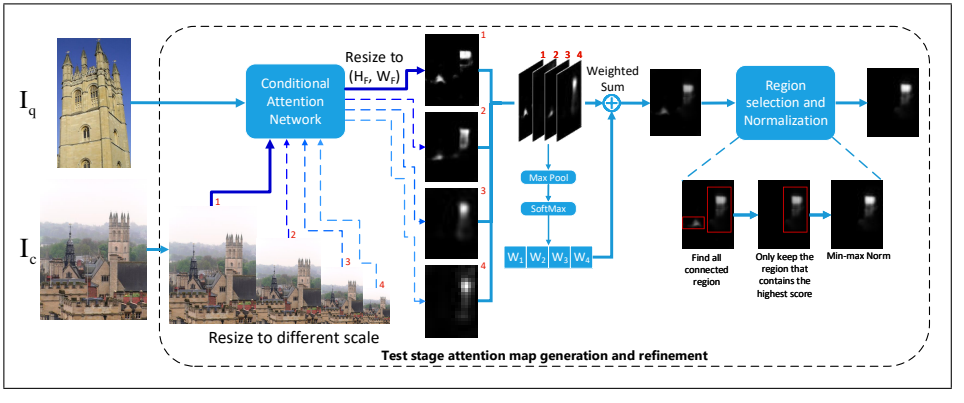

Figure 3: Attention map generation and refinement with the multi-scale scheme during the testing stage.

During the testing stage, in order to obtain more accurate attention maps, we refine the initially generated attention map before combining it with GeM features [25]. In Fig. 3, we show that the query ROI image I_q is fed into the Conditional Attention Network, together

with the candidate image I_c represented at 4 different scales $\{362, 512, 1024, 2048\}$ for the long side, while preserving the initial aspect ratio. All attention maps generated at different scales are resized to the same resolution and then weighted and their values added together. The weights are evaluated by implementing max pooling and the Softmax activation function on the attention maps. After that, we preserve the connected highlighted region that contains the highest score, as shown in Fig. 3, and normalize the attention map using the following min-max normalization equation:

$$\mathbf{X}' = \frac{\mathbf{X} - \mathbf{X}_{\min}}{\mathbf{X}_{\max} - \mathbf{X}_{\min}}, \tag{4}$$

where $\mathbf{X}$ represents the original tensor, $\mathbf{X}_{min}$ and $\mathbf{X}_{max}$ are the minimum and maximum value in $\mathbf{X}$, respectively.

In addition, as query expansion has been widely used for image retrieval (e.g [6, 9, 30, 34]), we use the α-weighted query expansion (αQE) [25] for retrieval result reranking. αQE acts on feature vectors of top-ranked nQE images from the initial retrieval result by applying weighting averaging and renormalization. The weight of the i-th ranked image descriptor is defined by $(\mathrm{V_q}^{\mathrm{T}}\mathrm{V}_i)^{\alpha}$ where $\mathrm{V_q}$ and V_i are the feature vectors corresponding to the query image and the i-th ranked image. The aggregated feature vector serves as a query descriptor for the second-round retrieval and produces the final retrieval result.

5 Experiments

In this section we discuss the implementation details of training, evaluation setting and compare the results obtained by the proposed Conditional Attention model to other CBIR approaches.

5.1 Training setup and implementation details

Our conditional attention model is trained with Adam [10], using an initial learning rate $l_0 = 10^{-3}$, an exponential decay $\exp(-0.1i)$ over epoch i, momentum = 0.9 and weight decay = 5×10^{-4}. The experiments are performed on an NVIDIA Titan XP GPU. In order to find the correspondences between images representing the same scene, we consider the image tuple dataset from [23, 25], which contains 91,642 images divided into 551 clusters, while 181,697 matching image pairs are annotated. During the training, at each training step we input a tuple of images. Each training image tuple consists of 1 query image, 1 positive image and 5 negative images. In other words, 1 training tuple contains 6 image pairs. Within each tuple, given the positive image pair, we can generate several pairs of query ROI and corresponding ground-truth attention maps $\mathrm{I_q}$ and $\hat{\mathrm{A}}$, respectively, as described in Section 3.2. These query ROIs and ground-truth attention maps are then used for training. When considering each negative image pair, we enforce that $\mathrm{I_q}$, defined through positive matches, would not match any region within the negative image. In this case $\hat{\mathrm{A}} = \mathbf{0}$. The training is performed for 100 epochs. For each epoch, 1200 image tuples are randomly selected from the image tuple dataset with a batch size of 5 training tuples.

For the GeM feature extraction model, we directly use the pre-trained GeM network provided in [25], which has been fine-tuned on the image tuple dataset. The GeM network with VGG16 [28], and Resnet101 [7], are tested as backbone architectures in our experiments.

5.2 Evaluation datasets

For the evaluation experiments, we consider 6 benchmark databases for image retrieval performance evaluation: Oxford5k [21], Paris6k [22], Oxford105k [21], Paris106k [22], ROxford5k [24] and RParis6k [24]. Oxford5k contains 5062 images which are collected from Flickr with 17 tags of buildings from Oxford. All images are manually annotated and this gives 55 images in all as queries for the image retrieval evaluation. Paris6k, consisting of 6412 images, is also collected from Flickr by searching for 12 Paris landmark tags and it also gives 55 query images. Oxford105k and Paris106k are expanded versions of the Oxford5k and Paris6k by adding additional 100K distractor images from Flickr. ROxford5k and RParis6k are revisited versions of Oxford5k and Paris6k, with each containing 15 extra new challenging queries while the potential positive images of each query are arranged into 3 groups with different difficulty levels of *Easy, Medium, Hard*. All these 6 evaluation datasets provide bounding boxes with the ROI for each query image. Following the standard evaluation protocol, we crop each query image with its bounding box and the cropped query image is fed into the GeM [25] network to get the feature vector for each query image. For each candidate image, when comparing its similarity with each query image, its attention maps conditioned by each cropped query image are separately generated and combined with its convolution feature, as described in Section 4. As the output feature vector of the GeM network is L2-normalized, the inner product is used for calculating the similarity measure. During the evaluation, all input images are limited to a maximum size of 1024×1024. We also implement the learned whitening and the multi-scale representation schemes, proposed in [25], for better image retrieval performance. The mean average precision (mAP) [21] is used as a performance measure for the results on all datasets.

5.3 Content Based Image Retrieval (CBIR) results

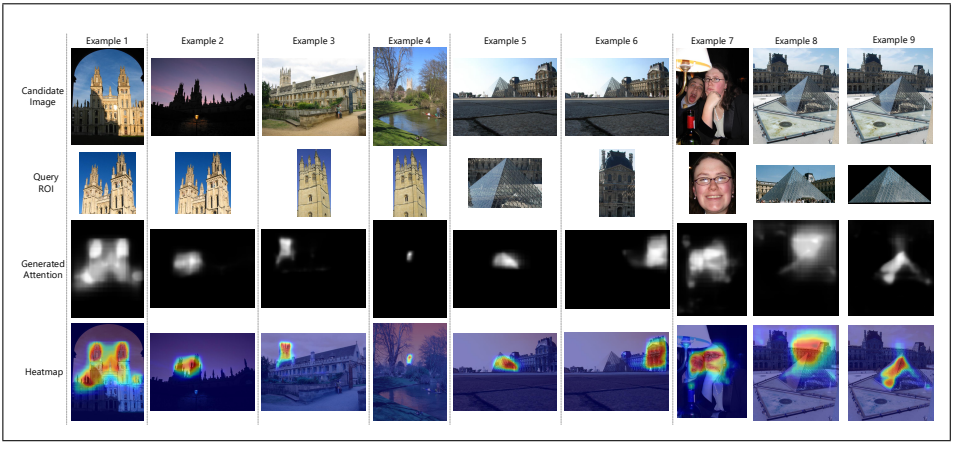

Figure 4: Attention map results for the proposed conditional attention model. Candidate images and the query ROIs are displayed in the first and second rows, respectively. Third and fourth rows represent the generated attention maps and corresponding heatmaps, after refining, min-max normalization and up-sampling to the original image size.

In Fig. 4 we show some examples of generated attention maps when considering various candidate image and query ROI pairs. Scene examples 1-4 show that our attention model can accurately locate the target object under a variety of challenging situations, such as when the

images are characterised by different acquisition parameters, changes in the light condition or when the object of interest is small and far away. In the examples 5 and 6 we can see the generated attention map for the same candidate image but considering different query ROIs. Unlike the wGeM failure example, shown in [33], our model can correctly highlight the target query object based on the input query ROI, even when there are two potential objects of interest in the same image. The example 7 shows how the proposed conditional attention model works with unseen image content in the training procedure. Although the network was trained with architecture images, it can be also used to retrieve human faces. Example 8 shows an example where our model fails. Because we use global average pooling to extract the features from query ROI, if the query ROI contains too much distraction content, the retrieval could fail. As shown in example 9, if we manually crop the target pyramid and then set its background to zero, the generated attention map improves.

Net	Method	Attention map refined	Oxford5k	Paris6k
VGG16	GeM [34]	-	87.9	87.7
	*GeM+CA	No	88.5	88.8
	*GeM+CA	Yes	88.7	88.9

Table 1: Image retrieval performance (mAP) comparison when considering the attention map refinement and without.

For the retrieval performance evaluation, we first evaluate the effect of attention map refinement as shown in Fig. 3. Image retrieval performance (mAP) results are provided in Table 1, and we can observe that the attention map refinement improves mAP by 0.2 on Oxford5k and 0.1 on Paris6k.

Net	Method	Fine-tuned	Oxford5k	Oxford105k	Paris6k	Paris106k
VGG16	SPoC [34]	No	68.1	61.1	78.2	68.4
	CroW [9]	No	70.8	65.3	79.7	72.2
	BoW-CNN [16]	No	73.9	59.3	82.0	64.8
	NetVLAD [1]	Yes	71.6	–	79.7	–
	R-MAC [6]	Yes	83.1	78.6	87.1	79.7
	GeM [25]	Yes	87.9	83.3	87.7	81.3
	*GeM+CA	Yes	**88.7**	**84.5**	**88.9**	**84.1**
Res50	DELF [19]	Yes	83.8	82.6	85.0	81.7
Res101	R-MAC [6]	Yes	86.1	82.8	**94.5**	**90.6**
	GeM [25]	Yes	87.8	84.6	92.7	86.9
	WGeM [33]	Yes	88.8	85.6	92.5	–
	*GeM+CA	Yes	**89.4**	**86.2**	93.0	87.1
Re-Ranking (R) and Query Expansion (QE)						
VGG16	CroW+QE [9]	No	74.9	70.6	84.8	79.4
	BoW-CNN+R+QE [16]	No	78.8	65.1	84.8	64.1
	R-MAC+QE [6]	Yes	89.1	87.3	91.2	86.8
	GeM+αQE [25]	Yes	91.9	89.6	91.9	87.6
	*GeM+CA+αQE	Yes	**93.1**	**90.1**	**92.9**	**88.9**
Res50	DELF+QE [19]	Yes	90.0	88.5	95.7	92.8
Res101	R-MAC+QE [6]	Yes	90.6	89.4	96.0	93.2
	GeM+αQE [25]	Yes	91.0	89.5	95.5	91.9
	WGeM+QE [33]	Yes	91.7	89.7	96.0	–
	*GeM+CA+αQE	Yes	**91.9**	**90.2**	**96.4**	**93.3**

Table 2: Image retrieval performance (mAP) comparison on Oxford5k, Oxford105k, Paris6k and Paris106k dataset. Fine-tuned indicate whether the model is only off-the-shelf, trained on ImageNet [27], or fine-tuned on other training datasets. * marks our method and it is always implemented with learned whitening, multi-scale representation scheme [25] and the attention map refinement from Fig. 3. The highest mAP score is highlighted in bold.

The retrieval results on Oxford5k, Paris6k, Oxford105k and Paris106k are shown in Table 2. The combination of our Conditional Attention Network (CA) with GeM feature extraction (GeM+CA) can always improve over the original GeM method's performance. When VGG16 [28] is used as the backbone network of GeM, our method outperforms other methods shown in the table. When Resnet101 [7] is implemented as the backbone network of GeM and combined with αQE [25], our method GeM+CA+αQE provides the best results on these four datasets. For the query expansion we set nQE = 10 for Oxford, nQE = 50 for Paris and α = 3. In Table 3 we provide the results on ROxford5k and RParis6k datasets. Our attention model can still boost the original GeM method's performance and was only outperformed by DELF [19] model. However, DELF is a local descriptor based feature representation method which is trained on Google landmark dataset [19], which is a much larger dataset than the image tuple dataset used for training our attention model and GeM.

Net	Method	Fine-Tuned	Roxford5k				Rparis6k			
			Medium		Hard		Medium		Hard	
			mAP	mAP@10	mAP	mAP@10	mAP	mAP@10	mAP	mAP@10
VGG16	SPoC	No	38.0	54.6	11.4	20.9	59.8	93.0	32.4	69.7
	CroW	No	41.4	58.8	13.9	25.7	62.9	94.4	36.9	77.9
	NetVLAD	Yes	37.1	56.5	13.8	23.3	59.8	94.0	35.0	73.7
	MAC	Yes	58.4	81.1	30.5	48.0	66.8	97.7	42.0	82.9
	GeM	Yes	61.9	82.7	33.7	51.0	69.3	97.9	44.3	83.7
	*GeM+CA	Yes	62.9	84.1	35.5	54.0	70.8	98.3	46.0	85.0
Res101	SPoC	No	39.8	61.0	12.4	23.8	69.2	96.7	44.7	78.0
	CroW	No	42.4	61.9	13.3	27.7	70.4	97.1	47.2	83.6
	R-MAC	Yes	60.9	78.1	32.4	50.0	78.9	96.9	59.4	86.1
	GeM	Yes	64.7	84.7	38.5	53.0	77.2	98.1	56.3	89.1
	*GeM+CA	Yes	67.3	87.1	42.6	59.1	**77.5**	98.6	**56.5**	88.6
Res50	DELF–ASMK+SP	Yes	**67.8**	**87.9**	**43.1**	**62.4**	76.9	**99.3**	55.4	**93.4**
Query Expansion (QE)										
VGG16	GeM+αQE	Yes	66.6	85.7	38.9	57.3	74.0	98.4	51.0	88.4
	*GeM+CA+αQE	Yes	68.0	83.5	40.6	55.2	76.7	98.7	54.7	90.4
Res101	R-MAC+αQE	Yes	64.8	78.5	36.8	53.3	82.7	97.3	65.7	90.1
	GeM+αQE	Yes	67.2	86.0	40.8	54.9	80.7	98.9	61.8	90.6
	*GeM+CA+αQE	Yes	68.8	84.9	43.9	59.7	83.6	**99.0**	66.1	91.6
Res50	DELF-HQE+SP	Yes	**73.4**	**88.2**	**50.3**	**67.2**	**84.0**	98.3	**69.3**	**93.7**

Table 3: Image retrieval performance (mAP) comparison on Roxford5k and Rparis6k datasets. All compared other works' mAP in (b) are from [24].

We randomly select 50 image pairs and evaluate the time required for the feature extraction by the proposed GeM+CA with attention map refinement, and that for the original GeM [25]. In average our method took 700ms longer than the original GeM. The extra time cost is due to the calculation of the attention map refinement step from Fig. 3. Without the refinement step, our method took 260ms longer than original GeM.

6 Conclusion

In this research study, we propose an independent conditional attention model which does not require any manual annotation. Instead, the model is trained on automatically generated training data by finding correspondences from existing matching image pairs. As shown in the experiments, our attention model can accurately highlight the region, matching the content of the query image, on the candidate image. It performs well even in various challenging situations such as when significantly changing the illumination conditions or the image acquisition parameters. When combined with the GeM feature extraction method, it achieves the state of the art image retrieval results on Oxford5k and Oxford105k datasets.

Acknowledgement

The authors would like to thank NVIDIA for granting a Titan XP GPU, which was used for the experiments.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016.

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *Proc. European Conf. Computer Vision (ECCV), vol. LNCS 8689*, pages 584–599, 2014.

[3] S. Berretti, A. Del Bimbo, and P. Pala. Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Trans. on Multimedia*, 2(4):225–239, 2000.

[4] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPR-w)*, pages 224–236, 2018.

[5] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. European Conf. on Computer Vision (ECCV), vol LNCS 9910*, pages 241–257, 2016.

[6] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-End learning of deep visual representations for image retrieval. *Int. Jour. Computer Vision*, 124(2):237–254, Sep 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu. One-shot object detection with co-attention and co-excitation. In *Proc Advances in Neural Information Processing Systems (NIPE)*, pages 2725–2734, 2019.

[9] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *Proc. of European Conf. on Computer Vision (ECCV), vol. LNCS 9913*, pages 685–701, 2016.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Proc. Systems (NIPS)*, pages 1097–1105, 2012.

[12] L. Li, M. Xu, H. Liu, Y. Li, X. Wang, L. Jiang, Z. Wang, X. Fan, and N. Wang. A Large-Scale database and a CNN model for Attention-Based glaucoma detection. *IEEE Trans. Medical Imaging*, 39(2):413–424, February 2020.

[13] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen. Deep learning of binary hash codes for fast image retrieval. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPR-w)*, pages 27–35, 2015.

[14] D. G Lowe. Distinctive image features from scale-invariant keypoints. *Int. Jour. of Computer Vision*, 60(2):91–110, 2004.

[15] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8):837–842, August 1996.

[16] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-i Nieto. Bags of local convolutional features for scalable instance search. In *Proc. of ACM on Int. Conf. on Multimedia Retrieval*, pages 327–331, 2016.

[17] E. Mohedano, K. McGuinness, X. Giró-i Nieto, and N. E. O'Connor. Saliency weighted convolutional features for instance search. In *Proc. Int. Conf. on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, September 2018.

[18] B. Munjal, S. Amin, F. Tombari, and F. Galasso. Query-guided end-to-end person search. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 811–820, 2019.

[19] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pages 3456–3465, 2017.

[20] A. Papushoy and A. G. Bors. Image retrieval based on query by saliency content. *Digital Signal Processing*, 36(1):156–173, January 2015.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[23] F. Radenović, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised Fine-Tuning with hard examples. In *Proc. European Conf. Computer Vision (ECCV), vol. LNCS 9905*, pages 3–20, 2016.

[24] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-Scale image retrieval benchmarking. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5706–5715, 2018.

[25] F Radenović, G Tolias, and O Chum. Fine-Tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1655–1668, July 2019.

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

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. G. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Inter. Jour. of Computer Vision*, 115(3):211–252, 2015.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1409.1556*, 2014.

[29] M. J. Swain and D. H. Ballard. Color indexing. *Int. Jour. of Computer Vision*, 7(1): 11–32, November 1991.

[30] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879*, 2015.

[31] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. Torr. Fast online object tracking and segmentation: A unifying approach. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338, 2019.

[32] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[33] X. Wu, G. Irie, K. Hiramatsu, and K. Kashino. Weighted generalized mean pooling for deep image retrieval. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pages 495–499, October 2018.

[34] A. B. Yandex and V. Lempitsky. Aggregating deep convolutional features for image retrieval. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1269–1277, 2015.

[35] W. Zhou, H. Li, and Q. Tian. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*, 2017.