ENABLING LARGE-SCALE IMAGE SEARCH WITH CO-ATTENTION MECHANISM

*Zechao Hu and Adrian G. Bors**

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

ABSTRACT

Content-based image retrieval (CBIR) consists of searching the most similar images to a given query. Most existing attention mechanisms for CBIR are query non-sensitive and are only based on single candidate image's feature regardless of the actual query content. This can result in incorrect regions especially when the target object is not salient or surrounded by distractors. This paper proposes an efficient and effective query sensitive co-attention mechanism for large scale CBIR tasks. Local feature selection and clustering are employed to reduce the computation cost caused by the query sensitivity. Experimental results indicate that the proposed co-attention method can generate good co-attention maps even under challenging situations leading to a new state of the art performance on several benchmark datasets.

***Index Terms*—** Image retrieval, Co-attention mechanisms, Feature clustering.

1. INTRODUCTION

Deep Convolution Neural Network (CNN) based methods for CBIR can be divided into two categories: global and local feature methods. Global feature methods extract a compact feature vector from each image using a single forward passing through the network. It can be achieved by a fully connected layer [1] or by global spatial pooling [2, 3, 4]. In addition, several attention mechanisms have been proposed for feature refinement before global pooling. The Weighted Generalized Mean pooling (WGeM) [5] employs a trainable spatial weighting module for feature re-weighting. SOLAR [6] explores the correlation between each entry from the convolution feature tensor with the second order attention. Deep Orthogonal Local and Global (DOLG) [7] proposes an Orthogonal Fusion module to combine the global feature with critical local features for better image representation, while a dot-product fusion module is trained in [8]. Local feature methods treat each entry from the feature tensor as a local descriptor followed by a separate aggregation method to build the final image representation [9, 10, 11]. In recent works, selected local features are further used in spatial verification mechanisms for re-ranking [12, 13]. For example HOW [14] combines CNN-based local features with the Aggregated Selective Match Kernel (ASMK) [15] to directly perform many-to-many local feature matching for image retrieval.

Despite the successes of CNN-based methods, existing attention mechanisms for CBIR [5, 12, 13, 14] are all query non-sensitive; for the given candidate images they predict the regions 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 fail when the target object is not salient or surrounded by distractors. For example in Fig. 1, the query-nonsensitive attention mechanism from the WGeM [5] fails. As the Louvre Pyramid and Palace are both potential objects of interest, when treating the Louvre Pyramid as the query item, it is always ignored by the WGeM attention module while the adjacent Louvre Palace attracts more attention.

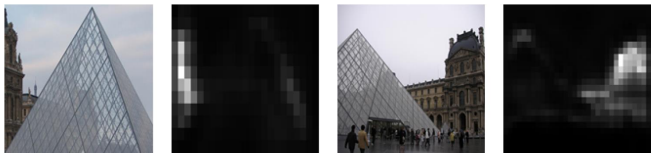

Fig. 1. Examples of query non-sensitive attention where WGeM approach fails. Images taken from [5].

Ideally, the attention should be query sensitive, consistent with the current query content. When the Louvre Pyramid is treated as query, it should be highlighted in the resulting co-attention map and vice versa, as shown in the examples 3-4 from Fig. 3. This kind of query sensitive attention, conditioned on the query content, is called co-attention in this paper. In some other co-attention works [16, 17, 18] the query pattern was shown to be essential for feature extraction.

Our contributions are : 1) we propose an efficient co-attention method based on local feature selection and clustering without the requirement of extra parameter training; 2) we show that our method could generate good co-attention maps even for some hard situations; 3) the retrieval performance is greatly improved with our co-attention method according to the experimental results and reaches new state of art performance on several benchmark datasets.

2. BASELINE MODEL STRUCTURE

The proposed co-attention method serves as a post-processing module for pre-trained spatial pooling models without requir-

* Dr. A. G. Bors acknowledges the partial support from the EPSRC, UK, project COUSIN (EP/V009591/1)

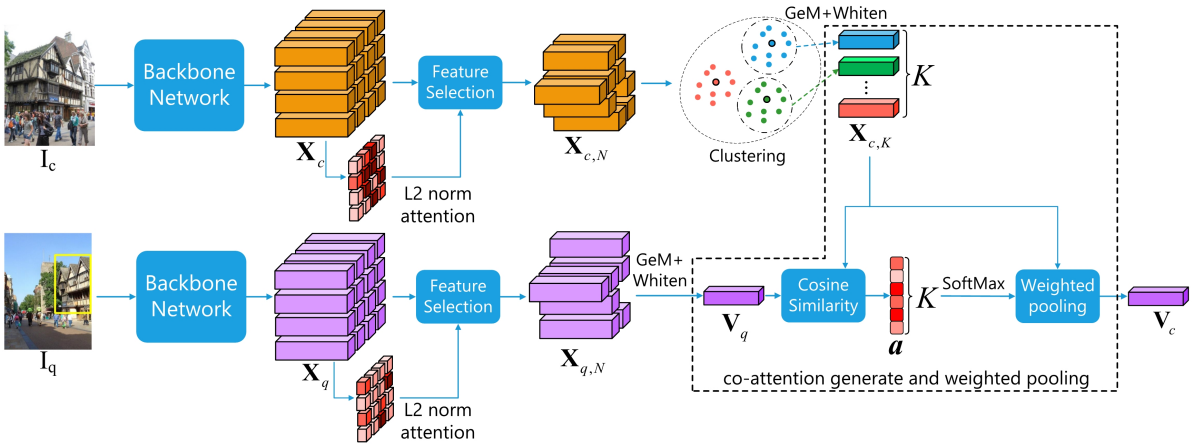

Fig. 2. Illustration of clustering based co-attention generation and weighted feature extraction.

ing the training of any extra parameters. Accordingly, in this paper, we follow the framework from [4] to construct the baseline GeM model. ResNet101 [19] is used as the backbone network for the feature tensor extraction. The output feature tensor is globally pooled by a GeM layer [4] followed by a fully connected layer for feature whitening. Let $\mathbf{X} = [\mathbf{x}_l] \in \mathbb{R}^{H \times W \times D}$ denote the feature tensor output by the backbone network before pooling, where H, W, D represent the height, width and the channel count ($D = 2048$ for ResNet101), $\mathbf{x}_l$ represents the local feature at location l from $\mathbf{X}$. According to the spatial pooling from [14], any loss function that optimizes the cosine similarity between global pooling features would implicitly optimize the following aspects : first, for irrelevant background locations $\mathbf{x}_{bg}$, the L2 norm is minimized, leading to little or no contribution to the final similarity score. On the contrary, for distinct foreground objects or region locations $\mathbf{x}_{bg}$, the L2 norm is maximized. Accordingly, the L2 norm can be treated as a spatial attention that the model implicitly learns at the training stage [14].

3. ENABLING CBIR WITH CO-ATTENTION

In the following we consider the convolution feature tensor output by the pre-trained GeM model from Section 2 for enabling the co-attention generation process.

Local feature selection and clustering. The first challenge for using co-attention is the computation cost required by the large number of local features that are extracted from a single image. Hundreds of local features could be extracted from a single high resolution image. However, it is impractical to cache all these features. An intuitive way to reduce the memory cost is to discard the irrelevant background local features. L2 norm of each entry on the CNN feature tensor can be used as an indicator of feature importance. Then, feature selection can be performed by keeping the top N local features with the highest L2 norm from feature tensor $\mathbf{X}$, resulting in a selected local feature set $\mathbf{X}_N \in \mathbb{R}^{N \times D}$. To further reduce the number of local features, k-means clustering is employed on $\mathbf{X}_N$, grouping them into K clusters. Within each cluster, after performing GeM pooling in order to select the representative local features centers, followed by whitening with the fully connected layer, results in the clustered local features $\mathbf{X}_K \in \mathbb{R}^{K \times D}$, $K << N$.

Co-attention generation with local features. The pipeline of co-attention generation and weighted feature extraction is illustrated in Fig. 2. After extracting the representative features by feeding the query image $\mathbf{I}_q$ and the candidate image $\mathbf{I}_c$ through the backbone network, we consider L2 norm for the feature selection. The selected query local features $\mathbf{X}_{q,N}$ are then directly GeM pooled and whitened to obtain the query global feature $\mathbf{V}_q$. In order to extract representative feature vectors, selected candidate local features $\mathbf{X}_{c,N}$ are clustered and then whitened, resulting in the local feature set $\mathbf{X}_{c,K}$. Then, the co-attention weights $\mathbf{a} = [a_i] \in \mathbb{R}^K$ are obtained by calculating the cosine similarity between $\mathbf{V}_q$ and each local feature from $\mathbf{X}_{c,K}$. However, the range of the attention weights is within $[-1, 1]$, which may not ensure a high contrast among the locations. For better controlling the weight distribution and normalizing them into the range $[0, 1]$, the SoftMax function is then applied on $\mathbf{a}$:

$$a'_i = \frac{\exp(a_i T)}{\sum_j^K \exp(a_j T)}, \tag{1}$$

where T is a temperature parameter. The final co-attention weighted candidate global feature vector $\mathbf{V}_c$ is defined by weighted sum pooling :

$$\mathbf{V}_c = \frac{1}{K} \sum_i^K a_i \mathbf{X}_{c,i}. \tag{2}$$

The similarity measure is performed by evaluating the cosine similarity between $\mathbf{V}_q$ and $\mathbf{V}_c$.

Further computation cost reducing. In order to make the co-attention practical to large-scale image retrieval and for further reducing the required computation costs we propose two extra processing steps during the retrieval stage. First, PCA dimension reduction is performed on both query global feature $\mathbf{V}_q$ and the candidate image local features from $\mathbf{X}_{c,K}$. Second, an inverted file indexing [23] module is applied to reduce the candidate image count that need to be compared with

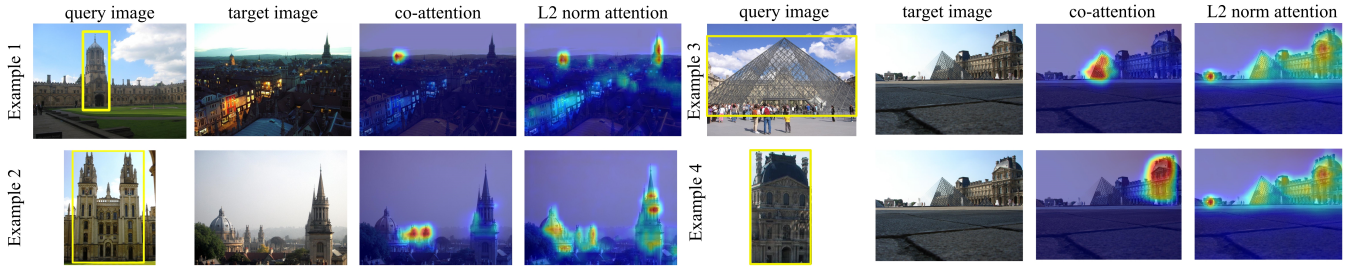

Fig. 3. Attention map visualizations on target images.

Method	*Medium (%)*				*Hard (%)*			
	ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1M
(A) Local feature								
DELF-D2R-R-ASMK*+SP [20]	76.0	64.0	80.2	59.7	52.4	38.1	58.6	29.4
R50^{-}-HOW [14]	79.4	65.8	81.6	61.8	56.9	38.9	62.4	33.7
R101^{-}-HOW (GLDv2)[14] †[2]	83.9	77.9	87.9	76.4	71.3	52.8	76.0	56.4
(B) Global feature								
R101-R-MAC [21]	60.9	39.3	78.9	54.8	32.4	12.5	59.4	28.0
R101-GeM (GLD) [6]	67.3	49.5	80.6	57.3	44.3	25.7	61.5	29.8
R101-DSM [22]	65.3	47.6	77.4	52.8	39.2	23.2	56.2	25.0
R101-SOLAR [6]	69.9	53.5	81.6	59.2	47.9	29.9	64.5	33.4
R50-DELG + SP [13]	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7
R101-DELG + SP [13]	81.2	69.1	87.2	71.5	64.0	47.5	72.8	48.7
R101-DELG + SP [13]†	84.1	75.9	91.0	79.2	68.8	53.6	83.0	62.3
R50-DOLG [7][1]	81.2	71.4	90.1	79.0	62.6	47.3	79.2	59.8
R101-DOLG [7][1]	82.3	73.6	90.9	80.4	64.9	51.6	81.7	62.9
(C) the proposed co-attention method								
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**

Table 1. Image retrieval results on ROxf/RPar datasets and when adding the 1 million distractor set R1M, for the *Medium* and *Hard* evaluation protocols. "†" indicates re-implemented model under the training details from Section 4.

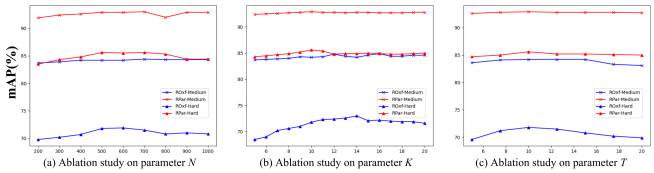

Fig. 4. Ablation results for the hyper-parameters.

the query image at the retrieval stage. At the feature extraction stage, both selected query image 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 [23] from the codebook while recording the visual word indices that each image is assigned to. Then during the retrieval stage, for each query image we only pick out those candidate database images that share at least one visual word with the query image to perform co-attention generation and assess their similarity. The other images are no longer considered.

4. EXPERIMENTS

Experiment setup. For a fair comparison with the current state-of-art (SOTA) work Deep Orthogonal Local and Global (DOLG) [7], the baseline model described in Section 2 is trained on GLDv2 dataset [24] with ArcFace margin loss [13]. The batch-size is set to 128. The initial learning rate is considered as 0.05 together with a cosine learning rate decay strategy [7]. The model is trained for no more than 50 epochs. 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. (1). The feature dimension is reduced to 512 by PCA. For the inverted file index, we use single scale 60,000 images from the training dataset (GLDv2) to train the codebook. From each image, 300 local features are picked out and compressed to a dimension of 128 by the PCA. The visual word count of the codebook is set to 65,536. In addition, the multi-scale feature extraction scheme [4] is applied, where we

consider 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. We consider ROxf/RPar datasets [25] along with a 1 million images distractor set R1M [25] for the performance evaluation.

Visualization results. Visualization examples of the proposed co-attention are shown in Fig. 3. For comparison, the query non-sensitive L2 norm attention is shown in the forth column. As we can observe, the L2 norm attention tends to highlight regions relevant to the training data, while our co-attention can accurately highlight regions that match the query content.

Image retrieval results[1]. 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 works are re-implemented and marked with "†". The group (A) of results from Table 1 shows the results of local feature methods. R101$^-$-HOW (GLDv2)†[2] is the reimplementation of HOW [14] on GLDv2 dataset with ResNet101 backbone and ArcFace loss. It reaches 71.3% mAP on ROxf *hard* set before but it has relatively weak performance with the 1 million distractors. Group (B) shows the results of the global feature methods. The original DEep Local and Global features (DELG) model [13] was trained on GLDv2 with a small batch size of 32. R101-DELG† is its re-implemented version with ResNet101 as the backbone network. 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 for the baseline model (GeM†) as described in Section 2 and when it is combined with the proposed co-attention method (GeM†-CA). For the results of GeM† and GeM†+CA, they share the same exact GeM model with that described in Section 2, the only difference is that GeM†+CA implements the co-attention method as in Section 3 as well as PCA dimension reduction and inverted file indexing from Section 3) to re-weight the candidate image feature tensor before global GeM pooling. We observe that by introducing the co-attention to the CBIR pipeline greatly improves the retrieval performance. Especially, on the *hard* set of ROxf (RPar), GeM†+CA reaches the best results of 72.6% (85.6%). Also the proposed co-attention method still gives the best retrieval results when considering the 1 million distractor set.

5. ABLATION EXPERIMENT AND DISCUSSION

Impact of clustering parameters. We evaluate in the plots from Figures 4 (a), (b) and (c) the impact of cluster hyper-parameters features N, clusters k, and temperature T from Eq. (1), on the model retrieval performance. The proposed method is robust to changes in these hyper-parameters. Varying the number of clusters k, has implications not only on the performance but also on the computation cost. The setting described in the beginning of Section 4 reaches a good balance between performance and computation costs.

Computation cost and speed. For the memory cost, it takes around 21GB to cache the whole ROxf/RPar database with the 1 million distractor set. For the time cost, the feature extraction takes in average 240ms to cache one candidate image's local features but it can be performed offline and it is only done once. It takes on average 530ms with acceleration on a NVIDIA Tesla GPU, when searching on ROxf/RPar with the 1 million distractor dataset for one query image. Detailed computation cost comparision is provided in Table 2. The proposed method "GeM†+CA" requires a similar memory cost as DELG [13]. Although the proposed co-attention method requires more time cost than those simple global feature methods, like GeM [4] and DOLG [7], it provides the best retrieval performance.

Method	Device	Memory (GB) ROxf/RPar+1M	Retrieval time (ms) in average
HOW [14]	CPU	14	750
GeM [4]	Tesla GPU	8	250
DOLG [7]	Tesla GPU	2	220
DELG+SP [13]	Tesla GPU	22	383
GeM†+CA (ours)	Tesla GPU	21	530

Table 2. Computation cost comparison.

6. CONCLUSION

In this paper, we enable large-scale content-based image retrieval with co-attention mechanisms. 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 selection and clustering while also considering the inverted file indexing to speed up the retrieval procedure. While straightforward, the proposed co-attention method generates good co-attention maps even in some challenging cases. By simply adding our co-attention method to 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 white requirying comparable computation costs to other models.

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

[2] R101$^-$ represents the ResNet101 without the final convolution block. According to the study from [14], HOW gives better results when discarding the final block and we follow this setting for our reimplementation.

7. REFERENCES

[1] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky, "Neural codes for image retrieval," in *Proc. of the European Conference on Computer Vision (ECCV), vol. LNCS 8689*, 2014, pp. 584–599.

[2] Artem Babenko and Victor Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1269–1277.

[3] Giorgos Tolias, Ronan Sicre, and Hervé Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *Proc. of the International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1511.05879*, 2016.

[4] Filip Radenović, Giorgos Tolias, and Ondřej Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 7, pp. 1655–1668, 2018.

[5] Xiaomeng Wu, Go Irie, Kaoru Hiramatsu, and Kunio Kashino, "Weighted generalized mean pooling for deep image retrieval," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 495–499.

[6] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk, "SOLAR: second-order loss and attention for image retrieval," in *Proc. of the European Conference on Computer Vision (ECCV), vol. LNCS 12370)*, 2020, pp. 253–270.

[7] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang, "DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 11772–11781.

[8] Zechao Hu and Adrian G. Bors, "Dot-product based global and local feature fusion for image search," in *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2022, pp. 1911–1915.

[9] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. of the IEEE CVPR-workshops*, 2015, pp. 685–701.

[10] Eva Mohedano, Kevin McGuinness, Noel E O'Connor, Amaia Salvador, Ferran Marques, and Xavier Giró-i Nieto, "Bags of local convolutional features for scalable instance search," in *Proc. of the ACM International Conference on Multimedia Retrieval (ICMR)*, 2016, pp. 327–331.

[11] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[12] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han, "Large-scale image retrieval with attentive deep local features," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3456–3465.

[13] Bingyi Cao, André Araujo, and Jack Sim, "Unifying deep local and global features for image search," in *Proc. of the European Conference on Computer Vision (ECCV), vol. LNCS 12365*, 2020, pp. 726–743.

[14] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum, "Learning and aggregating deep local descriptors for instance-level recognition," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2020, pp. 460–477.

[15] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou, "Image search with selective match kernels: aggregation across single and multiple images," *International Journal of Computer Vision (IJCV)*, vol. 116, no. 3, pp. 247–261, 2016.

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

[17] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 811–820.

[18] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1328–1338.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5109–5118.

[21] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2016, pp. 241–257.

[22] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum, "Local features and visual words emerge in activations," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11651–11660.

[23] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2003, vol. 3, pp. 1470–1470.

[24] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2575–2584.

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