Utilizing Query-Adaptive Attention Mechanisms in Deep Learning for Content-Based Image Retrieval

Kunal[1], Shiwali Yadav[1], Hardik[1], Jatin Bansal[2], Chanchal Rai[1], Aaryan Aneja[1]
[1]Chandigarh University, Mohali India.

Kunalsingla009@gmail.com , shiwali.e13277@cumail.in , chughhardik2@gmail.com , jatinbansal0709@gmail.com , 21bcs5599@cuchd.in

***Abstract*—**This research paper delves into the domain of Content-based Image Retrieval (CBIR) by introducing innovative query-sensitive attention mechanisms to augment the process of feature extraction for image retrieval. Many existing CBIR methods overlook the specifics of the query pattern, leading to a focus on irrelevant regions within the image database. To address this challenge, this paper presents a series of novel contributions. Firstly, it introduces the Conditional Attention Network (CANet), which takes both the query image and a candidate image as input, generating a co-attention map for the candidate image based on the query content. This co-attention map effectively highlights the target object and enhances image retrieval performance when embedded within a convolutional neural network (CNN)-based feature extraction pipeline. Additionally, a more efficient co-attention method is proposed, leveraging local feature selection and clustering techniques to significantly reduce computational costs while maintaining accurate co-attention maps. This clustering-based co-attention method achieves state-of-the-art performance on various benchmark datasets. Lastly, the paper explores the use of clustered expressive local features for many-to-many local feature matching in CBIR. This method implicitly generates local matching maps akin to co-attention and incorporates a trainable binary encoding layer for network fine-tuning. This allows the model to generate compact binary codes with minimal performance degradation and substantially reduces computation costs. In summary, this research underscores the crucial role of query information in feature extraction for CBIR and demonstrates the practicality and effectiveness of co-attention mechanisms, even in the context of large-scale image retrieval tasks.

Keywords*—*Content-based Image Retrieval (CBIR), Query-sensitive Attention Mechanisms, Conditional Attention Network (CANet), Co-attention, Feature Extraction, Local Feature Selection, Clustering, Image Retrieval Performance, Convolutional Neural Network (CNN), Many-to-Many Local Feature Matching, Binary Encoding, Benchmark Datasets.

DOI: 10.48047/ecb/2023.12.1025

I. INTRODUCTION

The advent of modern technology has brought about a proliferation of photo-capture devices, including cameras and smartphones, which are now widely used in various aspects of life, generating a vast amount of image data across domains such as social media, healthcare, industry, and education. With the rise of the Internet, images are being created, stored, and shared globally on a daily basis. This growing need for image organization has led to the development of image retrieval systems aimed at efficiently locating and retrieving images that match a given query based on their content. The roots of image retrieval can be traced back to the 1970s when it gained traction after the Database Techniques for Pictorial Applications conference. Image retrieval systems can be categorized based on query formats, with text-based image retrieval (TBIR) being an early approach that uses textual annotations like keywords and descriptions to search for images. While TBIR is computationally efficient, it has limitations in terms of scalability, subjectivity, and language constraints. To address these limitations, content-based image retrieval (CBIR), which relies on image content rather than text, was introduced, allowing users to query with images directly. Other specialized query formats like sketch-based retrieval and color layout retrieval have also emerged. In this paper, we focus on generic CBIR, which involves retrieving images that share the same content as the query image.

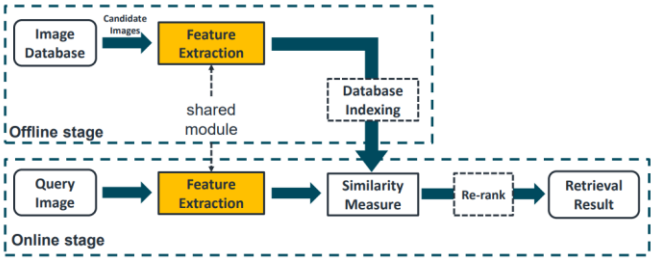


Figure 1. Illustration of content-based image retrieval pipeline.

The CBIR process comprises two main stages: the offline stage and the online stage. In the offline stage, features are extracted and cached for candidate images from the image database. These features serve as compact representations of the images and are computed only once. The online stage involves processing the query image, extracting its features, and measuring the similarity between the query and each database image's features. To enhance CBIR performance, additional modules such as database indexing and re-ranking can be employed. Database indexing optimizes the organization of the image database to expedite retrieval, while the re-ranking module refines initial retrieval results. The core components of a CBIR pipeline are feature extraction and similarity measurement. Deep learning has revolutionized CBIR by

enabling the automatic extraction of complex features from images. Convolutional neural networks (CNNs) are widely used for feature extraction due to their ability to learn rich representations. Attention mechanisms have been incorporated into CNN-based CBIR systems to refine feature output, but most of these mechanisms are query non-sensitive, which can lead to suboptimal results, especially when the query content is not salient. To address this issue, this paper introduces the concept of query-sensitive co-attention mechanisms, which dynamically adapt to the query content, improving feature selection or re-weighting for CBIR. Co-attention, while beneficial, can introduce additional computational costs, making it impractical for large-scale image retrieval. This research focuses on embedding effective and efficient co-attention mechanisms into the feature extraction procedure within the CNN-based image retrieval pipeline to enhance performance.

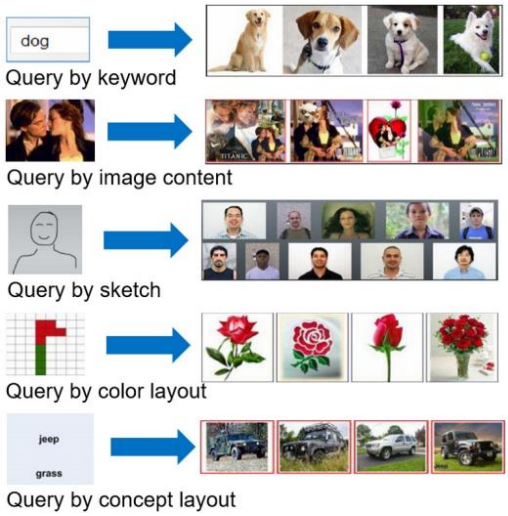


Figure 2. Illustration of different query formats and corresponding results.

II. Related Work

In this section, we provide an overview of several typical conventional CBIR approaches. To begin with, we discuss early techniques that involve manually crafting feature extractors, relying on various low-level feature characteristics. The primary objective of these methods is to convert raw pixel images into concise feature representations, optimizing the efficiency of image retrieval. Following this, we delve into more intricate methods for feature aggregation. These aggregation techniques can be considered as distinct post-processing modules, taking locally extracted features as their input and constructing more significant and streamlined representations tailored for CBIR applications.

A. Low-level Features

The early-stage conventional CBIR methods predominantly rely on manually engineered feature extractors that utilize low-level feature information, including color, texture, shape, and gradient. We will now delve into representative works for each of these feature categories.

Color: Color, being a fundamental aspect of image content, is invariant to various image acquisition parameters such as scale, rotation, and camera viewpoint changes. Early works employed color histograms, which represent the color distribution of an image in terms of probability density functions. These histograms provide global image descriptions, but they may not capture complex content. Techniques like Color Coherence Vector (CCV) and Color Correlograms introduced spatial information to improve color histogram representations.

Texture: Texture features describe image characteristics related to color, shape, structure, randomness, and more. Gabor wavelet features were among the pioneering methods for using texture in image retrieval, employing a set of Gabor filters with various orientations to capture patterns in the image. Other methods, such as Edge Histogram Descriptor (EHD) and Discrete Wavelet Transform, have been developed to extract texture-based features. However, texture features can be sensitive to noise and computationally demanding.

Shape: Shape information, conveying strong semantic meaning, is crucial for recognizing objects based on their contours. Shape-based features are typically used in conjunction with color and texture features to enhance image retrieval systems. Pseudo-Zernike moments are employed in to describe shape information. In this work, a CBIR pipeline combines color, texture, and shape features to calculate similarity scores.

Gradient: Gradient information, encompassing the magnitude and orientation of features, is used for local feature extraction. Scale-Invariant Feature Transform (SIFT) is a popular method that identifies interest points and extracts local descriptors based on gradient information. SIFT is robust to various transformations and has been applied in image retrieval. However, SIFT can be computationally intensive. Edge-SIFT simplifies SIFT by generating binary-coded features, making it suitable for large-scale retrieval. Speeded Up Robust Features (SURF) provides a more efficient alternative to SIFT by representing image content using Haar wavelet responses and offering faster feature extraction with lower-dimensional feature vectors.

FigurF

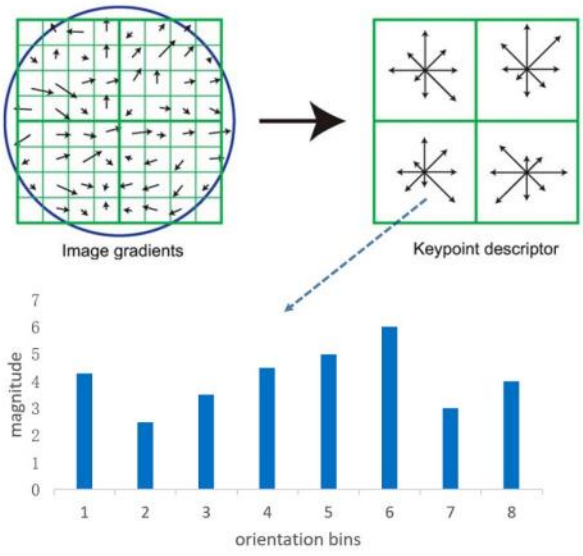


Figure 3. Illustration of building gradient orientation histogram from region blocks

These methods form the foundation of conventional CBIR approaches, each with its strengths and limitations in capturing image content based on various low-level feature characteristics.

B. Feature Aggregation

In addition to the previously discussed methods that primarily focus on extracting and encoding low-level feature information from raw input images into concise feature vectors, there exists another category of approaches dedicated to feature aggregation. These methods aim to leverage the extracted feature vectors more effectively, thereby creating comprehensive image representations and exploring suitable similarity measures for retrieval. Typically, these techniques work alongside a predefined local feature extractor and serve as a post-processing module for generating improved and compact image representations from the pre-extracted local features.

One prominent example of feature aggregation is the "Bag of Visual Words" (BoV) method, which draws inspiration from the Bag of Words (BoW) concept used in text processing. In BoV, local descriptors are computed from sample images during the training stage, and k-means clustering is applied to these descriptors to generate "visual words," forming a codebook. During the retrieval stage, local descriptors from each image are clustered based on these visual words, and frequency histograms for each visual word are created to represent the image. The global feature vector for each image is constructed by incorporating the frequency of each visual word. The similarity between the query image and each candidate image in the database is calculated using cosine similarity, yielding the retrieval results. While the BoV idea essentially extends the text-processing concept to image retrieval, it has influenced numerous other methods that focus on local feature aggregation.

Another notable approach is the "Vector of Locally Aggregated Descriptors" (VLAD), which differs from BoV by accumulating and concatenating residuals between each image's local descriptor and the corresponding visual word to construct the final compact global descriptor. Additionally, the "Fisher Vector" method is successful in aggregating features into a fixed-size vector by computing the gradient of the log-likelihood function with respect to a set of parameter vectors. In this approach, a Gaussian Mixture Model (GMM) is employed to aggregate normalized gradient vectors of all local descriptors into a uniform Fisher Vector, using an average pooling scheme. It can be considered a generalized representation of BoV or a probabilistic version of VLAD.

Furthermore, the "Aggregated Selective Match Kernel" (ASMK) encompasses many-to-many matching techniques with pre-extracted local feature vectors. Like VLAD, ASMK calculates residual vectors between local feature vectors and the corresponding visual word. However, instead of concatenating these residuals into a compact global feature, ASMK aggregates the residual vectors associated with the same visual word through summation, resulting in a set of aggregated local feature vectors as the final representation of the original image. A matching kernel is then applied to these local features to perform many-to-many similarity evaluations between images, yielding retrieval results. These methods collectively offer efficient ways to represent and retrieve images by leveraging aggregated features effectively.

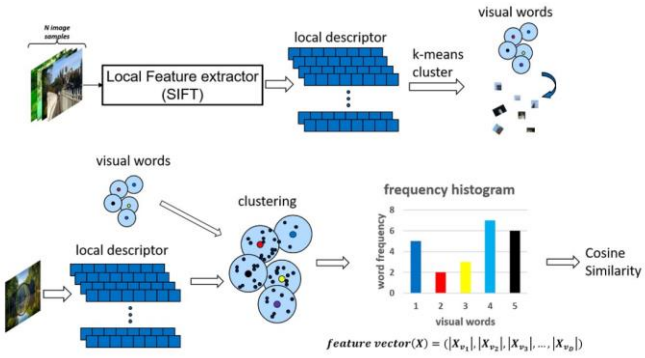


Figure 4. Illustration of bag of visual word pipeline

Benchmark Datasets: Several commonly used benchmark datasets are employed to evaluate the performance of Content-Based Image Retrieval (CBIR) models. These datasets include:

1. INSTRE: This dataset is for instance-level retrieval and comprises two subsets, INSTRE-S and INSTRE-M. INSTRE-S contains 23,070 images distributed across 200 categories, each containing a single object of interest. INSTRE-M consists of 5,473 images, each featuring two instances from 100 object categories.

2. University of Kentucky Benchmark Dataset (UKB): UKB contains 10,200 images grouped into 2,550 categories. Each group contains four images of the same object under various acquisition conditions. Evaluation on UKB is based on the average number of same-object images within the top four results.

3. INRIA Holidays Dataset (Holiday): Comprising 1,491 images from personal holiday photo albums, this dataset includes images taken intentionally with different acquisition conditions, such as rotation, illumination changes, and various

viewpoints. The images are divided into 500 groups, with each group corresponding to a different scene or object.

4. Oxford Building Dataset (Oxford5k): Oxford5k contains 5,062 images from Flickr, annotated with 17 tags related to specific landmarks in Oxford, such as Balliol Oxford, Christ Church Oxford, and others. It includes retrieval ground-truth images in 11 categories, with five query images per category, resulting in a total of 55 query images for evaluation.

5. Pairs Building Dataset (Paris6k): Paris6k dataset comprises 6,412 images collected from Flickr, focusing on 12 particular Paris landmarks, including the Eiffel Tower Paris and Louvre Paris. It also provides 55 queries for evaluating image retrieval models.

To make evaluation more challenging, Oxford5k and Paris6k can be expanded to Oxford105k and Paris106k by adding an additional set of 100,000 distractor images collected from Flickr.

Revisited Oxford (ROxf) and Paris (RPar) Datasets: ROxf and RPar are extended versions of the Oxford and Paris datasets, featuring corrected annotations and additional query images. ROxf contains 4,993 images, while RPar includes 6,322 images. Both datasets consist of 70 query images, categorized as Easy, Medium, or Hard based on the difficulty of assessing the similarity of their image representations with the corresponding query. Additionally, R1M is a new distractor set containing 1 million unbiased high-resolution images for ROxf and RPar. These datasets provide a comprehensive range of challenges for assessing CBIR model performance.

Dataset	Method	Result
INSTRE	BLCF-SalGAN	69.8
UKBench	R-MAC	3.90
Holiday	R-MAC	94.0
Oxford-5k	WGeM	88.8
Oxford-105k	WGeM	85.6
Paris-6k	R-MAC	93.6
Paris-106k	DELF	81.7
ROxf-5k	DOLG	64.9
ROxf-5k+1M	DOLG	51.6
RPar-6k	DOLG	81.7
RPar-6k+1M	DOLG	62.9

Table 1. Quantitative retrieval results on common benchmark datasets.

III. Conditional attention network Architecture

The CANet architecture effectively leverages conditional attention to focus on specific regions of interest within candidate images, making it particularly suitable for content-based image retrieval and object recognition tasks. By combining visual encoding, feature fusion, and co-attention map generation, CANet enhances the ability to assess the relevance of candidate images to a given query. This enables the model to provide more accurate and context-aware results, making it a valuable tool for applications where precise localization and retrieval of objects or regions are crucial, such as image-based search and recognition systems. The multi-scale convolution blocks within the fusion module enhance its flexibility in handling variations in object sizes and image conditions. Overall, CANet's conditional attention mechanism and feature fusion make it a promising architecture for advancing the state of the art in computer vision tasks that involve matching and localization of objects in images.

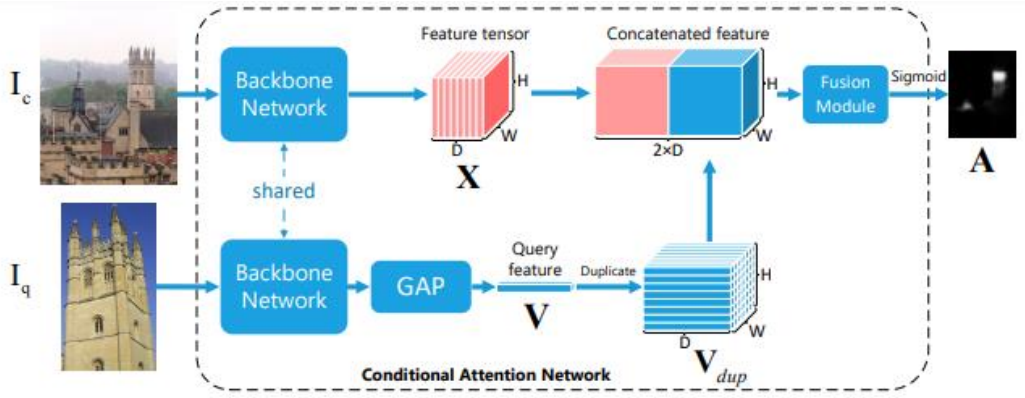
The Conditional Attention Network (CANet) architecture, aimed at defining Regions Of Interest (ROI) in candidate images based on the content in a query image, is described as follows:

Network Architecture: CANet comprises three main stages: visual encoding, feature fusion, and attention map generation. A convolutional neural network serves as the backbone network to encode features from both the query and candidate images. The query image is globally pooled to obtain a query global feature vector, which will be compared with features from the candidate image.

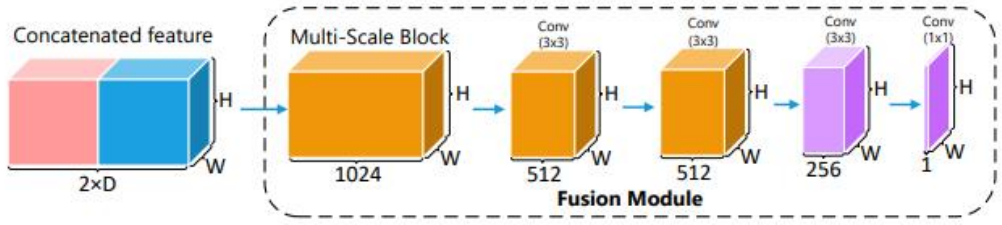
Feature Fusion: The attention model in CANet fuses the feature tensor of the candidate image with the global query feature vector. The feature tensor and query feature vector are L2 normalized and then concatenated. They pass through a fusion module, which includes multi-scale convolution blocks, to generate a co-attention map. The fusion module ensures that the model can evaluate the consistency between the candidate image's local features and the global query feature across different locations in a trainable way. The multi-scale convolution blocks allow the model to consider different context information at various locations.

Co-attention Generation: After the fusion step, a Sigmoid activation function is used to normalize values at each location within the range of 0 to 1, producing a one-channel co-attention map for the candidate image. This co-attention map represents the likelihood of a match between each location in the candidate image and the query image.

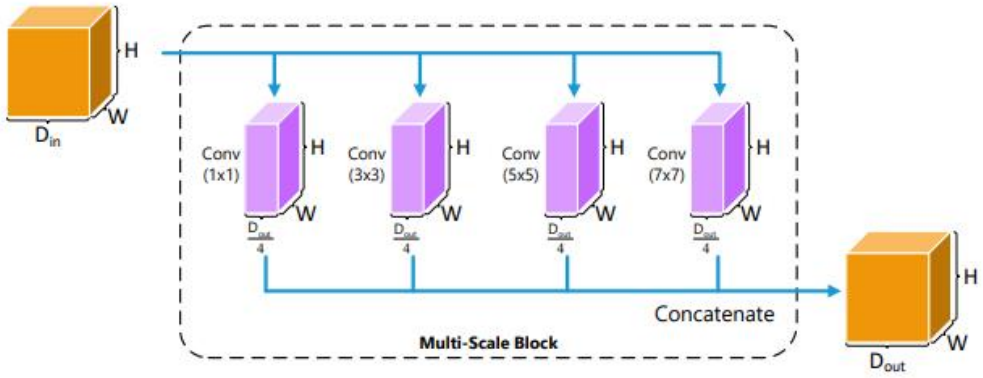
In summary, CANet is a conditional attention network designed to determine Regions Of Interest in candidate images based on the content in a query image. It achieves this through a series of stages including visual feature encoding, feature fusion with multi-scale convolution blocks, and co-attention map generation to assess the likelihood of matches between the candidate and query images.



(a) Illustration of the Conditional Attention Network global structure.



(b) Illustration of the feature fusion module.



(c) Illustration of the multi-scale convolution block.

Figure 5. The architecture of the proposed Conditional Attention Network

IV. Results

Quantitative Results: We present the image retrieval outcomes obtained through our local match approach, along with results from existing methods for reference. We re-implemented recent state-of-the-art techniques, which are indicated by "†" in the table.

Table 2 is organized into three groups for clarity. Group (A) displays the outcomes of local feature methods, while Group (B) shows the results of global feature methods. The lowermost group, Group (C), showcases the results of our proposed local match method, which incorporates PCA dimension reduction and Bi-half fine-tuning (LM-BiHalf). This method effectively serves as a post-processing module applied to the pre-trained baseline GeM model.

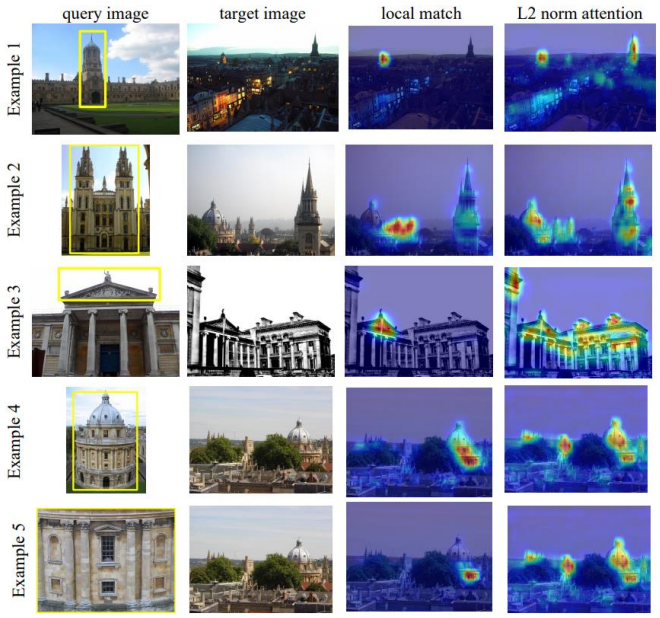


Figure 6. Visualization of the proposed local match and comparison with L2 norm attention.

Method	*Medium (%)*				*Hard (%)*			
	ROxf	ROxf+1M	RPar	RPar+1M	ROxf	ROxf+1M	RPar	RPar+1
(A) Local feature								
HesAff-rSIFT-ASMK*+SP [121]	60.6	46.8	61.4	42.3	36.7	26.9	35.0	16.8
HardNet-ASMK*+SP [79]	65.6	-	65.2	-	41.1	-	38.5	-
DELF-ASMK*+SP [120]	67.8	53.8	76.9	57.3	43.1	31.2	55.4	26.4
DELF-D2R-R-ASMK*+SP [120]	76.0	64.0	80.2	59.7	52.4	38.1	58.6	29.4
R50^{-}-HOW-MDA [137]	82.0	68.7	83.3	64.7	62.2	45.3	66.2	38.9
R50^{-}-HOW [123]	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 [39]	60.9	39.3	78.9	54.8	32.4	12.5	59.4	28.0
AlexNet-GeM [98]	43.3	24.2	58.0	29.9	17.1	9.4	29.7	8.4
VGG16-GeM [98]	61.9	42.6	69.3	45.4	33.7	19.0	44.3	19.1
R101-GeM [98]	64.7	45.2	77.2	52.3	38.5	19.9	56.3	24.7
R101-GeM-AP [101]	67.5	47.5	80.1	52.5	42.8	23.2	60.5	25.1
R101-GeM† [110]	67.3	49.5	80.6	57.3	44.3	25.7	61.5	29.8
R101-GeM (GLD) [85]	67.3	49.5	80.6	57.3	44.3	25.7	61.5	29.8
R101-DSM [110]	65.3	47.6	77.4	52.8	39.2	23.2	56.2	25.0
R101-SOLAR [85]	69.9	53.5	81.6	59.2	47.9	29.9	64.5	33.4
R50-DELG [14]	73.6	60.6	85.7	68.6	51.0	32.7	71.5	44.4
R50-DELG + SP [14]	78.3	67.2	85.7	69.6	57.9	43.6	71.0	45.7
R101-DELG [14]	76.3	63.7	86.6	70.6	55.6	37.5	72.4	46.9
R101-DELG + SP [14]	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 [146][8]	81.2	71.4	90.1	79.0	62.6	47.3	79.2	59.8
R101-DOLG [146][8]	82.3	73.6	90.9	**80.4**	64.9	51.6	81.7	**62.9**
(C) Our method								
R50-GeM†	79.8	69.0	87.3	73.1	60.4	44.2	74.0	52.0
R50-GeM†-LM-BiHalf	84.4	72.4	91.0	74.8	67.9	50.7	81.6	53.9
R101-GeM†	83.0	72.8	90.2	77.6	65.5	49.8	80.7	59.1
R101-GeM†-LM-BiHalf	**86.7**	**76.6**	**92.0**	79.3	**72.0**	**54.8**	**83.6**	61.4

Table 2. : Image retrieval results on ROxf/RPar datasets (and their extended version +1M distractor set R1M), considering Medium and Hard evaluation protocols.

Our observations reveal a significant enhancement in the retrieval performance of the baseline model when employing the local match method. Notably, when employing ResNet101 as the backbone network and evaluating on the Hard set of ROxf (RPar), the mAP of the local match method reaches an impressive 72.0% (83.6%). Furthermore, even when considering a dataset with 1 million distractors, the local match method consistently outperforms current state-of-the-art approaches such as DELG and DOLG on the ROxf+1M dataset, while delivering comparable results on the RPar+1M dataset. These results underscore the effectiveness of our local match approach in improving retrieval performance, especially in challenging scenarios.

V. Conclusion

The integration of query-sensitive attention mechanisms into content-based image retrieval (CBIR). They proposed different approaches for co-attention generation and local feature matching to enhance retrieval performance. The Conditional Attention Network (CANet) was introduced as a trainable co-attention generation branch for each candidate image, but it incurred high computational costs. To address this, a non-trainable co-attention generation method based on local feature clustering was developed and proved to significantly improve retrieval accuracy without the computational burden. The third approach, expressive local feature matching, utilized feature clustering with binary encoding and Bi-half fine-tuning to efficiently extract characteristic features for image retrieval. This method offered comparable retrieval results to co-attention techniques but with substantially reduced computational demands and eliminated the need for GPU support during online retrieval. The experimental results showed that embedding co-attention mechanisms in the feature extraction process markedly improved retrieval accuracy, setting new state-of-the-art results on benchmark datasets (ROxf/RPar). The interaction between query global features and local features output by the convolutional neural network was identified as an effective approach for co-attention generation. The clustering of local features automatically grouped those belonging to the same object, creating expressive local feature representations. The local feature matching method, while not explicitly generating co-attention maps, achieved query-sensitive local match maps and offered comparable retrieval accuracy to co-attention techniques, but with significantly reduced computational costs during retrieval.

Overall, the research concluded that query-sensitive attention mechanisms can substantially enhance the performance of CBIR systems, with co-attention methods and local feature matching offering effective solutions to this end.

References

[1] R Achanta, S Hemami, F Estrada, and S Susstrunk. Frequency-tuned salient region detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1597–1604, 2009.

[2] Swati Agarwal, Anil Kumar Verma, and Preetvanti Singh. Content based image retrieval using discrete wavelet transform and edge histogram descriptor. In Proceedings of the International Conference on Information Systems and Computer Networks, pages 19–23, 2013.

[3] Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan. Semantic content-based image retrieval: A comprehensive study. Journal of Visual Communication and Image Representation, 32:20–54, 2015.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place

recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5297–5307, 2016.

[5] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), volume 8, pages 1027–1035, 2007.

[6] Yannis Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry reranking for large scale image retrieval. International journal of computer vision, 107(1):1–19, 2014.

[7] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1269–1277, 2015.

[8] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), pages 584–599, 2014.

[9] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In Proceedings of the European Conference on Computer Vision (ECCV), pages 404–417, 2006.

[10] Yoshua Bengio, Nicholas L´eonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.

[11] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2):157–166, 1994.

[12] Albrecht Blaser. Data Base Techniques for Pictorial Applications. Springer, 1980.

[13] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), pages 677–694. Springer, 2020.

[14] Bingyi Cao, Andr´e Araujo, and Jack Sim. Unifying deep local and global features for image search. In Proceedings of the European Conference on Computer Vision (ECCV), pages 726–743, 2020.

[15] Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 761–768, 2011.

[16] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang.Mindfinder: interactive sketch-based image search on millions of images. In Proceedings of ACM International Conference on Multimedia (MM), pages 1605–1608, 2010.

[17] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Hashnet: Deep learning to hash by continuation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 5608–5617, 2017.

[18] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems, 33:9912–9924, 2020.

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning (ICML), pages 1597–1607, 2020.

[20] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. arXiv preprint arXiv:2101.11282, 2021.

[21] Yudong Chen, Zhihui Lai, Yujuan Ding, Kaiyi Lin, and Wai Keung Wong. Deep supervised hashing with anchor graph. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 9796–9804, 2019.

[22] Yizong Cheng. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 17(8):790–799, 1995.

[23] Fran¸cois Chollet. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1251–1258, 2017.

[24] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 539–546, 2005.

[25] Ondˇrej Chum, Andrej Mikulik, Michal Perdoch, and Jiˇr´ı Matas. Total recall ii: Query expansion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 889–896, 2011.

[26] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1–8. IEEE, 2007.

[27] Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, and Le Song. Stochastic generative hashing. In Proceedings of the International Conference on Machine Learning (ICML), pages 913–922, 2017.

[28] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems, 34:3965–3977, 2021.

[29] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4690–4699, 2019.

[30] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops, pages 224–236, 2018.

[31] Kamran Ghasedi Dizaji, Feng Zheng, Najmeh Sadoughi, Yanhua Yang, Cheng Deng, and Heng Huang. Unsupervised deep generative adversarial hashing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3664–3673, 2018.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.