Saliency Inside: Learning Attentive CNNs for Content-Based Image Retrieval

Shikui Wei, *Senior Member, IEEE*, Lixin Liao, Jia Li, *Senior Member, IEEE*, Qinjie Zheng, Fei Yang, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— In content-based image retrieval (CBIR), one of the most challenging and ambiguous tasks is to correctly understand the human query intention and measure its semantic relevance with images in the database. Due to the impressive capability of visual saliency in predicting human visual attention that is closely related to the query intention, this paper attempts to explicitly discover the essential effect of visual saliency in CBIR via qualitative and quantitative experiments. Toward this end, we first generate the fixation density maps of images from a widely used CBIR dataset by using an eye-tracking apparatus. These ground-truth saliency maps are then used to measure the influence of visual saliency to the task of CBIR by exploring several probable ways of incorporating such saliency cues into the retrieval process. We find that visual saliency is indeed beneficial to the CBIR task, and the best saliency involving scheme is possibly different for different image retrieval models. Inspired by the findings, this paper presents two-stream attentive convolutional neural networks (CNNs) with saliency embedded inside for CBIR. The proposed network has two streams that simultaneously handle two tasks. The main stream focuses on extracting discriminative visual features that are tightly related to semantic attributes. Meanwhile, the auxiliary stream aims to facilitate the main stream by redirecting the feature extraction to the salient image content that a human may pay attention to. By fusing these two streams into the Main and Auxiliary CNNs (MAC), image similarity can be computed as the human being does by reserving conspicuous content and suppressing irrelevant regions. Extensive experiments show that the proposed model achieves impressive performance in image retrieval on four public datasets.**

Manuscript received December 15, 2017; revised September 2, 2018 and March 14, 2019; accepted April 22, 2019. Date of publication May 2, 2019; date of current version July 16, 2019. This work was supported in part by the National Key Research and Development of China under Grant 2017YFC1703503, in part by the National Natural Science Foundation of China under Grant 61572065, Grant 61532005, and Grant 61672072, in part by the Beijing Nova Program under Grant Z181100006218063, and in part by the Fundamental Research Funds for the Central Universities under Grant 2018JBZ001. An early version of this paper was presented at ACM MM'17 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kalpana Seshadrinathan. *(Corresponding author: Jia Li.)*

S. Wei, L. Liao, Q. Zheng, F. Yang, and Y. Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: shkwei@bjtu.edu.cn; 16112056@bjtu.edu.cn; 1512035@bjtu.edu.cn; 15120352@bjtu.edu.cn; yzhao@bjtu.edu.cn).

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: jiali@buaa.edu.cn).

Digital Object Identifier 10.1109/TIP.2019.2913513

***Index Terms*— Visual saliency, content-based image retrieval, bag-of-word, convolutional neural networks.**

I. Introduction

With the booming of smart phones and digital cameras, the amount of images grows surprisingly fast in our daily life. To maximize the value of such big visual data, it is necessary to develop an image search approach that is capable of efficiently and accurately retrieving images with the desired content. For such a content-based image retrieval (CBIR) approach, one of the key challenges is to infer the inherent query intention expressed by a query image. As shown in Fig. 1, confusion may arise in determining what is the desired content, while the similarity between images may be defined in visual and/or semantic levels [2], [3]. Actually, the ambiguity in capturing the *inherent query intention* acts as a major obstacle in CBIR.

In the past decades, hundreds of approaches have been proposed for fast and reliable CBIR [4]–[8]. Generally speaking, most of existing image retrieval methods attempt to improve image retrieval performance from the following three aspects: 1) constructing discriminative image features [9]–[11], 2) designing good similarity estimation schemes [12]–[15], and 3) handling large-scale issues [16]–[20]. For example, many hashing methods [21]–[26] have been proposed to make the similarity computation faster and more semantic. In particular, recent advances in deep learning [27]–[30] provide an opportunity to overcome the well-known semantic gap in CBIR [31]–[35]. In [31], Razavian *et al.* extracted subpatches from different locations in an image and characterized them with deep features. Such features are then compressed to compute patch-based similarity. Gong *et al.* [32] extracted deep features from patches at different scales and locations by using Convolutional Neural Networks (CNNs) as well as orderless pooling strategies. In [33], local deep features were aggregated to produce compact global descriptors for image retrieval. Typically, such CNN-based approaches can outperform classic SIFT- or GIST-based approaches since the features extracted by CNNs are generally considered to be closer to the semantic attributes of images. However, such features are extracted from the whole image, making them sometimes inaccurate to capture and characterize the inherent query intention (*e.g.*, the desired content).

To develop a CBIR approach that is capable of capturing inherent query intention, we first turn to a fundamental

1057-7149 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

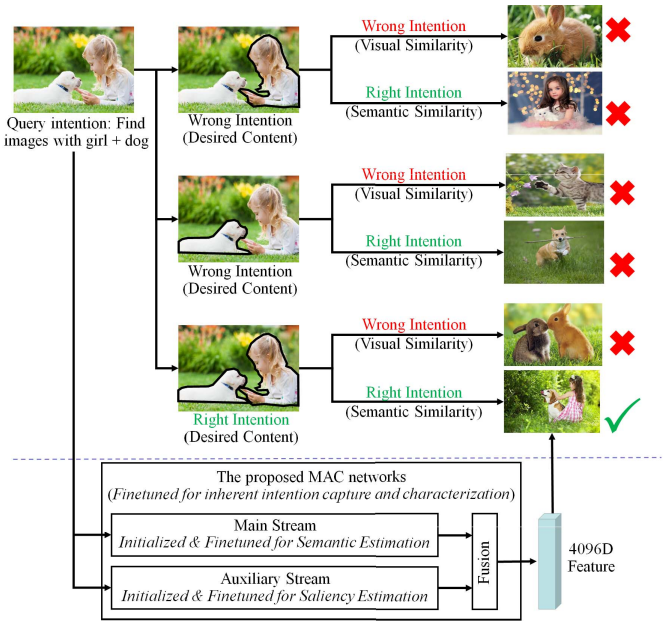

Fig. 1. Capturing the query intention is very important to correctly retrieve the desired images. Toward this end, we propose attentive CNNs that start from a Main stream for semantic prediction and an Auxiliary stream for saliency/attention prediction, which are fused and simultaneously fine-tuned so as to capture and characterize such inherent intention.

question: how does the human being perceive and understand the images to be retrieved? With this question in mind, we first conduct extensive eye-tracking experiments, from which we collect human eye-tracking data in free-viewing images from the Holidays dataset [10], a widely used dataset in CBIR. From these experiments, we obtain a human fixation density map for each image, which reveals the distribution of human visual attention in observing the image. From the perspective of visual saliency, such fixation density maps can be viewed as the ground-truth saliency maps, which we believe are tightly related to the inherent query intention when the corresponding images are involved in CBIR. For the sake of simplification, we use the term saliency and attention interchangeably in the rest parts of this paper.

Given the ground-truth saliency map of each image to be retrieved, we can further investigate an advanced question: is visual saliency really useful in image retrieval? To answer this question, we conduct extensive qualitative and quantitative experiments that incorporate visual saliency into two CBIR systems that both adopt the classic bag-of-word (BoW) framework. In these experiments, the ground-truth saliency maps generated by eye-tracking experiments (*i.e.*, an ideal saliency model) are used to find out the best ways to make use of visual saliency in CBIR. In this process, different saliency involving schemes are adopted, such as saliency filter, saliency intensity embedding, saliency representation embedding and re-ranking. Through these experiments, we find that visual saliency is indeed beneficial to the image retrieval task. However, the best saliency involving scheme is possibly different for different image retrieval models. Therefore, it is necessary to design or even learn a model-specific scheme that can embed saliency cues inside the desired image retrieval model.

To address this problem, we propose two-stream attentive CNNs with visual saliency embedded inside for image retrieval. As shown in Fig. 1, the Main and Auxiliary CNNs, denoted as MAC, start from two separate streams that handle different cognitive tasks. The Main stream is initialized with image recognition model VGG16 [29], while the Auxiliary stream is initialized with fixation prediction model DeepFixNet [36]. In other words, the Main and Auxiliary streams start from the tasks of semantic attribute prediction and visual saliency prediction, respectively. Considering that such semantic and saliency cues are tightly related to but not equivalent to the inherent intention in a query image, we further fuse them and fine-tune the entire networks on existing image retrieval datasets. In this manner, the semantic and saliency cues can be gradually modulated to reflect the inherent query intention. As a result, we can obtain reliable similarity scores between a query image and all candidate images by using the output features of MAC, even with a very simple $\ell 2$ distance measure. Extensive experiments show that our approach achieves impressive performance on four public datasets. In particular, our approach further validates its effectiveness in many synthesized challenging scenarios such as rain/snow and low-resolution/low-quality, implying that it can be suitable for many real-world applications.

The main contributions of this paper are summarized as follows: 1) we extend a classic image retrieval dataset with ground-truth saliency maps, which can be useful to further study the effect of visual saliency in image retrieval; 2) we explicitly evaluate the effect of the ground-truth saliency on image retrieval and discover some effective schemes of involving the saliency information; 3) we propose two-stream attentive CNNs for image retrieval, which achieve impressive performance in image retrieval on four public datasets. The first two new contributions also make the paper remarkably different from its conference version [1].

The rest of this paper is organized as follows: Section II reviews related works and Section III extends a classic image retrieval dataset with the ground-truth saliency maps. In Section IV, we investigate several schemes to involve saliency into image retrieval to find out whether visual saliency is really useful in image retrieval. Section V introduces the proposed attentive CNNs for CBIR. Section VI tests the proposed approach, and the paper is concluded in Section VII.

II. Related Work

In this Section, we review saliency-guided and CNN-based CBIR models that are tightly related to our work.

A. Saliency-Guided CBIR

Visual saliency has been used by many CBIR systems since it can depict the most conspicuous image content. Generally speaking, existing saliency-guided CBIR systems can be divided into two categories: *i.e.*, saliency filtering systems and saliency weighting systems. For the saliency filtering systems, only regions with high saliency values are used for subsequent feature extraction. In these systems, the negative effect of background regions is completely suppressed by directly

discarding non-salient regions. For example, Acharya and Devi [37] directly employed the classic saliency model [38] to generate saliency maps and then extracted feature vectors with respect to these saliency maps for image retrieval. Wen *et al.* [39] extracted SIFT and color features from salient regions to retrieve images. Giouvanakis and Kotropoulos [40] combined a classic attention model with the BoW model by reserving SIFT features only in attention regions.

For the saliency weighting systems, a saliency map is employed to re-weight the features extracted from the whole images. For example, Papushoy and Bors [41] employed the graph-based saliency model [42] to extract saliency maps and introduced the saliency information into region-based image retrieval system. The saliency information was involved by weighting different regions according to their saliency scores. In [43], [44], a histogram of saliency map was extracted as separated image features, and it was integrated into original similarity measure of image retrieval system. In [45], salient contour maps were extracted to localize the objects in images.

In the systems discussed above, visual saliency is used as a global constraint to enhance similarity estimation between two images. Actually, many such saliency-guided models have been proposed, but the performance gained from the straightforward usage of visual saliency is usually not as high as expected. It is still not clear what is the best way to involve visual saliency into CBIR. Moreover, existing saliency models, which are mainly developed for predicting human fixations in free-viewing conditions, may be not suitable for the CBIR task if they are directly used without being fine-tuned with respect to the task. Furthermore, most of such attentive CBIR models are developed based on classic features (*e.g.*, SIFT), which often perform worse than deep features in depicting semantic attributes of the desired image content.

B. CNN-Based CBIR

Due to the remarkable success of deep learning models in many vision tasks [46]–[54], they have been incorporated into image retrieval in many different ways such as local feature extraction [31]–[33], [55], [56], global feature extraction [57]–[60], hashing [61]–[63] and similarity computation [64]–[68]. In local feature extraction, the commonly used paradigm is to replace the traditional local descriptors (*e.g.*, SIFT) with deep features. For example, Paulin *et al.* [55] proposed to learn patch descriptors without supervision. In their approach, the convolutional kernel networks were adopted to extract patch features for matching and instance-level retrieval.

In global feature extraction, some additional cues are generally introduced to enhance the discriminative capability of the original deep features. For example, Razavian *et al.* [31] used Structure-from-Motion (SfM) method to get 3D models, which can guide the selection of deep features. Zheng *et al.* [58] fused various features by extracting the output of pooling layers in VGG and Alexnet for image retrieval.

In CNN-based hashing, a key step is to project deep features into more compact binary codes as well as preserving distance invariance. For example, Xia *et al.* [61] proposed a CNN-based hashing method that broke down similarity matrix and generated the binary encoding results. Zhao *et al.* [63] proposed a hashing method with deep semantic ranking. They used CNNs to learn the ranking of retrieval results and optimize the evaluation index.

In similarity computation, the goal is to reliably estimate the similarity between two images. For example, Zagoruyko and Komodakis [65] proposed to directly learn visual similarity from image pairs by using two-stream networks. Bontar and Lecun [66] learned the similarity measure on small image patches by using CNNs. Zhou *et al.* [64] used the matching function to integrate SIFT and deep features. A threshold exponential match kernel method was proposed to calculate the scores of similar images.

To sum up, CNN-based CBIR approaches demonstrate impressive capabilities in extracting features or learning similarity measures that are closer to "semantic." However, a key challenge for these approaches is: how to extract features only from the desired image content so as to avoid the influence of irrelevant distractors that are beyond the query intention? In other words, existing CNN-based approaches can extract powerful features with unexpected noise beyond the desired image content. On the contrary, attentive CBIR approaches can filter out irrelevant regions. But the classic features used by most attentive models are relatively weak. Moreover, most saliency models are developed for the fixation prediction task in free-viewing conditions. It may be inappropriate to directly use them in the CBIR task without revision. Inspired by these facts, we propose two-stream attentive networks with saliency embedded inside for CBIR, in which the two streams are initialized for fixation prediction and semantic recognition, respectively. These two streams are then fused and fine-tuned together on image retrieval datasets so that the extracted saliency cues and semantic features become more suitable for the CBIR task. The novelty and contribution mainly reside in the way we use saliency cues. Instead of computing the saliency maps with a pre-trained saliency model first and then incorporating them into the retrieval model, we use the saliency extraction model as an auxiliary CNN stream. By fine-tuning this stream on the image retrieval dataset, the saliency cues extracted in this stream become more suitable for the image retrieval task.

III. A CBIR Dataset With Ground-Truth Saliency

To explore the role of visual saliency in CBIR, we need to associate each image with a "ground-truth" saliency map that is accurate enough to discover the inherent relationship between saliency and CBIR. Therefore, we extend the Holidays dataset [10] by associating each image with a fixation density map generated in eye-tracking experiments, which can be used as the ground-truth saliency map that performs the best in depicting the most conspicuous image contents than existing computational saliency models.

The INRIA Holidays dataset is widely used in CBIR. It contains 500 image groups with different scenes or objects. Each group contains multiple images, and the total number of images in all groups is 1491. To evaluate an image retrieval model, the first image in each group is often used as query,

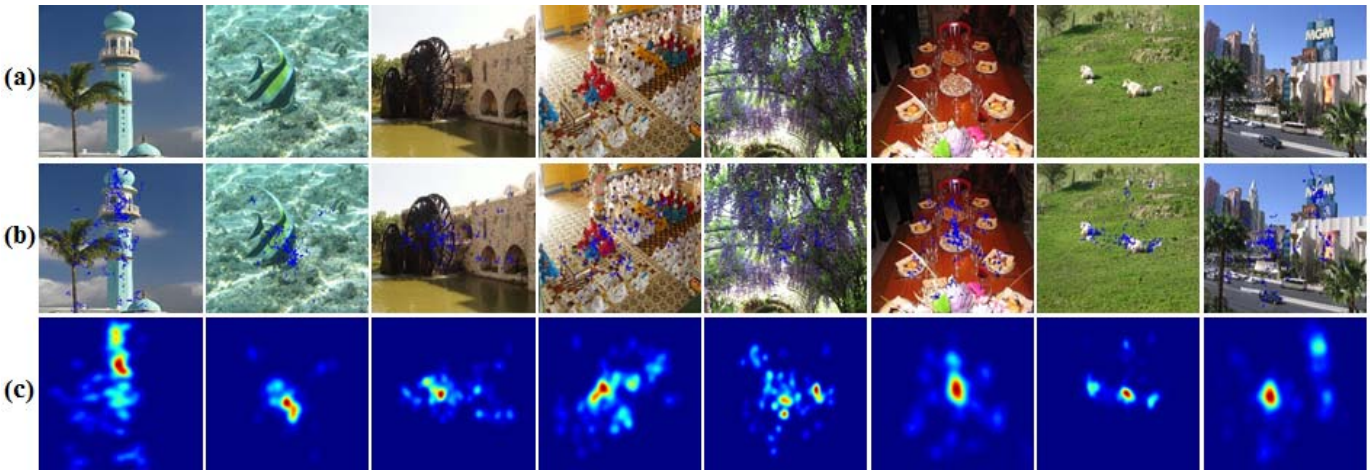

Fig. 2. Ground-truth saliency obtained from eye-tracking experiments. (a) Original images from the Holidays dataset; (b) Fixations of multiple subjects (blue dots); (c) Fixation density maps that can be viewed as the ground-truth saliency maps.

which results in a query set with 500 images. The other 991 images are treated as database images.

To extend the dataset with ground-truth saliency maps of all images, we first manually divide the 1491 images into 30 subsets with about 50 images per subset. In the division, images in each subset are expected to contain different scenes or objects so that the visual fatigue in eye-tracking experiments can be alleviated. Note that different subjects may have different visual attention regions when watching the same image. However, the spatial distribution of the attention regions from multiple subjects can become stable when the number of subjects free-viewing the same image grows large, leading to a stable ground-truth saliency map. In this study, we follow the common settings of many previous visual saliency estimation works that determine the ground-truth saliency map of an image by recording the visual attention of multiple subjects. For each subset of images, we request 12 subjects (6 males, 6 females, aged between 21 and 27) to free-view all the images, and their eye movements are recorded with a SMI RED 500 eye-tracking apparatus. Note that we divide the 12 subjects into several groups, and subjects in each group free-view image subsets in an interlacing manner so that they can avoid getting tired in eye-tracking experiments.

In the experiments, a subject sits in a dark room and uses a chin rest for head stabilization, and a calibration operation is conducted before the first image from each subset is free-viewed. Each image will be displayed for three seconds, and we display a one-second gray screen between any two images to clean up human visual memory. By collecting the fixations of all the 12 subjects in their three-second free-viewing process, we can obtain a fixation density map for each image by replacing each fixation with a small Gaussian blob and accumulating all these blobs. As shown in Fig. 2, such fixation density maps can be viewed as the ground-truth saliency map for images from the Holidays dataset. For the sake of simplification, we normalize each map to the dynamic range of [0, 1].

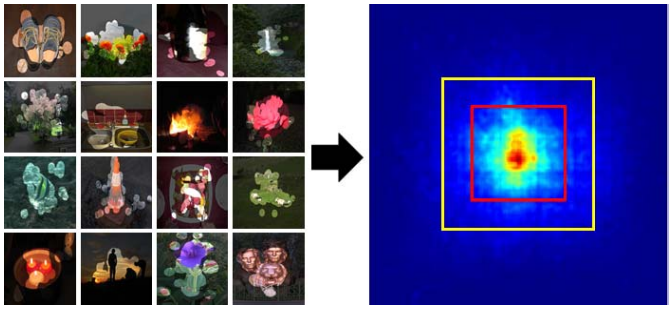

Fig. 3. Fixated regions often distribute around image centers. Left: Salient image regions being free-viewed in eye-tracking experiments. Right: the average annotation map that accumulates per-image fixation density maps. We find that the fixations have a strong center-bias. Approximately 52% and 79% of fixations lie in the red (center 10% area) and yellow (center 25% area) square boxes, respectively.

To better understand the attributes of ground-truth saliency maps, we resize the fixation density maps of all images into the same resolution and overlay them to form a single average annotation map. As shown in Fig. 3, the fixated image regions often distribute near image centers, and the average annotation map in Fig. 3 shows a strong bias towards image centers too. We find that 52% of fixations from all images distribute within a small square box in the average annotation map (*i.e.*, the center 10% area), and 79% of fixations fall in a bigger square box (*i.e.*, the center 25% area). These statistical results are consistent with the analysis of many previous saliency models [69]–[71], implying that many computational saliency models can be reused on the Holidays dataset to enhance the CBIR performance.

IV. Is Visual Saliency Really Useful in CBIR?

Given the ground-truth saliency, we can turn to an advanced question: is visual saliency really useful in image retrieval? Although there already exist many previous works [37], [41], [44] that try to seek an answer by using computational saliency models [38], [42], the ground-truth saliency

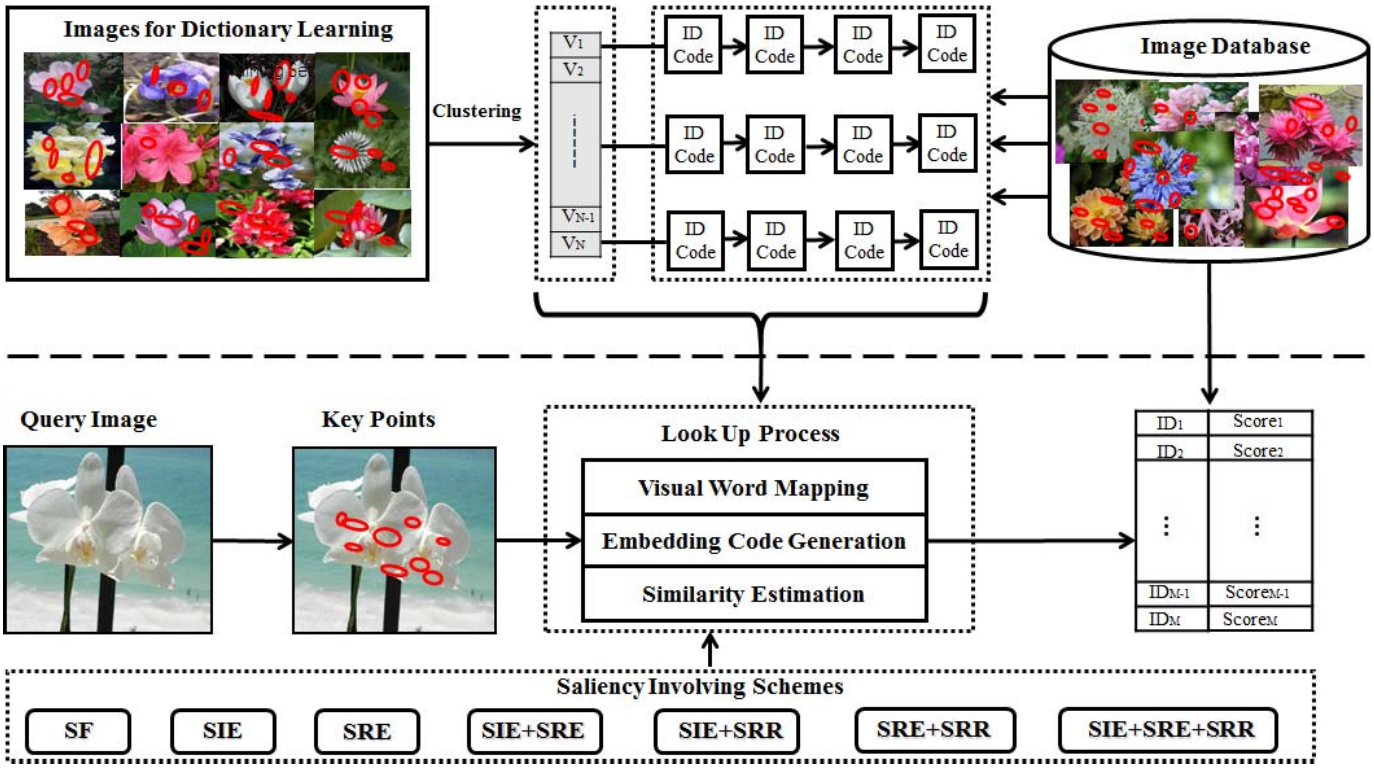

Fig. 4. The CBIR framework for testing various saliency involving schemes. Under this framework, two local feature-based CBIR models are implemented, denoted as BoW (*i.e.*, the classic Bag-of-word model) and BoW-HE (*i.e.*, the BoW model with Hamming embedding code [11]), respectively. In the off-line indexing stage, each key point in database images is first mapped to the nearest visual word, and then its image ID without (BoW) or with (BoW-HE) embedding code is inserted into the list corresponding to the visual word. In the online query stage, each key point in the query image is also mapped to the nearest visual word, and the items in the corresponding list are returned as matches. If embedding codes are employed, the returned list will be further refined and only top n items whose are most similar to query key point in distances among their codes are returned as matches. In this process, four saliency involving schemes (and their combinations) are tested, including SF (saliency filter), SIE (saliency intensity embedding), SRE (saliency representation embedding) and SRR (saliency representation re-ranking).

provides a unique opportunity to look deeper into this question: finding the best ways to make use of visual saliency in CBIR since the ground-truth saliency can be viewed as the predictions of an ideal saliency model. Toward this end, we conduct extensive experiments in this section to discover the relationship between visual saliency and image retrieval on the extended Holidays dataset with the ground-truth saliency. More specifically, a popular image retrieval framework based on local image features is employed as the baseline, and four heuristic schemes are designed to involve saliency into CBIR.

A. The Image Retrieval Framework

In conducting the experiments, we employ the classic BoW-based framework since it is more explainable and easier to understand. In addition, it can also provide a high flexibility for involving saliency information into CBIR with balanced effectiveness and efficiency, which will provide an intuitive impression on the influence of different saliency involving schema. As shown in Fig. 4, a dictionary of visual words is first constructed by employing previous clustering methods [10], [11]. Based on the dictionary, an image can be represented as an orderless collection of visual words by replacing its local features with the nearest visual words. After inserting the visual words and their corresponding image IDs into the inverted table, the image retrieval process can be efficiently performed. Beyond such a classic BoW model, we also adopt the BoW-based approach proposed in [11] that further makes use of the Hamming embedding code (denoted as BoW-HE). The Hamming embedding code encodes the quantization error between a local feature and its visual word, which can be also inserted into the inverted table to further refine the retrieval results. Based on these two retrieval models, we wish to find out some common trends when the ground-truth saliency is involved into the classic BoW model and its improved descendants.

B. Saliency Involving Schemes

Typically, there are many ways that saliency can be involved into the BoW-based image retrieval framework. By investigating and summarizing the solutions adopted in previous works, we design four saliency involving schemes, including:

1) Saliency Filter (SF): For both query and database images, only the key points with saliency scores above a predefined threshold are reserved for retrieval. All the other key points are directly abandoned.

2) Saliency Intensity Embedding (SIE): Instead of directly filtering out non-salient key points, the saliency values of key points can be also embedded into the similarity computation process as re-weighting factors. Similar to the SF scheme, the SIE scheme only utilizes key points with saliency scores

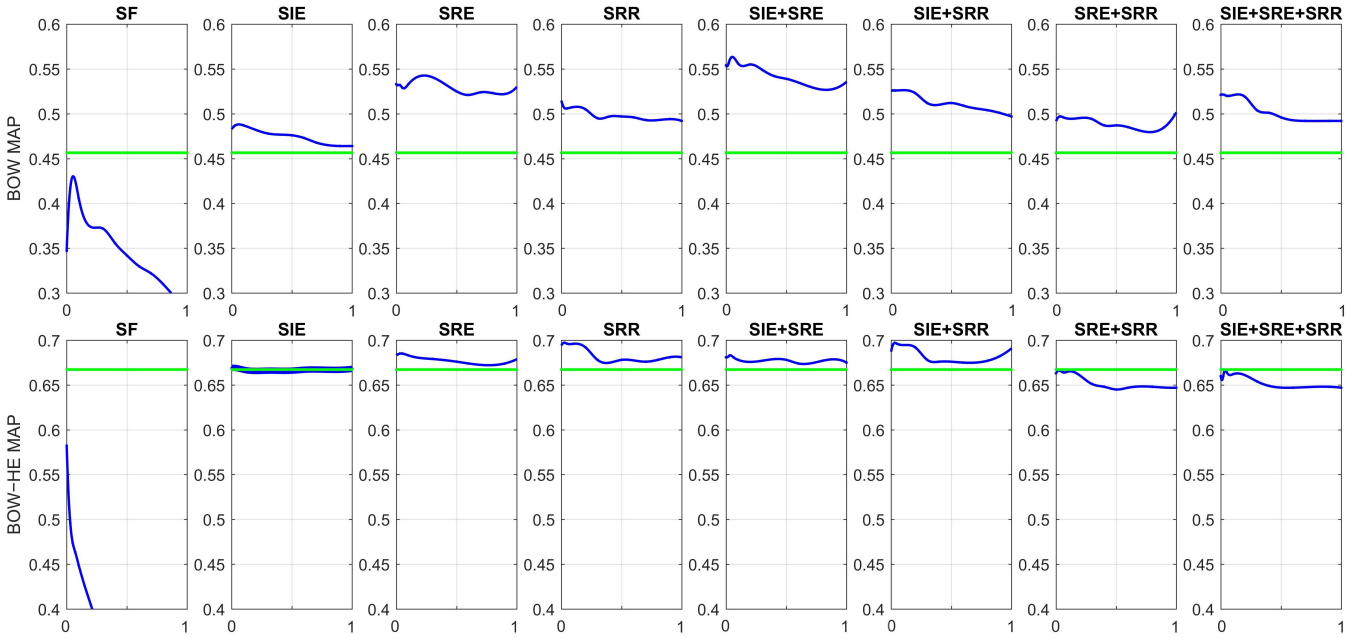

Fig. 5. Performance of BoW and BoW-HE models when using various saliency involving schemes. The green lines in figures from the first and second rows denote the performance of BoW and BoW-HE models, respectively. The blue curves show the performance after using various saliency involving schemes at different thresholds (*i.e.*, the horizontal axis).

above a predefined threshold. However, it assumes that two key points with similar saliency scores should be assigned with a large weight in the similarity computation. Let P_1 and P_2 be two key points with saliency scores $S(P_1)$ and $S(P_2)$, respectively. We adopt a linear function to map the saliency difference $|S(P_1) - S(P_2)|$ to a non-negative weight. Note that the parameters of the function is manually fine-tuned in experiments to maximize the performance of the SIE scheme.

3) Saliency Representation Embedding (SRE): In the third scheme, we extract additional features for both query and database images. We first find out the regions whose ground-truth saliency values are higher than the predefined threshold. After that, we compute the average values from color channels of these regions (*e.g.*, hue, saturation and intensity). These color values are then combined to form an image-specific saliency representation, which is then embedded into the feature vectors of all key points within the same image.

4) Saliency Representation Re-Ranking (SRR): Similar to the SRE scheme, the SRR scheme also computes an image-specific saliency representation for each image by using a predefined threshold. Instead of embedding this saliency representation into key points, we use such saliency representations to re-rank the images retrieved from the database.

Beyond these four saliency involving schemes, we also test their combinations, including SIE + SRE, SIE + SRR, SRE + SRR and SIE + SRE + SRR. Note that the SF scheme is not combined with the other schemes since it can be viewed as a special case of SIE (*i.e.*, a constant re-weighting function that outputs the same weights for all keypoint pairs).

C. Comparisons of Saliency Involving Schemes

Given the BoW and BoW-HE models as well as the saliency involving schemes, we perform quantitative comparisons on the Holidays dataset. Different from previous works that adopt imperfect computational saliency models, we use the ground-truth saliency maps in the experiments. Considering that the ground-truth saliency maps are the ultimate objective all existing saliency models wish to approximate, we can safely assume that the saliency model used in the experiments is "perfect" and the performance variation is only influenced by the ways we use visual saliency in CBIR. In these experiments, the Mean Average Precision (MAP) is employed as the evaluation measure. By employing the saliency schemes and varying the threshold, we demonstrate the performance variation of the two baseline models in Fig. 5.

From the curves at the top row of Fig. 5, we can clearly find that almost all saliency involving schemes can boost the accuracy of the classic BoW model in image retrieval. These results imply that the saliency information may have the capability to identify intention regions, localize regions-of-interest or suppress the interference of irrelevant regions. An outlier is the SF scheme, which even makes the retrieval performance worse than the baseline BoW model. A possible explanation is that visual saliency is over emphasized in the SF scheme, while non-salient regions such as the visual context of targets can also help to retrieve the desired content (*e.g.*, the grassland can help to retrieve a cow). By directly filtering out these visual contexts other than suppressing them with smaller weights, the SF scheme leads to a severe loss of some high discriminative key points from visual context of the desired content and thus significantly reduce the image retrieval performance. We also find that the best performance is achieved by the SIE + SRE scheme, which improves the performance of baseline model from 0.456 to 0.560. These results imply that visual saliency should be used by emphasizing conspicuous image contents and extracting

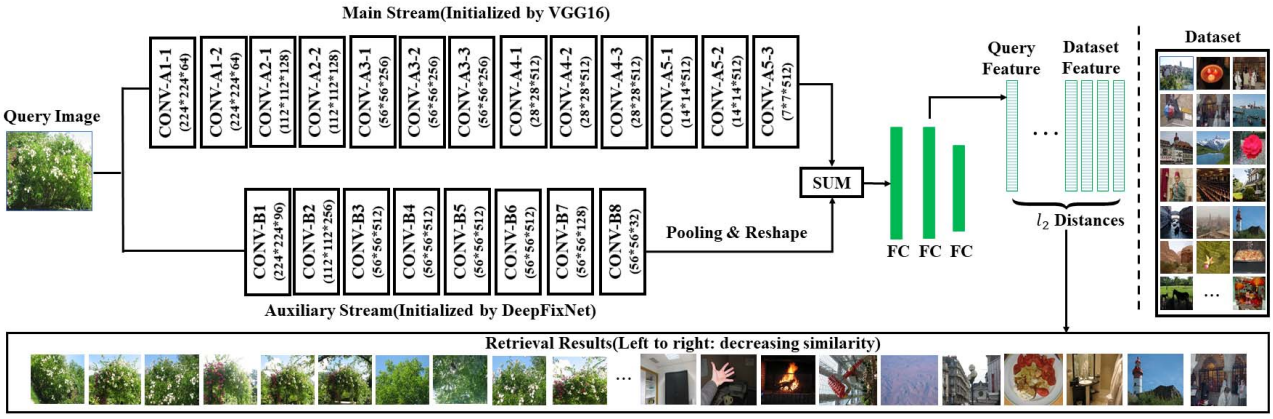

Fig. 6. The framework of the proposed CBIR system. The core of this framework is the attentive CNNs with saliency cues embedded inside. The proposed CNNs contain two separate streams. The Main stream is initialized with the semantic prediction networks VGG16, while the Auxiliary stream is initialized with the fixation prediction networks DeepFixNet. These two streams are then fused and simultaneously fine-tuned on image retrieval datasets so as to extract features that can well capture and characterize the inherent query intention. Finally, such features are used to measure the similarity (computed as the $\ell 2$ distance) between a query image and all images in the database.

additional feature representations so as to improve the CBIR performance.

From the curves at the bottom row of Fig. 5, we find that sometimes the performance variations of BoW-HE are inconsistent with the BoW model. The best performance is achieved by the SRR scheme, which improves the performance of the baseline BoW-HE from 0.667 to 0.697. Moreover, both the SF, SRE + SRR and SIE + SRE + SRR schemes are worse than the original BoW-HE model. This can be explained by the fact that embedding codes used in the BoW-HE model have screened the items (content information) returned for a query key point, and the final list is already quite reliable and relatively short. In essence, the saliency involving schemes attempt to further filter out or suppress some irrelevant items in the list. However, if the number of returned items is too small, involving such a saliency scheme will cause a loss of discriminative information and thus lead to performance degradation. For example, the MAP values from the SF scheme drop sharply when increasing the threshold (*i.e.*, reducing candidate items). Actually, in designing saliency involving schemes, we are seeking a trade-off between saliency cues and non-salient context.

From these experiments, we can tentatively draw two conclusions: 1) visual saliency is indeed beneficial to image retrieval task, and our extension to the Holidays dataset provide an opportunity to assess saliency involving schemes with an ideal saliency model; 2) the best saliency involving scheme is possibly different for different image retrieval models. In other words, it is necessary to carefully select or even learn a best scheme that fits for the adopted CBIR model.

V. Two-Stream Attentive Networks for CBIR

All experiments presented above are conducted on the new benchmark. That is, each image is associated with a fixation density map. In real-world CBIR systems, however, it is very difficult to obtain such ground-truth saliency maps for images from large-scale databases. Therefore, most of saliency-based CBIR methods directly extracted saliency maps with existing saliency models. However, different saliency models will result in quite different saliency maps, which can be treated as different approximations of the ground-truth saliency map. In most cases, such saliency maps are quite different from the ground-truth, and it is also not clear which schemes perform the best in involving such imperfect saliency cues into the retrieval process. To address these problems, we propose attentive CNNs with saliency embedded inside for CBIR. In this section, we first introduce the architecture of the proposed Main and Auxiliary CNNs (MAC), followed by the details about how to train such a model and how to use it in CBIR.

A. System Framework

As shown in the framework of Fig. 6, the core of the proposed CBIR approach is the two-stream attentive networks with saliency embedded inside, which can be denoted as Main and Auxiliary CNNs (MAC). Different from previous works, MAC has two separate streams that are initialized for different cognitive tasks. Both streams take a 224×224 image with three channels as the input.

The Main stream is initialized as the first five major convolutional and pooling groups of the VGG16 networks (we display only the convolution layers in Fig. 6 due to space limitation). Finally, the major stream will output a 7×7 map with 512 channels, while such a map, denoted as a 3D matrix $\mathbf{X}_{main}$, contains high-level cues extracted from both the desired image content and the irrelevant regions. As a result, such feature maps need to be further refined to obtain cleaner *semantic* features that can characterize the inherent query intention.

Toward this end, we incorporate the Auxiliary stream to filter out the unexpected *noise* from the original features extracted by the Main stream. Saliency cues are also embedded into MAC by using this stream. To maintain the features that may be useful for the task of image retrieval, we initialize this

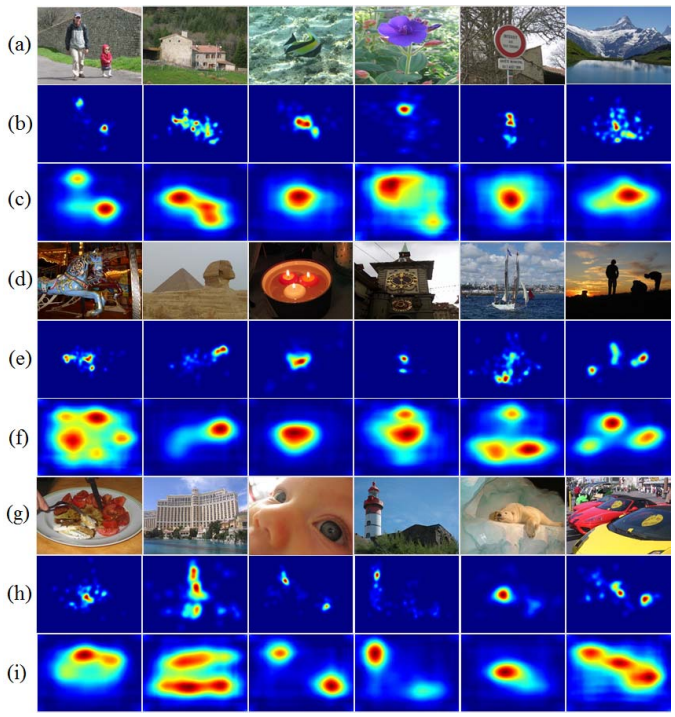

Fig. 7. DeepFixNet achieves impressive performance in predicting the most salient targets. However, it also pops-out a large portion of the visual context near the most salient targets and needs to be further fine-tuned on image retrieval datasets before being applied in CBIR. (a,d,g) Images from the Holidays dataset; (b,e,h) Ground-truth saliency maps; (c,f,i) Saliency maps predicted by DeepFixNet.

stream with the DeepFixNet [36], the CNNs that are developed for the fixation prediction task. In the initialization, we select the first eight convolution layers together with the related pooling layers and generate a 56×56 map with 32 channels. After that, such a feature map enters a pooling layer and is then reshaped to 7×7 maps with 512 channels. Note that in the image retrieval task, the effect of brute transformation is not as remarkable as in many location-sensitive tasks such as object detection and segmentation. In other words, image retrieval is more sensitive to the occurrences of some specific visual patterns other than their location. Similar to the Main stream, the output map is denoted as $\mathbf{X}_{aux}$.

Compared with heuristic saliency models, DeepFixNet gains impressive performance in predicting the most salient locations (see Fig. 7). With this stream, we can filter out the features from regions that are irrelevant to the inherent query intention. However, Fig. 7 also shows that the results of DeepFixNet still look different from the ground-truth saliency maps (*e.g.*, it also pops-out a large portion of the visual context near the most salient targets. As a result, it may not perfectly meet the specific requirement of visual saliency in the image retrieval task. Therefore, the parameters of this stream, as well as the Main stream, need to be further fine-tuned on image retrieval datasets. Toward this end, we first conduct element-wise fusion of the output maps from the Main and Auxiliary streams:

$$\mathbf{X}_{\text{fuse}} = \lambda \mathbf{X}_{\text{main}} + (1 - \lambda)\mathbf{X}_{\text{aux}}, \tag{1}$$

where λ is empirically set to 0.6 to balance the output features from the two streams (the influence of λ will be discussed in experiments). In this manner, the fused feature map contains both semantic and saliency cues, which is converted to a lower dimensional feature vector via three consecutive Fully Connected layers, denoted as `FC6`, `FC7` and `FC8`, respectively. Note that both `FC6` and `FC7` output 4096D feature vectors, while `FC8` outputs a vector with N components. For an image retrieval dataset, N denotes the number of categories formed by aggregating training images with similar contents (such similarity is manually annotated by the human being). By applying a softmax layer after `FC8` to turn its output to a probability vector, we can train a classification network on image retrieval datasets by solving the minimization problem:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \sum_{n=1}^{N} log(p_{kn}, y_{kn}) + \beta \|\mathbf{W}\|_2^2, \tag{2}$$

where $\mathbf{W}$ denote the set of network parameters of MAC. p_{kn} is the nth component of the probability vector generated by the final softmax layer of MAC in processing the kth training image. y_{kn} is a binary indicator which equals to 1 only if the kth training image belongs to the nth category of similar training images. β is a constant that controls the norm of parameters in MAC.

By minimizing the prediction error (2), MAC gains the capability to aggregate similar images and separate dissimilar images, while such similarity is defined from the perspective of image retrieval. In this manner, the features generated by MAC can well capture and characterize the inherent query intention for CBIR. In training MAC, we adopt the Caffe platform [72] and utilize a batch size of 16. The learning rate is initialized as 10^{-6}, which will decrease twice, by a factor of 10, after the 33% and 66% of the maximum iteration number have been reached, respectively. Moreover, a weight decay of 0.0005 and momentum of 0.9 are used.

After the training stage, the two-stream attentive CNNs can be used for image retrieval. Considering that the feature dimension of `FC8` varies with respect to different training data, we adopt the 4096D feature vector generated by the `FC7` layer of MAC to characterize the inherent query intention of a new query image. After that, the similarity scores between this feature vector and those pre-computed for the images in the database can be computed. Since the main objectiveness is to demonstrate the powerfulness of the proposed two-stream attentive CNNs, we only use the simplest $\ell 2$ distance as the similarity measure, which can already generate impressive performance by using the powerful features from MAC.

TABLE I
DETAILS OF THE 4 BENCHMARKING DATASETS

Datasets	Total	Training	Testing	Categories
Paris	6,392	5,192	1,200	12
UKBench	10,200	5,100	5,100	2,550
Flower	7,169	6,149	1,020	102
Bird	11,788	6,788	5,000	200

VI. EXPERIMENTS

We conduct extensive experiments to validate the effectiveness and scalability of MAC from multiple perspectives,

including: (1) Effectiveness test that compares MAC with other CBIR models; (2) Scalability test that adds one million images into testing datasets or synthesizes more challenging real-world scenarios like rain/snow and low-resolution/low-quality; (3) Performance analysis that investigates the performance variation of MAC by changing key parameters. Detailed experiments are described as follows.

A. Settings

To conduct comprehensive evaluation of MAC, we adopt four datasets from the areas of image retrieval and image classification. Among these datasets, Paris [73] and UKBench [74] are two image retrieval datasets that are widely used in the literature. Flower [75] and Bird [76] are two fine-grained image classification datasets which can be also used to benchmark the image retrieval models [12], [77]. For each dataset, we split them into a training set and a testing set. It is worth noting that the manner employing the datasets for evaluation is quite different between the proposed method and previous works. Taking the Paris dataset (6392 images, except 20 corrupted images) as an example, previous methods extract 55 images corresponding to 11 distinct buildings as query images, and all the 6392 images are treated as the image database for retrieval. In contrast, our approach first randomly selects 5192 images from the Paris dataset to train the network parameters and leave the rest 1200 images as the query dataset as well as the image database. That is, we choose each of the 1200 images as the query image to retrieve the image from the 1200 images. The main reason for the division lies in that the proposed approach is a supervised model and labeled training data is needed. In the settings, the retrieval operations are much larger and the retrieval scenario is more challenging, leading to much lower mAP. Details of the four datasets can be found in Tab. I.

As in Sect. IV, we adopt the Mean Average Precision (mAP) as the evaluation metric. Instead of using the official evaluation measure on UKBench dataset, we also choose mAP as the evaluation metric on the UKBench dataset due to our experimental settings. In experiments, we randomly choose two images from the four images of each of the 5100 classes of UKBench to construct training dataset, while the rest images are used as testing. In this way, the official evaluation measure becomes inappropriate since the best performance approaches to 2 other than 4 in this settings. To provide a unified and consistent evaluation, we employ mAP as evaluation metric for all datasets.

One problem in comparing image retrieval models is that the performance of learning-based models may vary remarkably before/after being fine-tuned on specific training data. Therefore, we compare MAC with three state-of-the-art models and four baselines, including:

(1) BOW-HE [11]: A non-deep approach that jointly optimize Bag-of-Words and embedding methods for CBIR.

(2) Siamese [65]: Two-stream CNNs that take a pair of images as the input and output the similarity scores.

(3) Base-VGG: A baseline model formed by directly using the 4096D features from the original VGG16 networks and the same retrieval settings with MAC.

TABLE II
EFFECTIVENESS TEST OF FIVE MODELS ON FOUR DATASETS

	Paris	UKBench	Flower	Bird
BOW-HE [11]	0.12	0.81	0.05	0.004
Siamese [65]	0.11	0.26	0.03	0.01
Base-VGG	0.29	0.87	0.31	0.17
Base-VGG-F	0.46	**0.92**	0.60	0.26
Crow [51]	0.25	0.88	0.34	0.15
Selective [52]	0.30	0.79	0.40	0.05
R-MAC [53]	0.12	0.29	0.04	0.01
MAC	**0.48**	**0.92**	**0.64**	**0.28**

(4) Base-VGG-F: Different from Base-VGG, Base-VGG-F is further fine-tuned on the same training data used by MAC in all experiments so that the 4096D features it generated is refined for the retrieval tasks.

(5) Crow [51]: A state-of-the-art method on image retrieval which is based on aggregated deep convolutional features with cross-dimensional weighting.

(6) Selective [52]: A state-of-the-art method on image retrieval which is based on selective deep convolutional features.

(7) R-MAC [53]: A state-of-the-art method on image retrieval which is based on a global representation obtained by aggregating many region-wise descriptors based on the convolution maps.

B. Effectiveness Test

In the effectiveness test, we fine-tune MAC and Base-VGG-F on the training set of each dataset and compare them with other models on the testing set. Performance of all approaches can be found in Table II. Some representative retrieval results of MAC can be found in Fig. 8.

From Table II, we can see that the proposed MAC model achieves impressive performance on all the four datasets. In particular, the MAC network outperforms Base-VGG-F, even when they are fine-tuned on the same training data. This may be caused by the fact that, after incorporating the Auxiliary stream, the semantic features from irrelevant regions can be removed, and the retrieval process will mainly focus on comparing the "desired content" shared by query and target images. In other words, with the assistance of feature maps from the Auxiliary stream, the Main semantic stream perform better in distinguishing images from different categories by focusing on the right regions. Moreover, both the Main semantic stream and the Auxiliary stream are fine-tuned on image retrieval datasets. In this manner, we can assume that both the semantic features and the saliency cues extracted by the two streams become more suitable for the task of image retrieval. That also explains the remarkable performance enhancement from Base-VGG to Base-VGG-F after fine-tuning the original semantic attributes on image retrieval datasets. It is worth noting that both Crow [51] and Selective [52] methods employ provided bounding boxes to crop the query images in their original experiments. Similarly, the learned R-MAC [53] also directly treats the region in the manually labeled bounding box as the input of its network when processing the query images

Fig. 8. Representative results of MAC on four datasets. Red and blue borders indicate query images and the top three retrieval results, respectively. Rows 1-2: Paris; Rows 3-4: UKBench; Rows 5-6: Flower; Rows 7-8: Bird.

in the Oxford and Paris datasets. However, it is rare in the real-world application scenarios to provide bounding boxes of query images. For fair comparison, the query images are not cropped for all methods in our experiments. That is why the performances of Crow [51] and Selective [52] are lower than the base VGG model on some datasets.

Moreover, from Table II we find that the proposed MAC network is not only suitable for the classic image retrieval datasets but also fits for the fine-grained image classification datasets. As shown in Fig. 8, MAC can successfully retrieve images with fine-grained birds and flowers. This is an interesting findings, implying that the usage of the Auxiliary stream also helps maintain the unique attributes of the desired objects while refining the noisy features. Actually, the fine-grained classification/retrieval is much more sensitive to the noise from irrelevant regions, while visual saliency cues can help to "neglect" such regions in feature retrieval. In other words, the semantic stream mainly learns about what is a bird, while the Auxiliary stream may help in learning where is the right place to extract such features. From these results, we can safely conclude that incorporating an additional attention stream is effective for the task of image retrieval.

In addition, we also conduct an experiment to verify the consistency between the saliency cues extracted by the auxiliary stream and the human visual attention-based saliency maps. The experiment is conducted on the Holidays dataset. On this dataset, we randomly select 1191 images and their ground-truth saliency maps from the Holidays dataset to train a simple deconvolution layer for the auxiliary stream, while the rest 300 images are used for testing. In training the deconvolution layer, we take the saliency cues of the auxiliary stream as the input and learn to directly output a saliency map. On the rest 300 images of the Holidays dataset, we use the shuffled AUC (denoted as sAUC) as the evaluation metric to measure the consistency between the predicted and ground-truth saliency maps. We find that the average sAUC score of the saliency maps recovered from the saliency cues of the auxiliary stream reaches up to 0.64, while the original DeepFixNet only reaches 0.53. This indicates the saliency cues extracted by the auxiliary stream are consistent with the human visual attention-based saliency maps.

TABLE III
SCALABILITY TEST ON FOUR DATASETS AFTER ADDING A ONE-MILLION CONFUSION IMAGES FROM FLICKR1M

	Paris	UKBench	Flower	Bird
Base-VGG-F	0.10	**0.88**	0.20	**0.12**
MAC	**0.11**	**0.88**	**0.25**	**0.12**

C. Scalability Test

Beyond effectiveness test, we also conduct several experiments to validate the scalability of MAC (and the baseline models). Toward this end, we first incorporate the one million images from the Flickr1M dataset [10] into the testing

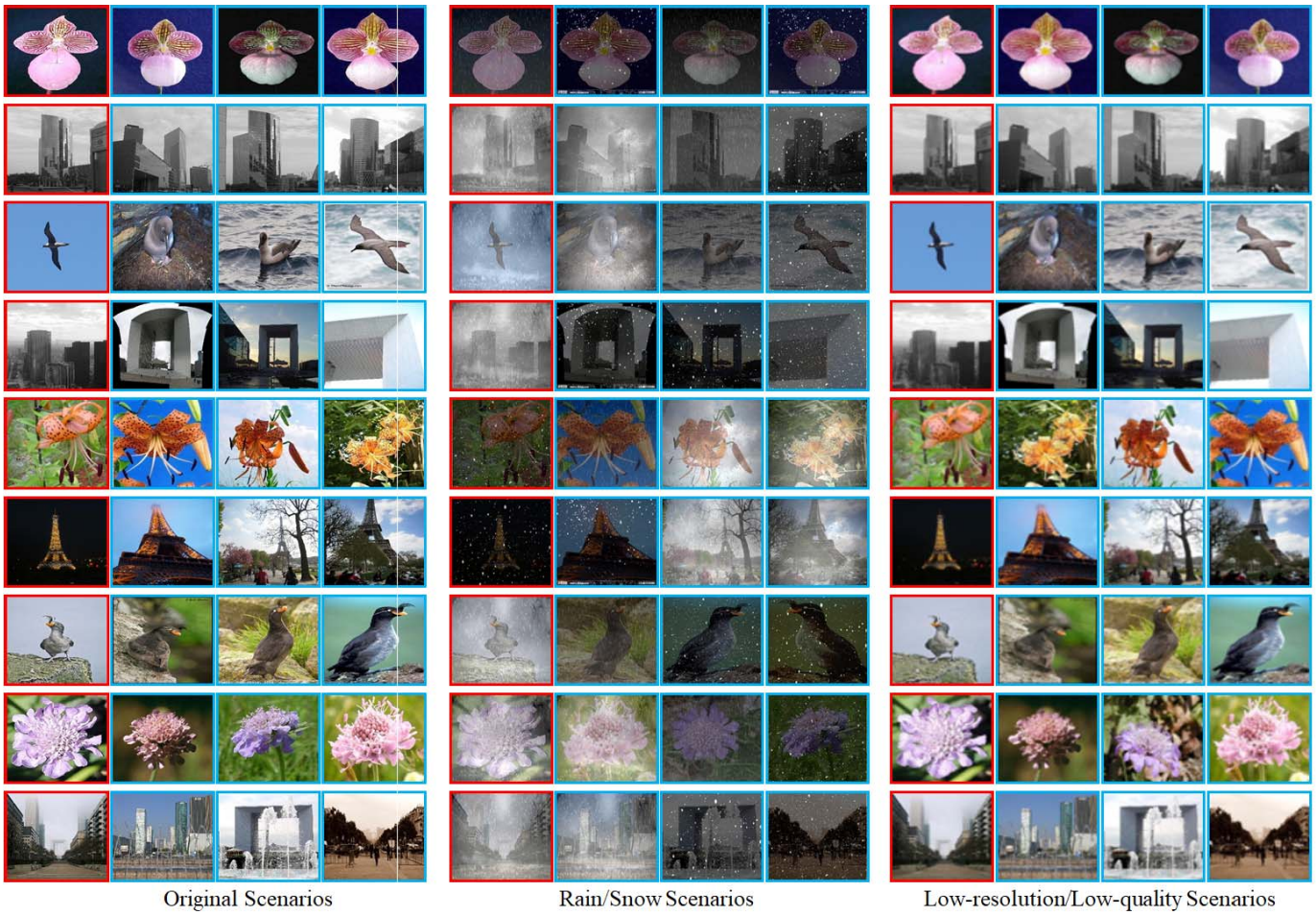

Fig. 9. Representative retrieval results of MAC in rain/snow and low-resolution/low-quality scenarios. Red and blue borders indicate query images and the top three retrieval results, respectively. Left column: results on original datasets; Middle column: results on datasets with synthesized rain/snow; Right column: results on datasets with degraded resolution/quality.

TABLE IV
PERFORMANCE OF FOUR MODELS ON SYNTHESIZED LOW-RESOLUTION/LOW-QUALITY SCENARIOS

Models	Paris	UKBench	Flower	Bird
Base-VGG	0.24	0.84	0.26	0.07
Base-VGG-F	0.38	0.87	0.49	0.15
MAC	**0.41**	**0.88**	**0.53**	**0.17**

sets of each dataset and run the retrieval experiments again. Experimental results are shown in Table III. By comparing the results in Table III and Table II, we can see that even with so many confusion images the retrieval performance of both MAC and Base-VGG-F drop sharply on most datasets. However, in such a challenging setting the performance on UKBench still reaches 0.88. Considering that Flickr1M contains many objects like flower and bird, the performance of MAC on the fine-grained datasets Flower and Bird are still acceptable, implying that the MAC is a scalable network.

Beyond adding confusion images, in actual life many images uploaded to the Internet are low-quality/low-resolution ones. To further validate the effectiveness of our approach, from the four datasets we generate their low-resolution version and test the performance of MAC and baseline models Base-VGG and Base-VGG-F. The performance scores are shown in Table IV, while some representative results are shown in Fig. 9. By comparing Table IV and Table II, we can see that the performance only slightly decreases, while the results in Fig. 9 validates that the proposed approach is scalable to low-quality and low-resolution scenarios. Moreover, in such scenarios MAC still outperforms Base-VGG and Base-VGG-F. This may be caused by the fact that the saliency maps are less sensitive to resolution variation, and many saliency models will resize the input image to an extremely low resolution (*e.g.*, 32×32 in [78]) to speed up the computation process. When the resolution decreases, the Auxiliary stream still outputs reliable cues that assists the localization of desired content, making the whole network more reliable.

Moreover, many images in our daily life are taken in rain or snow, and it is necessary to develop a model that can effectively retrieve such images. To test the performance of image retrieval models in such scenarios, we add synthesized rain/snow to the four datasets. As shown in Tab. V and Tab. II, the performance scores of both MAC and the two baseline models decrease in rain/snow scenarios. In particular, the

TABLE V
PERFORMANCE ON SYNTHESIZED RAIN/SNOW SCENARIOS

Models	Paris	UKBench	Flower	Bird
Base-VGG	0.17	0.35	0.10	0.02
Base-VGG-F	0.36	0.47	0.38	0.09
MAC	**0.40**	**0.48**	**0.46**	**0.11**

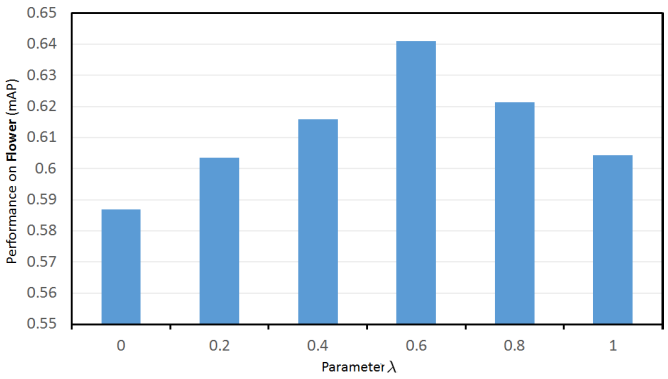

Fig. 10. Influence of fusion weight λ on Flower dataset.

performance on UKBench drops remarkably since it contains many large-scale scenes, while the other three datasets contain large objects that are less influenced by rain and snow, leading to smaller performance drop. Actually, rain and snow can be viewed as additive noise to the original images, while such noises can be viewed as outliers in a local region. In the convolutional operations of CNNs, such outlier will lead to unexpected local maximum or minimum, while such wrongly extracted local extremum will lead to inaccurate semantic features in the Base-VGG-F. Surprisingly, the performance decrease in MAC is often less than Base-VGG-F, which may be caused by the fact that the Auxiliary stream is capable to ignore such frequently appeared fake local extremum and enforces the semantic streams focus on the attractive regions. These results further validate the scalability of the proposed two-stream attentive CNNs.

D. Performance Analysis

Finally, we conduct an experiment to see the influence of the parameter λ, which controls the way that the two streams are fused. By varying λ from 0.0 to 1.0 with a step of 0.2, we test the performance of MAC on the Flower dataset and obtain a performance curve (as shown in Fig. 10). We find that over-emphasizing either stream will lead to degraded performance, and the best performance is achieved at $\lambda = 0.6$, indicating the Main stream has weight 0.6 and the Auxiliary stream has weight 0.4. This also implies that semantic attributes play the most important role in CBIR, while saliency cues provide supplementary cues to improve the retrieval effect.

VII. CONCLUSION

In this paper, we make comprehensive and systematic study to explicitly discover the effect of visual saliency on image retrieval in a quantitative manner. The key finding is that salient information indeed has positive effect on image retrieval and it is difficult for hand-crafted involving schemes to best adapt a specific image retrieval model. To naturally involve salient information into image retrieval in a self-learning and optimal manner, we propose two-stream attentive CNNs for image retrieval. By initializing a Main stream for semantic feature extraction and an Auxiliary stream for saliency prediction, the two-streams fused and fine-tuned on image retrieval datasets. In this manner, the capability of the whole network in capturing inherent query intention can be improved. Experimental results show that the proposed approach has impressive performance on two image retrieval datasets and two fine-grained image classification datasets. Moreover, its performance on retrieving low-resolution/ low-quality and rain/snow images are also very promising.

In our future work, we will seek to train a network with saliency cues embedded in several locations of semantic feature extraction so as to extract more discriminative features for outdoor scenes. Moreover, the hashing operations will be embedded into the network so that the retrieval process can become much faster.

REFERENCES

[1] F. Yang, J. Li, S. Wei, Q. Zheng, T. Liu, and Y. Zhao, "Two-stream attentive CNNs for image retrieval," in *Proc. ACM Multimedia Conf. (MM)*. New York, NY, USA: ACM, 2017, pp. 1513–1521. doi: 10.1145/3123266.3123396.

[2] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1224–1244, May 2017.

[3] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1276–1288, Jun. 2012.

[4] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[5] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[6] J. Faruque, C. F. Beaulieu, J. Rosenberg, D. L. Rubin, D. Yao, and S. Napel, "Content-based image retrieval in radiology: Analysis of variability in human perception of similarity," *J. Med. Imag.*, vol. 2, no. 2, 2015, Art. no. 025501.

[7] S. Wei, Y. Zhao, Z. Zhu, and N. Liu, "Multimodal fusion for video search reranking," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 8, pp. 1191–1199, Aug. 2010.

[8] S. Wei, Y. Zhao, C. Zhu, C. Xu, and Z. Zhu, "Frame fusion for video copy detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 1, pp. 15–28, Jan. 2011.

[9] L. Duan, W. Ma, J. Miao, and X. Zhang, "Visual saliency based bag of phrases for image retrival," in *Proc. ACM Siggraph Int. Conf.*, 2014, pp. 243–246.

[10] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[11] S. Wei, D. Xu, X. Li, and Y. Zhao, "Joint optimization toward effective and efficient image search," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2216–2227, Dec. 2013.

[12] L. Liao, S. Wei, Y. Zhao, and G. Gu, "Improving the similarity estimation via score distribution," in *Proc. ICME*, Jul. 2016, pp. 1–6.

[13] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 1–13, 2016.

[14] N. Murray, H. Jégou, F. Perronnin, and A. Zisserman, "Interferences in match kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1797–1810, Sep. 2017.

[15] M. Jain, H. Jégou, and P. Gros, "Asymmetric Hamming embedding: taking the best of our bits for large scale image search," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1441–1444.

[16] A. B. Yandex and V. Lempitsky, "Efficient indexing of billion-scale datasets of deep descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2055–2063.

[17] Q. Dai, J. Li, J. Wang, and Y.-G. Jiang, "Binary optimized hashing," in *Proc. ACM Multimedia Conf.*, 2016, pp. 1247–1256.

[18] R. Liu, Y. Zhao, S. Wei, and Y. Yang. (2015). "Indexing of CNN features for large scale image search." [Online]. Available: https://arxiv.org/abs/1508.00217

[19] W. Zhou, H. Li, J. Sun, and Q. Tian, "Collaborative index embedding for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1154–1166, May 2017.

[20] J. Johnson, M. Douze, and H. Jégou. (2017). "Billion-scale similarity search with GPUs." [Online]. Available: https://arxiv.org/abs/1702.08734

[21] X. Wang, T. Zhang, G.-J. Qi, J. Tang, and J. Wang, "Supervised quantization for similarity search," in *Proc. CVPR*, Jun. 2016, pp. 2018–2026.

[22] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. CVPR*, Jun. 2015, pp. 37–45.

[23] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. CVPR*, Jun. 2012, pp. 2074–2081.

[24] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[25] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards large-scale histopathological image analysis: Hashing-based image retrieval," *IEEE Trans. Med. Imag.*, vol. 34, no. 2, pp. 496–506, Feb. 2015.

[26] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. CVPR*, Jun. 2010, pp. 3424–3431.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[29] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2014.

[31] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. CVPR Workshops*, 2014, pp. 512–519.

[32] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 392–407.

[33] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1269–1277.

[34] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," *Comput. Sci.*, 2015.

[35] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 584–599.

[36] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. CVPR*, Jun. 2016, pp. 598–606.

[37] S. Acharya and M. R. V. Devi, "Image retrieval based on visual attention model," *Procedia Eng.*, vol. 30, pp. 542–545, Mar. 2012.

[38] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun./Jul. 2004, p. 2.

[39] Z. Wen, J. Gao, R. Luo, and H. Wu, "Image retrieval based on saliency attention," in *Foundations of Intelligent Systems*. Berlin, Germany: Springer, 2014, pp. 177–188. doi: 10.1007/978-3-642-54924-3_17.

[40] E. Giouvanakis and C. Kotropoulos, "Saliency map driven image retrieval combining the bag-of-words model and PLSA," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 280–285.

[41] A. Papushoy and A. G. Bors, "Image retrieval based on query by saliency content," *Digit. Signal Process.*, vol. 36, pp. 156–173, Jan. 2015.

[42] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2007, pp. 545–552.

[43] S. Wan, P. Jin, and L. Yue, "An approach for image retrieval based on visual saliency," in *Proc. Int. Conf. Image Anal. Signal Process.*, 2009, pp. 172–175.

[44] G.-H. Liu, J.-Y. Yang, and Z. Y. Li, "Content-based image retrieval using computational visual attention model," *Pattern Recognit.*, vol. 48, no. 8, pp. 2554–2566, 2015.

[45] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.

[46] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Jun. 2015.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[48] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2016.

[49] Y. Wei *et al.*, "Learning to segment with image-level annotations," *Pattern Recognit.*, vol. 59, pp. 234–244, Nov. 2015.

[50] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1568–1576.

[51] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 685–701.

[52] T. Hoang, T.-T. Do, D.-K. Le Tan, and N.-M. Cheung, "Selective deep convolutional features for image retrieval," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1600–1608.

[53] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 241–257.

[54] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, 2017.

[55] M. Paulin, J. Mairal, M. Douze, Z. Harchaoui, F. Perronnin, and C. Schmid, "Convolutional patch representations for image retrieval: An unsupervised approach," in *Proc. IJCV*, 2016, pp. 1–20.

[56] W.-L. Ku, H.-C. Chou, and W.-H. Peng, "Discriminatively-learned global image representation using CNN as a local feature extractor for image retrieval," in *Proc. VCIP*, Dec. 2015, pp. 1–4.

[57] F. Radenović, G. Tolias, and O. Chum. (2016). "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples." [Online]. Available: https://arxiv.org/abs/1604.02426

[58] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian. (2016). "Good practice in CNN feature transfer." [Online]. Available: https://arxiv.org/abs/1604.00133

[59] W. Yu, K. Yang, H. Yao, X. Sun, and P. Xu, "Exploiting the complementary strengths of multi-layer CNN features for image retrieval," *Neurocomputing*, vol. 237, pp. 235–241, May 2016.

[60] Y. Li, H. Su, C. R. Qi, N. Fish, D. Cohen-Or, and L. J. Guibas, "Joint embeddings of shapes and images via CNN image purification," *ACM Trans. Graph.*, vol. 34, no. 6, 2015, Art. no. 234.

[61] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," *Assoc. Adv. Artif. Intell.*, vol. 1, no. 1, pp. 2156–2162, 2014.

[62] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. CVPR*, Jun. 2015, pp. 3270–3278.

[63] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. CVPR*, Jun. 2015, pp. 1556–1564.

[64] D. Zhou, X. Li, and Y.-J. Zhang, "A novel CNN-based match kernel for image retrieval," in *Proc. IEEE ICIP*, Sep. 2016, pp. 2445–2449.

[65] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 4353–4361.

[66] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, Jan. 2016.

[67] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk. (2016). "PN-Net: Conjoined triple deep network for learning local image descriptors." [Online]. Available: https://arxiv.org/abs/1601.05030
[68] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," in *Proc. CVPR*, Jun./Jul. 2016, pp. 267–275.
[69] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009
[70] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.
[71] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–17, 2007.
[72] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093
[73] 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. CVPR*, Jun. 2008, pp. 1–8.
[74] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, Jun. 2006, pp. 2161–2168.
[75] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
[76] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[77] A. Iscen, M. Rabbat, and T. Furon, "Efficient large-scale similarity search using matrix factorization," in *Proc. CVPR*, Jun. 2016, pp. 2073–2081.
[78] J. Li, L.-Y. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the secret of image saliency in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2428–2440, Dec. 2015.

Shikui Wei received the B.E. degree from Hebei University in 2003 and the Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), Beijing, China, in 2010. From 2010 to 2011, he was a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Full Professor with the Institute of Information Science, BJTU. His research interests include computer vision, image/video analysis and retrieval, and machine learning. More information can be found at http://mic.bjtu.edu.cn.

Lixin Liao received the B.E. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the Institute of Information Science. His research interests include image retrieval, machine learning, and video understanding.

Jia Li (M'12–SM'15) received the B.E. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University, he used to serve in Nanyang Technological University, Peking University, and Shanda Innovations. His research interests include computer vision and multimedia big data, especially the cognitive vision toward evolvable algorithms and models. He is the author or coauthor of over 60 technical articles in refereed journals and conferences, such as TPAMI, TIP, IJCV, ICCV, and CVPR. More information can be found at http://cvteam.net.

Qinjie Zheng received the master's degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2018. Her research interests include image retrieval and machine learning.

Fei Yang received the master's degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2018. Her research interests include image retrieval and machine learning.

Yao Zhao received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and a Full Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. He is also leading several national research projects of the 973 Program, the 863 Program, and the National Science Foundation of China. His research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010. He serves on the editorial boards of several international journals, including as an Area Editor for *Signal Processing: Image Communication* (Elsevier) and an Associate Editor for *Circuits, Systems and Signal Processing* (Springer).