Saliency Weighted Convolutional Features for Instance Search

Eva Mohedano*, Kevin McGuinness*, Xavier Giró-i-Nieto† and Noel E. O'Connor*
*Insight Centre for Data Analytics, Dublin City University, Ireland
†Image Processing Group, Universitat Politècnica de Catalunya, Spain
Email: eva.mohedano@insight-centre.org, kevin.mcguinness@insight-centre.org,
xavier.giro@upd.edu, noel.oconnor@insight-centre.org

***Abstract*—This work explores attention models to weight the contribution of local convolutional representations for the instance search task. We present a retrieval framework based on bags of local convolutional features (BLCF) that benefits from saliency weighting to build an efficient image representation. The use of human visual attention models (saliency) allows significant improvements in retrieval performance without the need to conduct region analysis or spatial verification, and without requiring any feature fine tuning. We investigate the impact of different saliency models, finding that higher performance on saliency benchmarks does not necessarily equate to improved performance when used in instance search tasks. The proposed approach outperforms the state-of-the-art on the challenging INSTRE benchmark by a large margin, and provides similar performance on the Oxford and Paris benchmarks compared to more complex methods that use off-the-shelf representations. Source code is publicly available at https://github.com/imatge-upc/salbow.**

***Index Terms*—Instance Retrieval, Convolutional Neural Networks, Bag of Words, Saliency weighting**

Fig. 1. Top 5 retrieved images for a query in INSTRE dataset using different spatial weighting schemes with BLCF. Correct results are highlighted in green, irrelevant images in red.

I. Introduction

Retrieval methods based on convolutional neural networks (CNN) have become increasingly popular in recent years. Off-the-shelf CNN representations extracted from convolution layers can be directly aggregated via spatial max/sum-pooling generating a global image representation that achieves a good precision and recall performance for a reasonable footprint [1], [2]. Moreover, when suitable training data is available, CNN models can be trained for similarity learning to adapt the learned representations to the end task of retrieval as well as to the target image domain (e.g. landmarks as in [3], [4]). However, the fine-tuning process involves the collection, annotation, and cleaning of a suitable training dataset, which is not always feasible. Furthermore, in generic instance search scenarios the target instances are unknown, which potentially makes off-the-shelf CNN representations more adequate for the task.

In this work we propose using state-of-the-art saliency models to weight the contribution of local convolutional features prior to aggregation. The notion of saliency in recent CNN-based image retrieval works has referred to the most active regions obtained from a specific filter within a CNN such as in [5], [6], or to a 2D feature saliency map derived from the convolutional layers [7]–[9]. In these cases, salient regions are detected with the same CNN from which image representations are extracted. On the other hand, human-based saliency models, such as those derived from the popular Itti and Koch model [10], have been applied only in image retrieval pipelines based in local handcrafted features [11]. To the best of our knowledge, it has not yet been explored how the most recent CNN-based retrieval pipelines can benefit from modern saliency models, and how different models affect the retrieval performance.

The contributions of this paper are the following:

- We propose a novel approach to instance search combining saliency weighting over off-the-shelf convolutional features which are aggregated using a large vocabulary with a bag of words model (BLCF).
- We demonstrate that this weighting scheme outperforms all other state of the art on the challenging INSTRE benchmark, without requiring any feature fine-tuning. Furthermore, it also offers comparable or better performance to more complicated encoding schemes for off-the-shelf features on Oxford and Paris benchmarks.
- We investigate the impact of the specific visual attention model used, and show that, perhaps surprisingly, higher

978-1-5386-7021-7/18/$31.00

performance on saliency benchmarks does not necessary equate to improved performance when used in the instance search task.

The remainder of the paper is organized as follows: Section II presents recent CNN-based retrieval methods, Section III introduces the different saliency models evaluated, Section IV describes the proposed retrieval pipeline, and finally, Section V presents the datasets and retrieval results obtained.

II. Related Work

While early approaches using CNN for retrieval were mainly focused in the usage of fully connected layers as global image representations [12], [13], most of the recent work is focused in exploring convolutional layers to derive those representations. Babenko [1] shown that a simple spatial sum pooling on a convolutional layer outperformed fully connected layers using pre-trained CNN models for retrieval for a reasonable footprint of 512D. Furthermore, the performance of this representations was enhanced by applying a simple Gaussian center prior scheme prior to aggregation on retrieval buildings dataset. Following a similar line, Kalantidis [8] proposed a cross dimensional weighting scheme (CroW) that consisted of the channel sparsity of a convolutional layer and a spatial weighting based in the L^2-norm of the local convolutional features. The proposed convolutional weighting was shown to outperform the centre prior scheme since the approach was able to better focus in the relevant instances.

Tolias *et al.* [14] focused in performing a region analysis in a convolutional layer prior the sum-pooling aggregation. Particularly, they proposed the regional of maximum activation (R-MAC) descriptor where different region vectors where extracted from a CNN layer (by max-pooling the activation of a particular area) following a multi-scale sliding window approach. The method has the advantage of roughly locating the instance of interest within an image. Jimenez *et al.* [6] explored the Class Activation Maps (CAMs) to improve over the fix region sampling strategy of R-MAC. CAMs generates a set of spatial maps highlighting the contribution of the areas within an image that are most relevant to classify an image as a particular class. In their approach, each of the maps is used to weight the convolutional features and generate a set class vectors, which are post-processed and aggregated as the region vectors in R-MAC. Some other recent works [7], [9] propose using a saliency maps which are derived directly from convolutional features. Lanskar and Kannala [7] use the saliency measure to weight the contribution of each R-MAC region prior to aggregation, while Simeoni *et al* [9] propose a method to detect a set of rectangular regions based on the derived saliency maps.

The current state-of-the-art retrieval approaches [3], [4] use models fine-tuned with a ranking loss. In particular, Gordo *et al.* [3] improves over the original R-MAC encoding by learning a region proposal network [15] for the task of landmark retrieval. Region proposal and feature learning are optimized end-to-end achieving excellent performance in the popular Paris, Oxford, and Holidays datasets. This approach, however, requires the construction of a suitable training dataset, which is usually domain specific, time consuming to construct and it is unclear how those models generalize in more generic scenarios such as the INSTRE dataset.

Intuitively, one can see that pooling descriptors from different regions and then subsequently pooling all of the resulting together is similar to applying a weighting scheme to the original convolutional features since region descriptors and globally pooled descriptors are built from the same set of local features. In this work, we propose a method to directly aggregate features from a convolutional layer by exploring a saliency weighting scheme as alternative to perform region analysis. Specifically, we utilize human-based saliency models as they were shown in the literature to be an effective weighting scheme on traditional SIFT-based BoW [16]–[18] and to the best of our knowledge they have not yet been investigated in combination with CNN representations in an instance search retrieval scenario. We demonstrate the generalization of our approach across different retrieval domains, resulting in a scalable retrieval system that does not require any additional feature fine-tuning.

III. Saliency models

Visual attention models detect regions that stand out from their surroundings, producing saliency maps that highlight the most prominent regions within an image. Saliency has been shown beneficial in content-based image retrieval when using traditional the Bag of Words model and SIFT features [11]. Some works have investigated the usage of saliency as a mechanism to reduce the number of local features and reduce the computational complexity of SIFT-based BoW frameworks [16]–[18]. Other works, rather than completely discard the background information, have used saliency to weight the contribution of the foreground and the background simultaneously [11], [19]. However, the usage of saliency models in the task of image retrieval has thus far been restricted to handcrafted CBIR features and handcrafted saliency models [10].

With the emergence of challenges such as the MIT saliency benchmark [20] and the Large-Scale Scene Understanding Challenge [21] (LSUN), and the appearance of large-scale public annotated datasets such as iSUN [22], SALICON [23], MIT300 [20], or CAT2000 [24], data-driven approaches based on CNN models trained end-to-end have become dominant, generating more accurate models each year. Having access to large datasets has not only allowed the development of deep learning based saliency models, but has also enabled a better comparison of the performance between models.

Our goals are to first evaluate whether saliency is useful in a CNN-based retrieval system, and second, to examine how the performance of different saliency models on saliency benchmarks relates to performance of instance search. To this end, we select two handcrafted models and four deep learning based models. The handcrafted models selected are the classic approach proposed by Itti and Koch [10], and the

Boolean Map based saliency (BMS) [25], which represents the best performing non-deep learning algorithm on the MIT300 benchmark. The four deep learning based models differ in CNN architecture, the loss functions used in training, and the training strategy [26]–[28], providing a reasonable cross section of the current state-of-the-art.

IV. Bags of Local Convolutional Features

Our pipeline extends the work proposed in [29], that consists of using the traditional bag of visual words encoding on local pre-trained convolutional features. The bags of local convolutional features (BLCF) has the advantage of generating high-dimensional and sparse representations. The high-dimensionality makes the aggregated local information more likely to be linearly separable, while relatively few non-zero elements means that they are efficient in terms of storage and computation. A maximum of 300 non-zero elements are required per image (174, 163 and 289 average words per image in practice in the Oxford, Paris, and INSTRE dataset respectively).

A convolutional layer generates a tensor of activations $\mathcal{X} \in \mathbb{R}^{H \times W \times D}$, where (H, W) is the spatial dimension of the feature maps and D the total number of feature maps. $\mathcal{X}_{ijk}$ refers to an activation located in the spatial location (i, j) in the feature map k. The volume of activations can be re-interpreted as $N = H \times W$ local descriptors $\mathbf{f}(i, j) \in \mathbb{R}^D$ arranged in a 2D space. A visual vocabulary is learned using k-means on the local CNN features, so each local feature can be mapped into an *assignment map*.

One of the main advantages of this representation is to preserve the spatial layout of the image, so it is possible to apply a spatial weighting scheme prior to constructing the final BoW representation. In particular, a saliency map is computed at the original image resolution, down-sampled to match the spatial resolution of the assignment map, and normalized to have values between 0 and 1. The final representation consists of an histogram $\mathbf{f}_{\text{BOW}} = (h_1, \ldots, h_K)$, where each component is the sum of the spatial weight assigned to a particular visual word $h_k = \sum_{i=1}^{W} \sum_{j=1}^{H} w_{ijk}$, being:

$$w_{ijk} = \begin{cases} \alpha(i,j) & \text{if } k = \underset{m}{\operatorname{argmin}} \parallel \mathbf{f}(i,j) - \mu_m \parallel \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

the spatial weight associated to a particular assignment, where $\alpha(i, j) \in \mathbb{R}$ is the normalized saliency map, and $\mu_m \in \mathbb{R}^D, m = \{1 \ldots K\}$ is one of the centroids of the visual vocabulary.

We use descriptors from *conv5_1* from a pre-trained VGG16 without fully connected layers. The VGG16 network was chosen following Jimenez *et al.* [6], where it was found that the features from this network outperform those from ResNet50 using the CAM approach. We confirmed this result in our own experiments, finding a similar result for BCLF encoding when using the local search scheme: there was a consistent reduction (> 10 points) in mAP when compared with VGG16 regardless of the weighting scheme used. Images are resized to have its maximum size of 340 pixels ($1/3$ of the original resolution) before performing mean subtraction prior to being forwarded through the network. This resolution is also used to extract the different saliency maps. Local features are post-processed with L^2-normalization, followed by a PCA whitening (512D), and L^2-normalized again prior to being clustered. Clustering is performed using approximate k-means with $k = 25,000$ words. The visual vocabulary and PCA models are fit on the target dataset. The obtained saliency maps are down-sampled by computing the maximum across 16×16 non-overlapped blocks.

For the query images, we interpolate the volume of activations to $(2H, 2W)$, as this interpolation of the feature maps has been shown to improve performance [29]. However, we only apply this strategy to the query images, and not to the entire dataset, so as not to increase dataset memory consumption. The bounding box of the queries is mapped to the assignment maps, and only the information relevant to the instance is used to create $\mathbf{f}_{\text{BOW}}$. This procedure is similar to the one used in [4], where a bounding box is mapped to the convolutional feature maps instead of cropping the original image, which was shown to provide better results.

V. Experiments

We experiment with three different datasets: the challenging INSTRE dataset for object retrieval [30], and two well-known landmark-related benchmarks, the Oxford [31] and Paris [32] datasets. The former dataset was specifically designed to evaluate instance retrieval approaches. It consists in 23,070 manually annotated images, including 200 different instances from diverse domains, including logos, 3D objects, buildings, and sculptures. Performance is evaluated using mean average precision (mAP) following the same standard protocols as in Oxford and Paris benchmarks, and evaluating mAP over 1,200 images as described in [33].

A. Weighting schemes

In this section we evaluate different spatial weighting schemes with the BLCF framework. We find that saliency weighting schemes (BMS, SalNet, and SalGAN) improve retrieval performance across the evaluated datasets with respect

TABLE I
Performance (mAP) of different spatial weighting schemes using the BLCF approach.

Weighting	INSTRE	Oxford	Paris
None	0.636	0.722	0.798
Gaussian	0.656	0.728	0.809
L^2-norm	0.674	0.740	**0.817**
Itti-Koch [10]	0.633	0.693	0.785
BMS [25]	0.688	0.729	0.806
SalNet [26]	0.688	0.746	0.814
SalGAN [27]	**0.698**	**0.746**	0.812
SAM-VGG [28]	0.688	0.686	0.785
SAM-ResNet [28]	0.688	0.673	0.780

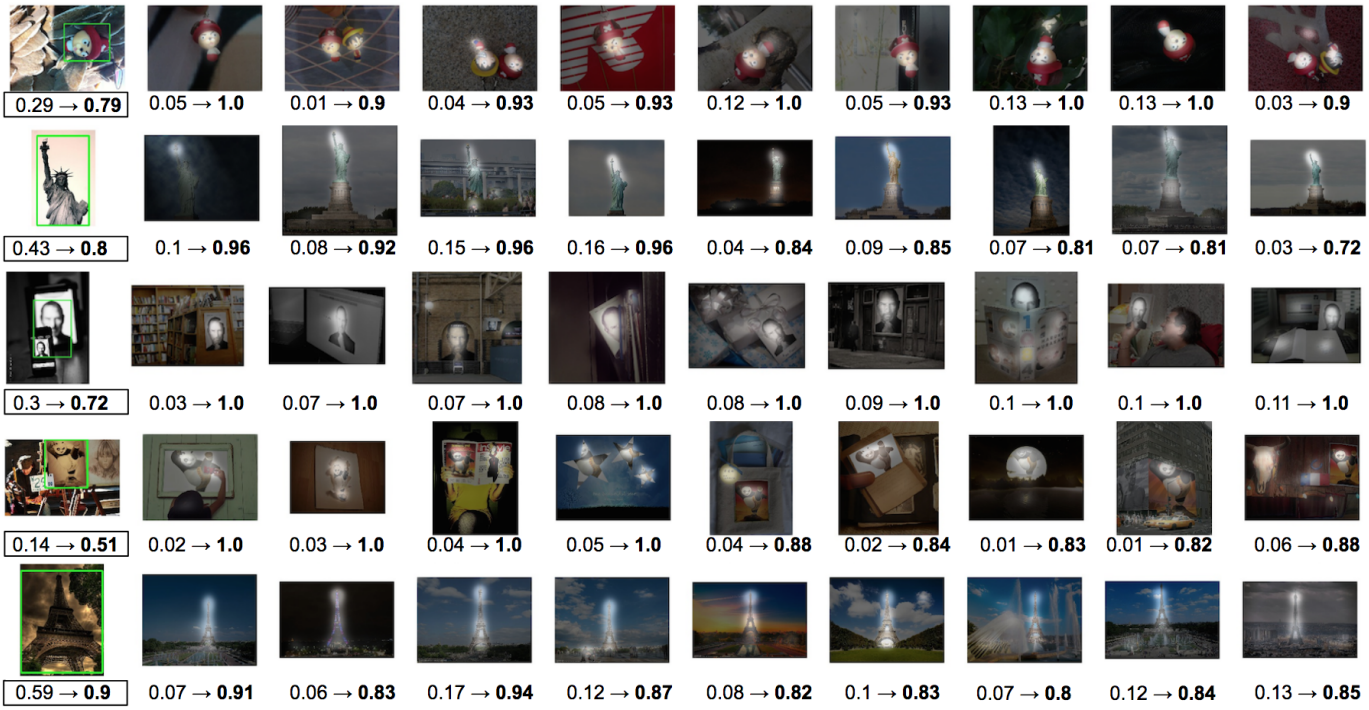

Fig. 2. Effect on performance for five examples from the INSTRE dataset after applying saliency weighting (SalGAN). The first column depicts the query with its associated average precision (AP). On the left, AP when performing unweighted BLCF, and on the right when performing saliency weighted BLCF. Retrieved images are ranked in decreasing order of ranking difference between unweighted BLCF and saliency weighted BLCF. For each image, precision at the position where each image is reported for unweighted BLCF (left) and saliency weighted BLCF (right).

to the unweighted assignment maps. Figure 2 shows some of the cases where saliency (SalGAN model) is most beneficial, allowing the efficient localisation of the target instances in most of the cases, despite of the high variability of the backgrounds and relative positions within the images in the INSTRE dataset.

Table I contains the performance of different weighting schemes on the BLCF approach. The simple Gaussian weighting achieves a boost in performance with respect to the baseline, which indicates a natural tendency of the instances to appear in the center of the images in all the datasets. The saliency measure derived from the convolutional features (L^2-norm weighting) allows to roughly localize the most relevant parts of the image, which represents a boost in performance with respect the Gaussian center prior weighting. However, the L^2 weighting appears in general to be much more noisy than the human attention-based saliency models.

Saliency models achieve the best performing results in the INSTRE dataset, with the exception of the Itti-Koch model, which decreases performance with respect the baseline in all datasets. This result is consistent with the quality of the saliency prediction achieved in the MIT300 benchmark, where it is rated in the second lowest rank of the evaluated models. The more accurate saliency models (BMS, SalNet, SalGAN, SAM-VGG, and SAM-ResNet), however, achieve almost equivalent performance on the INSTRE dataset. The relatively small size of the instances within this dataset and the high variance in their relative position within the images makes saliency a very efficient spatial weighting strategy for instance search, with no substantial difference between saliency models. This contrasts with results for the Oxford and Paris datasets, where a coarser saliency prediction (i.e the one provided by the SalNet model) achieves better results than the one obtained with the more accurate models of SAM-VGG and SAM-ResNet. Figure 3 illustrates this effect on three different instances of the Notre Dame cathedral from the Paris dataset. The most accurate saliency models (i.e. SAM-VGG and SAM-ResNet) detect saliency regions "within" the building, instead of detecting the full building as relevant. Also, SAM-VGG, SAM-ResNet, and SalGAN appear to be more sensitive to detecting people, omitting any other salient region of the image and thus decreasing the overall retrieval performance.

B. Aggregation methods

We also tested saliency weighting in combination with classic sum pooling over the spatial dimension to compare with the BCLF weighting. Here we report the best performing saliency models (SalNet and SalGAN) compared with a no weighting baseline using the VGG-16 network *pool5* layer at full image resolution. Although combining sum pooling with saliency weighting did give a performance improvement on the INSTRE dataset (mAP 0.527 for SalGAN vs 0.405 for no pooling), saliency weighting combined with BLCF gives substantially better mAP (see Table II). Furthermore, sum

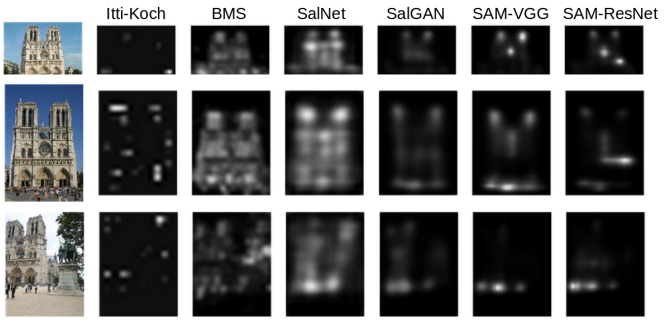

Fig. 3. Sample of saliency predictions for three examples of Notre Dame cathedral in the Paris dataset. Average precision (AP) for this query is 0.868, which is improved by BMS and SalNet models (achieving 0.880 and 0.874 AP respectively). More accurate saliency models decrease performance with respect the baseline, achieving an AP of 0.862 in the case of SalGAN, 0.857 SAM-VGG, 0.856 SAM-ResNet and 0.853 Itti-Koch models.

TABLE II
PERFORMANCE (MAP) OF DIFFERENT SPATIAL WEIGHTING SCHEMES USING THE SUM POOLING AGGREGATION ON (SUM) AND THE BLCF APPROACH.

Method	INSTRE		Oxford		Paris	
Weighting	SUM	BLCF	SUM	BLCF	SUM	BLCF
None	0.405	0.636	0.686	0.722	0.765	0.798
SalNet	0.519	0.688	0.681	0.746	0.766	0.814
SalGAN	0.527	0.698	0.612	0.746	0.749	0.812

pooling with saliency weighting gave little or no improvement on the Oxford and Paris datasets (mAP Oxford: 0.681 for SalNet vs 0.686 for no weighting, and Paris: 0.766 for SalNet vs 0.765 for no weighting). This result is perhaps unsurprising. Weighting the feature assignments prior to pooling in the BCLF framework can be interpreted as a simple importance weighting on each discrete feature. However, the interpretation for sum pooling is less clear, since the feature vectors at each spatial location usually have multiple activations of varying magnitude, and weighting alters the potentially semantically meaningful magnitudes of the activations. The BCLF-saliency approach also has the advantage of having two to three times fewer non-zero elements in the representation, giving substantially improved query times.

C. Comparison with the state-of-the-art

Our approach compares favorably with other methods exploiting pre-trained convolutional features, as shown in Table III. In particular, we achieve state-of-the-art performance using pre-trained features in Oxford and INSTRE datasets, when combining BLCF with saliency prediction from SalGAN. CAMs [6], and R-MAC [14] slightly outperform our approach in the Paris dataset, where they achieve 0.855, and 0.835 mAP respectively, whereas our method achieves 0.812 mAP. It is worth noting, however, that our system is much less complex (we do not perform region analysis), and more scalable, since the sparse representation only needs a maximum of 336 non-zero elements out of the 25K dimensions.

TABLE III
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART.

Method	Off-the-shelf	dim	INSTRE	Oxford	Paris
CroW [8]	yes	512	0.416	0.698	0.797
CAM [6]*	yes	512	0.325	0.736	0.855
R-MAC [14]	yes	512	0.523	0.691	0.835
R-MAC [4]†	No	512	0.477	0.777	0.841
R-MAC-ResNet [3]†	No	2048	0.626	**0.839**	**0.938**
(our) BLCF	yes	336	0.636	0.722	0.798
(our) BLCF-Gaussian	yes	336	0.656	0.728	0.809
(our) BLCF-SalGAN	yes	336	**0.698**	0.746	0.812

Results marked with (*) are provided by the authors. Those marked with (†) are reported in Iscen et al. [33]. Otherwise they are based on our own implementation.

TABLE IV
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART WITH AVERAGE QUERY EXPANSION.

Method	Off-the-shelf	dim	INSTRE	Oxford	Paris
CroW [8]	yes	512	0.613	0.741	0.855
CAM [6]*	yes	512		0.760	0.873
69 R-MAC [14]	yes	512	0.706	0.770	0.884
R-MAC [4]†	No	512	0.573	0.854	0.884
R-MAC-ResNet [3]†	No	2048	0.705	**0.896**	**0.953**
(ours) BLCF	yes	336	0.679	0.751	0.788
(ours) BLCF-Gaussian	yes	336	0.731	0.778	0.838
(ours) BLCF-SalGAN	yes	336	**0.757**	0.778	0.830

Results marked with (*) are provided by the original publications. Those marked with (†) are reported in Iscen et al. [33]. Otherwise they are based on our own implementation.

Fine-tuned CNN models [3], [4] significantly outperform our approach in Oxford and Paris. However, saliency weighted BLCF achieves state-of-the art performance in the INSTRE dataset with 0.698 mAP, outperforming the fined-tuned models. This is probably a consequence of restricting the training domain to landmarks images in [3], [4], since R-MAC with pre-trained features [14] (0.523 mAP) already achieves a higher performance in INSTRE than the fine-tuned R-MAC [3] (0.477 mAP). This provides evidence that similarity learning is a strategy to greatly improve performance on CNN representations where the target instance domain is restricted. In more generic scenarios, such as the INSTRE dataset, pre-trained ImageNet models achieve more general image representations.

One common technique to improve retrieval results is the average query expansion (AQE) [34]. Given a list of retrieved images, a new query is generated by sum aggregating the image representations of the top N images for a particular query. We select the top 10 images, and L^2-normalization on the new query representation. Table IV With this simple post processing we substantially improve the results in the INSTRE dataset (mAP increased from 0.698 to 0.757), achieving the best performance in comparison with other methods using the same post-processing strategy.

VI. CONCLUSIONS AND FUTURE WORK

We have proposed a generic method for instance search that relies on BoW encoding of local convolutional features

and attention models. We have demonstrated that saliency models are useful for the instance search task using a recently proposed CNN-based retrieval framework. Results indicate that different state-of-the-art saliency prediction models are equally beneficial for this task in the challenging INSTRE dataset. In landmark related datasets such as Oxford an Paris, however, a coarse saliency is more beneficial than highly accurate saliency models such as SAM-VGG or SAM-ResNet.

Better query expansion strategies can be applied to further improve results, such as the diffusion strategy proposed in [33], where global diffusion using $2,048$ dimensional descriptors from fine tuned R-MAC [3] achieves 0.805 mAP in INSTRE, and a regional diffusion 0.896 mAP. This strategy can be also applied to the proposed saliency weighted BLCF representation, potentially increasing retrieval performance of the proposed saliency weighted BLCF representations.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289 and SFI/15/SIRG/3283, and partially supported by the Spanish Ministry of Economy and Competitivity and the European Regional Development Fund (ERDF) under contract TEC2016-75976-R.

REFERENCES

[1] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1269–1277.

[2] A. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251–258, 2016.

[3] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 2017.

[4] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *European Conference on Computer Vision*. Springer International Publishing, 2016, pp. 3–20.

[5] J. Cao, L. Liu, P. Wang, Z. Huang, C. Shen, and H. T. Shen, "Where to focus: Query adaptive matching for instance retrieval using convolutional feature maps," *arXiv preprint arXiv:1606.06811*, 2016.

[6] A. Jimenez, J. Alvarez, and X. Giro-i Nieto, "Class-weighted convolutional features for visual instance search," in *28th British Machine Vision Conference (BMVC)*, September 2017.

[7] Z. Laskar and J. Kannala, "Context aware query image representation for particular object retrieval," in *Scandinavian Conference on Image Analysis*. Springer, 2017, pp. 88–99.

[8] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," *CoRR*, vol. abs/1512.04065, 2015. [Online]. Available: http://arxiv.org/abs/1512.04065

[9] O. Siméoni, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Unsupervised deep object discovery for instance recognition," *arXiv preprint arXiv:1709.04725*, 2017.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[11] Y. Wu, H. Liu, J. Yuan, and Q. Zhang, "Is visual saliency useful for content-based image retrieval?" *Multimedia Tools and Applications*, pp. 1–24, 2017.

[12] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision–ECCV 2014*, 2014, pp. 584–599.

[13] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.

[14] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *International Conference on Learning Representations*, 2016.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017.

[16] D. Awad, V. Courboulay, and A. Revel, "Saliency filtering of sift detectors: Application to cbir," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2012, pp. 290–300.

[17] Z. Liang, H. Fu, Z. Chi, and D. Feng, "Salient-sift for image retrieval," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2010, pp. 62–71.

[18] S. Nakamoto and T. Toriu, "Combination way of local properties classifiers and saliency in bag-of-keypoints approach for generic object recognition," *International Journal of Computer Science and Network Security*, vol. 11, no. 1, pp. 35–42, 2011.

[19] R. de Carvalho Soares, I. da Silva, and D. Guliato, "Spatial locality weighting of features using saliency map with a bag-of-visual-words approach," in *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, vol. 1. IEEE, 2012, pp. 1070–1075.

[20] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark."

[21] Y. Zhang, F. Yu, S. Song, P. Xu, A. Seff, and J. Xiao, "Large-scale scene understanding challenge: Eye tracking saliency estimation."

[22] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," *arXiv preprint arXiv:1504.06755*, 2015.

[23] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1072–1080.

[24] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[25] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 153–160.

[26] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.

[27] J. Pan, C. Canton-Ferrer, K. McGuinness, N. O'Connor, J. Torres, E. Sayrol, and X. Giró i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *CoRR*, vol. abs/1701.01081, 2017. [Online]. Available: http://arxiv.org/abs/1701.01081

[28] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *arXiv preprint arXiv:1611.09571*, 2016.

[29] E. Mohedano, K. McGuinness, N. O'Connor, A. Salvador, F. Marqués, and X. Giró-i Nieto, "Bags of local convolutional features for scalable instance search," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 327–331.

[30] S. Wang and S. Jiang, "INSTRE: A new benchmark for instance-level object retrieval and recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 3, pp. 37:1–37:21, Feb. 2015. [Online]. Available: http://doi.acm.org/10.1145/2700292

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[32] ——, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[33] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.