IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 4, 2021, accepted January 27, 2021, date of publication February 2, 2021, date of current version February 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056330

Colour Neural Descriptors for Instance Retrieval Using CNN Features and Colour Models

SURAJIT SAIKIA[1,2], LAURA FERNÁNDEZ-ROBLES[1,3], EDUARDO FIDALGO FERNÁNDEZ[1,2] AND ENRIQUE ALEGRE[1,2]

[1]Department of Electrical, Systems and Automation, University of León, 24071 León, Spain
[2]Researcher at INCIBE (Spanish National Cybersecurity Institute), 24005 León, Spain
[3]Department of Mechanical, Informatics and Aerospace Engineering, University of León, 24071 León, Spain

Corresponding author: Surajit Saikia (ssai@unileon.es)

This work was supported in part by the Junta de Castilla y Leon under Grant EDU/529/2017, and in part by the framework agreement between the University of Leon and INCIBE (Spanish National Cybersecurity Institute) under Addenda 22 and 01.

ABSTRACT Image representations in the form of neural activations derived from intermediate layers of deep neural networks are the state-of-the-art descriptors for instance based retrieval. However, the problem that persists consists of how to retrieve identical images as the most relevant ones from a large image or video corpus. In this work, we introduce colour neural descriptors that are made of convolutional neural networks (CNN) features obtained by combining different colour spaces and colour channels. In contrast to previous works, which rely on fine-tuning pre-trained networks, we compute the proposed descriptors based on the activations generated from a pretrained VGG-16 network without fine-tuning. Besides, we take advantage of an object detector to optimize our proposed instance retrieval architecture to generate features at both local and global scales. In addition, we introduce a stride based query expansion technique to retrieve objects from multi-view datasets. Finally, we experimentally proved that the proposed colour neural descriptors, obtain state-of-the-art results in Paris 6K, Revisiting-Paris 6k, INSTRE-M and COIL-100 datasets, with mAPs of 81.70, 82.02, 78.8 and 97.9, respectively.

INDEX TERMS Colour neural descriptors, CNN, image retrieval, image representation.

I. INTRODUCTION

In the last decade, significant progress has been made in the domain of computer vision and machine learning, especially in object detection, recognition and instance retrieval. The availability of massive amount of visual data in the form of images and videos have attracted many researchers to contribute new techniques and ideas in these domains. Specifically, we are interested in the problem of instance-level object retrieval from image datasets. Instance retrieval is a visual search task, that aims at retrieving all the images from a corpus of a large dataset that contain the same object instance as the query, see Fig. 1.

Instance-based retrieval systems have a large scope of potential applications for data-driven methods, such as secure retrieval in cloud environments [1], retrieving images with specific content [2], natural language description of images [3] and textile retrieval [4]. One important practical application where an instance retrieval system can play a major role is for crime scene evidence analysis [5]. To decipher a crime scene, the pieces of evidence in the form of images can be of great help for a forensic department. Most of the crime scenes are related to a specific location or place, and usually, there are several objects present or involved. For instance, when a crime has taken place in a specific location, looking for images related to that place and objects that can be found in those images may help in finding the location of the place and, ultimately, in solving the case. Also, a retrieval system similar to the one we are proposing can serve for GPS photo tagging, which associates an image with a specific location. Similarly, it could be possible to tag objects in images after identifying a queried object in them, which can be used for automatic database labelling and scene annotation. Looking in contributing to the solution of those problems, we propose a solution for instance retrieval using novel colour descriptors.

Nowadays, deep learning is being widely used in many computer vision applications and it obtains state of the art performance in various domains. Particularly, since Krizhevsky *et al.* [6] achieved the first place on the ImageNet

The associate editor coordinating the review of this manuscript and approving it for publication was Eyhab Al-Masri.

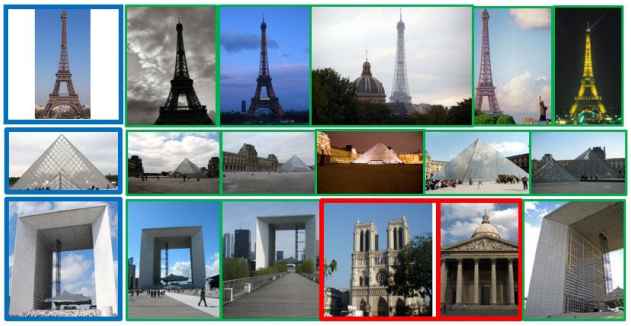

FIGURE 1. Three examples of image retrieval. For each row, the first image on the left represents the query image to look for in a dataset, in this case Paris 6k dataset. The rest of the images represent the hit list of retrieved images sorted according to a similarity metric to the query image. Images framed with a green rectangle correspond to correct retrievals whereas a red rectangle depicts the incorrect retrievals. These examples show the actual results of the proposed method.

classification and localization challenges in 2012, deep learning has been significantly applied to various problems, such as object detection [7], [8], image captioning [9], microscopic image analysis [10], analysis of skin marks [11], [12] and remote-sensing [13]. Moreover, the pre-trained CNNs can be used to generate high-level feature descriptors to represent visual objects in images. These activations generated at the intermediate layers of pre-trained CNNs could be used for instance search [14]–[16], in which given a query image, similar objects can be searched and retrieved from different images or videos. Region-based proposal methods [17], [18] based on CNNs, play a crucial role for object retrieval by easing up object localization process [19]. However, although much progress has been made, still significant issues and challenges persist related to query based instance retrieval. In real-world scenarios, the objects in images may appear partially occluded or in cluttered environments, which may lead to significant variations concerning viewpoints, scale and rotation. To address those challenges, most recent works focus on generating object proposals in images using end-to-end CNNs to learn the location of objects. But, the problem of retrieving the correct and the most similar ones persist as one of the main challenges in the instance retrieval domain.

To use an existing pre-trained CNN for instance retrieval, the network needs to be fine-tuned according to the type of images that are expected to be retrieved. The fine-tuning step is due to the fact that CNNs trained for image classification or recognition tend to encode the general information of a category, and they ignore the visual differences of instances belonging to the same category. For example, cars with different colours will be classified under the same category even if they are not visually similar. As a result, using those CNN features would not be viable for retrieval tasks. However, a fine-tuned network may not be effective for diverse retrieval tasks, in which the complexities of images are different from the ones that were used for training. The more challenging issue regarding the retrieval domain is to create a system that could retrieve images from a database even without fine-tuning. Therefore, to make a robust image retrieval system, the discriminative power of the features extracted from the CNNs needs to be increased by considering various factors.

For such retrieval tasks, the most crucial aspect is object localization that strongly depends on its appearance, and undoubtedly, the colour provides essential cues about object resemblance. The colour is the most basic and straightforward visual feature that represents the spectral content of images. Besides, colour based features are invariant to pixel translation or rotation in images. For query-based image search, to retrieve the relevant images, the query and the image features have to be nearly or completely similar. In fact, the same instance or object across different images may have similar outer appearances, which is represented by the colour. Using the colour as a feature, we can identify and discriminate between different images and the objects present in them. Since colour provides vital information on images, to increase the discriminative power of the neural features without fine-tuning, we propose to use different colour spaces and combinations of colour channels to transform the CNN features into robust descriptors. Therefore, our approach exploits the CNN features of a classification model by making the features discriminative by using colour models [20], but without applying fine-tuning.

Inspired by the success and recent breakthroughs of deep learning in the computer vision domain, we propose novel colour neural descriptors using Deep Convolutional neural networks Features (DCFs) or activations generated from a pre-trained CNN. Unlike most of the previous works that employ fine-tuning and require new training for image retrieval, we create new robust descriptors using several colour models without training and fine-tuning. We used the VGG-16 [21] pre-trained on the ImageNet [22] dataset to generate the neural activations from the last Fully Connected (FC8) layer. Those activations are generated concerning the three colour channels, Red (R), Green (G) and Blue (B), present in an RGB image. Instead of directly generating activations of an image, in our approach, we generate neural features for each of the colour channels -R, G and B- separately and we further pass them through a Colour neural Descriptor Generation (CDG) layer to construct the proposed colour neural descriptors.

In Fig. 2, we briefly illustrate our complete approach. To find a query instance present in an image, it is necessary to review the complete image, part by part, to check its presence. For this purpose, we employ Region Proposal Network (RPN) [17], which is an object detector and generates rectangular proposals of various sizes. To determine if the query region is present in the image, each proposal needs to be compared against the query image. Hence, we create colour neural descriptors for each of them and decided if they are similar using a distance metric. If the computed metric is higher than a given threshold, we retrieve that particular instance assuming that the query is present in that image. Moreover, under this instance retrieval scenario, if we are addressing datasets with multi-view or rotated objects, the retrieval task becomes even more challenging. Given that asymmetrical objects may

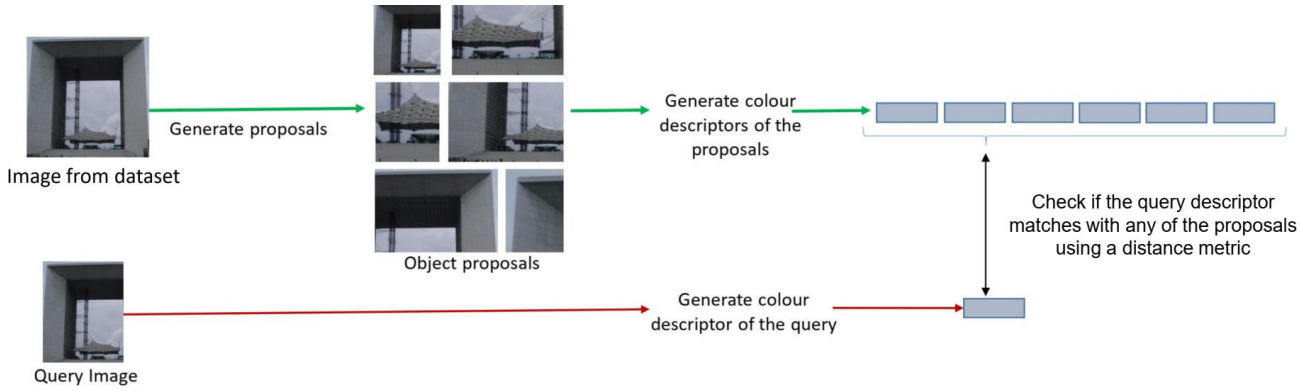

FIGURE 2. Overview of our proposed query-based instance retrieval framework using colour neural descriptors. The green line indicates the generation of descriptors for each of the proposals of an image from the dataset, and the red line represents the generation of the query descriptor.

appear rotated in images of the dataset, the method presented here would possibly fail to retrieve rotated views of the same object in which the appearance is remarkably different from the one of the query image. In order to mitigate this issue, we employ a query expansion technique.

We propose several configurations to obtain colour neural descriptors, and we compare them in terms of mAP and computation complexity. We evaluate our method on four benchmark datasets: COIL-100 [23], INSTRE-M [24], Paris 6K [25], and Revisiting-Paris 6K [26], and then we experimentally analyse the performance and distinctiveness of these descriptors.

To sum up, the main contributions of the paper are the following ones:

- We introduce new colour neural descriptors, based on the activations generated from a pre-trained Deep Convolutional Neural Network.
- We present a hybrid architecture composed of two different CNNs, and we use it successfully as the instance retrieval pipeline without employing fine-tuning techniques.
- We demonstrate experimentally that our proposed colour neural descriptors outperform the state-of-the-art in four datasets for image retrieval, COIL-100, INSTRE-M, Paris 6K and Revisiting-Paris 6k.

The rest of the paper is structured as follows. Section 2 briefly introduces the related works while Section 3 describes the method used to detect objects. The experiments and results are presented in Section 4, and finally, Section 5 discusses and 6 summarizes the conclusions and the future research lines derived from this work.

II. RELATED WORK

In this work, we are proposing discriminative colour neural descriptors obtained using CNNs, in order to improve retrieval in case of images in which colour is an important characteristic. Accordingly, in this section, we present the works related to our proposal divided in two topics: deep learning approaches and color descriptors.

A. DEEP LEARNING APPROACHES

Since the breakthrough of deep learning in the computer vision domain, the neural activations of a pre-trained network serves as a robust image descriptor. Several works [14], [27] used the neural activations extracted from the intermediate layers and achieved state-of-the-art results in instance retrieval tasks. Radenovic *et al*. [28] proposed to fine-tune CNNs for image retrieval by introducing trainable generalized-mean pooling layer that boosts the retrieval performance. Since the features obtained from CNNs demonstrated good performance, Simeoni *et al*. [29] proposed a method known as deep spatial matching for image retrieval which uses image descriptors extracted from convolutional neural network activations by global pooling. Similarly, Noh *et al*. [30] introduced Deep Local Feature (DELF), also based on CNNs which are trained with image-level annotations on a landmark dataset. Gordo *et al*. [31] presented a Siamese architecture that produces a global representation of images that is suitable for image retrieval. Recently, Wang *et al*. [32] introduced a deep cascaded neural network with deep representation for establishing multi-modal relationships for image retrieval tasks. For medical image retrieval, Y. Cai *et al*. [33] proposed a framework using CNN and supervised hashing, that adopts a Siamese network. Dubey *et al*. [34] proposed AlexNet descriptor for biomedical image retrieval, that is computed by fax-fusing RELU feature maps of a pretrained AlexNet, obtained from bit-plane decoded images. Recently, Maji *et al*. [16] proposed to use features derived from a CNN trained for a large image classification problem. Moreover, deep learning also plays a key role in remote sensing image retrieval. In [35], an autoencoder based framework was proposed to retrieve remote sensing aerial images using the encoded learned representation as a feature descriptor. Weixun *et al*. [36] investigated extracting deep CNN features for high-resolution remote sensing image retrieval using pretrained and a newly trained CNN. However, the main challenge is the unavailability of a large-scale dataset, and to address this issue, Weixun [37] introduced a large-scale remote sensing dataset known as PatternNet

which is suitable for training deep neural networks. Recently, Shao *et al*. [13] proposed a fully convolutional neural network for multi-label remote sensing image retrieval by extracting region convolutional features.

In addition, deep learning based object detectors can be leveraged for instance retrieval [19]. Salvador *et al*. [38] took advantage of the object proposals learned by an RPN [17] and their associated features to build an instance search pipeline. Furthermore, Mohedano *et al*. [39] explored local convolutional features for instance search task, and they built a retrieval framework based on those CNN features. Recently, Teichmann *et al*. [40] addressed query based image retrieval by extracting object regions and local features from images. This work introduced a regional aggregated selective match kernel (R-AMSK) to combine information from the detected regions to represent an image.

As the number of images has grown exponentially in the last few years, deep hashing plays a key role in making a retrieval system efficient in terms of computation and storage. With the development of deep learning-based approaches, various deep models are proposed to learn hash functions. Erin *et al*. [41] proposed two hash functions known as deep hashing and supervised deep hashing for learning binary codes. To preserve relative similarities between images, Lai *et al*. [42] presented a one-stage supervised hashing method using a deep architecture that generates pairwise hash codes. In most of the deep hashing methods, during discretization, the key category-level information may get lost. In order to address this issue, Lu *et al*. [43] introduced a method known as ranking optimization discrete hashing (RODH), that directly generates discrete hash codes. For avoiding information loss, Ding *et al*. [44] proposed discriminative dual-stream deep hashing (DDDH). Recently, to learn more effective binary codes, in [45], a new hashing method termed as DeepFuzzy Hashing Network is proposed, and Chen *et al*. [46] introduced Deep learning Supervised Hashing (DLSH) that learns features and binary codes together.

In our work, we employ R-FCN [18], which is a region-based convolutional neural network and VGG-16 to create colour neural descriptors. Furthermore, we don't do fine-tuning like other approaches as mentioned in the literature. Instead, we use colour models to create discriminative object descriptors for our retrieval approach. Currently, most of the approaches require fine-tuning a network with the specific type of data that a retrieval framework wants to address, whereas our proposal aims to create a common solution addressing all kinds of retrieval tasks without fine-tuning.

B. COLOUR DESCRIPTORS

In the literature, several colour-based descriptors have been proposed for image retrieval with a focus on increasing the illumination invariance and discriminative power. The earlier approaches used appearance models such as RGB colors histogram [47], YCbCr regional histogram [48], RGB spatiogram [49], and also the combination of texture with color descriptors [50]. From a general perspective, they focused on increasing the illumination invariance and discriminative power of such descriptors. Van *et al*. [20] studied the invariance properties and the distinctiveness of colour descriptors based on SIFT and Histograms, in which, apart from object recognition, the descriptors can be used for content-based image retrieval (CBIR) systems to search for similar images. Pujari *et al*. [51] presented a framework which uses colour and shape features from Lab and HSV spaces to retrieve edge features, and the experiments carried out in the Corel dataset demonstrated the efficiency of the method. Alzu *et al*. [52] introduced an optimized image descriptor that combines colour histogram in HSV space with the rootSIFT [53] descriptors and outperformed many state-of-the-art methods. Cortes *et al*. [54] evaluated 11 image descriptors and concluded that combinations of Gabor descriptors and dominant colour neural descriptors provide better performance. Lately, some works propose to combine colour with other texture or shape descriptors. In this line, Ahmed *et al*. [55] used canny edge histogram combined with discrete wavelets on YCbCr colour images or, more recently, Sotoodeh *et al*. [56] presented two approaches to extract discriminative features for colour image retrieval, based on Radial Mean Local Binary Pattern.

It is well known that colour is an important component for distinguishing objects in some specific problems. We are aware that the neural descriptors obtained from pre-trained models are not discriminative enough when there are similar objects with different colours. Therefore, in this work, we particularly focus on enhancing the neural activations using the colour information to make the descriptors more discriminative in those situations. We take advantage of the object detection architecture of R-FCN to extract region proposals for instance search.

III. METHOD

In this section, we present our approach that enhances image retrieval using colour neural descriptors and bounding boxes predicted by an object detector. In particular, our method builds on top of the object proposals and activations generated from the detector of the R-FCN algorithm. We first explain the backbone of the architecture and then we introduce our proposal for instance based retrieval using colour neural descriptors. Next, we present the overall instance retrieval method based on a given query, and finally, we describe a stride-based query expansion technique to retrieve multi-view objects.

A. BACKBONE ARCHITECTURE

The objective behind using the two different pre-trained networks is to facilitate the local instance search and the creation of discriminative descriptors. For local instance search, we use R-FCN to generate proposals on the dataset images in order to compare the query instance against those proposals. In contrast, VGG-16 serves as a feature extractor for both the query image and the proposals. Both networks serve as

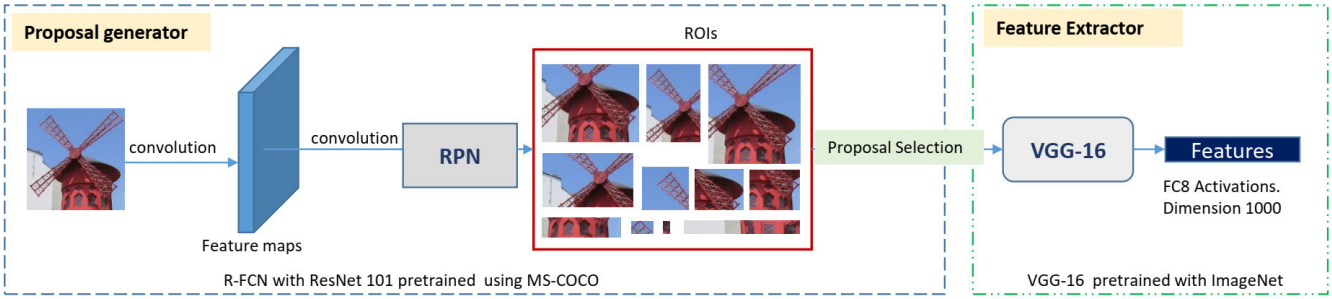

FIGURE 3. Backbone architecture of the instance retrieval approach. It is constituted in two parts: RPN from the R-FCN and the VGG-16 network for the feature extraction. The proposals are generated by the RPN, which are then given as an input to the VGG-16 net for region-based feature extraction.

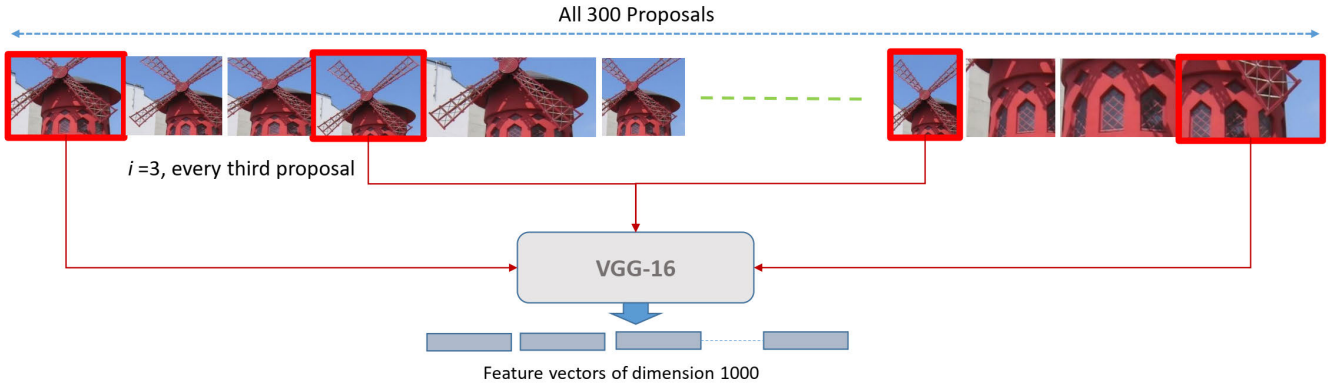

FIGURE 4. Detailed proposal selection. Every i^{th} proposals are selected out of the 300 proposals generated by the RPN to search the presence of query instance in a given image. In this case, we set $i = 3$ which results in 100 proposals per image. We next pass the proposals through the VGG-16 to obtain local features each of dimension 1000.

a single framework, where the candidate proposals generated by the R-FCN network are given directly as an input to the VGG-16 network to compute colour descriptors. In Fig. 3 we illustrate the architecture of the proposed method that constitutes the following two major stages.

- Generation of object proposals for regional search using R-FCN.
- Deep feature extraction using VGG-16 net.

1) OBJECT PROPOSAL GENERATION

To find a queried object or instance in an image, it is required to first detect and localize all possible objects for matching the query image features with each of the localized object features. The more similar the features are, the more likely the query and the proposals are from the same object.

In our approach, to search a query instance locally, we use the object detector of the R-FCN network to generate object proposals. R-FCN is faster than other region-based CNNs, such as Fast or Faster-RCNN [17], because it derives region proposals (ROIs) from the feature maps directly. In R-FCN, the RPN generates the object proposals using convolutional features maps, but unlike Fast and Faster RCNN, the fully connected layers after the ROI pooling are removed and hence no learnable layer is required after the ROI layer. As a result, R-FCN is up to twenty times faster than Faster R-CNN with a competitive mAP, and that is the reason we chose this architecture to generate region proposals. The total number of proposals obtained by the object detector is around 300, with lots of overlapping boxes covering the same object, what makes that during a query, it would be necessary to compare the same instance multiple times. Therefore, to reduce this cost, we define a set of candidate regions per image by selecting every i^{th} proposal, with i= 3 in this case, see Fig. 4. We store the proposal descriptors independently in a database to be used for our image retrieval system. The descriptors of each of the regions that the object detector select as a proposal are stored as well in the same database.

2) DEEP CNN FEATURES (DCFs) EXTRACTION

In order to create colour neural descriptors, we extract DCFs from the VGG-16 network pre-trained on the ImageNet dataset. We use the last fully connected layer (FC8) which contains 1000 neurons, resulting in a feature vector of 1000-D. In particular, the activations from the hidden layers represent low-level features, such as edges and contours, and the higher layers produce abstract features that fully represent images. Hence, we prefer to extract the DCFs at the penultimate layer. However, to generate colour neural descriptors, we extract the features corresponding to the three different colour channels (R, G and B). We represent the

DCFs obtained using R channel as R^*, G as G^* and B as B^*, which we will use to obtain the colour neural descriptors.

B. COLOUR NEURAL DESCRIPTORS

In this section, we first introduce the intuition behind the feasibility of colour neural descriptors, later, we explain how Deep Convolutional Features (DCFs) are generated using the colour channels, and finally, we present the proposed colour neural descriptors.

1) INTUITION BEHIND COLOUR NEURAL DESCRIPTORS

In some situations, the colour plays an essential role in obtaining visual information about objects present in images. Our idea is to leverage that information to create high-level discriminative colour feature vectors. An RGB image is composed of three channels, and the absence or presence of anyone would change the neural activations generated from an image.

For instance, in Fig. 5 we have a query representing a red box along with two other images: Image A is identical to the query and Image B differs only in the colour, which is yellow. First, for each image, we extract the DCFs of each colour channel, –R, G and B–, and then we concatenate them to create colour neural descriptors of the image. Next, we compute the similarity between the descriptor of the query image and the other two descriptors extracted from images A and B. Image A stands out in terms of similarity as compared to B, because the red-box on image A is similar to the query with respect to colour, and hence their respective colour neural descriptors are similar. Therefore, the channel-based activation of objects that have the same colours and textures are identical. As a result, colour neural descriptors made by deriving activation from individual channels and concatenating them are more robust. In this work, we present and evaluate different ways of fusing the DCFs obtained with respect to the colour channels to propose the colour neural descriptors.

2) COLOUR NEURAL DESCRIPTOR GENERATION LAYER

We define a Colour neural Descriptor Generation (CDG) layer, where we obtain colour neural descriptors from the DCFs extracted for each specific input colour channel passed to the network. In order to obtain a robust colour neural descriptor, we evaluated different colour spaces and combinations of colour channels, inspired by the work of [20]. Next, we present the different descriptors, and based on our preliminary tests, we chose the one that we consider more appropriate for the retrieval problem. Consequently, the CDG layer creates the colour models, which transforms the DCFs into colour neural descriptors.

a: NE-Raw.

The NEural Raw (*NE-Raw*) descriptor is generated by passing an image through the network without any modification of the input layer. We directly extract the activation with respect to the FC8 layer of the VGG-16 network without letting it pass through the CDG layer (we don't apply colour models). This descriptor posses no invariance to colour apart from the one conferred by the network. We use this descriptor mainly as a baseline, for comparison purposes against the other descriptors.

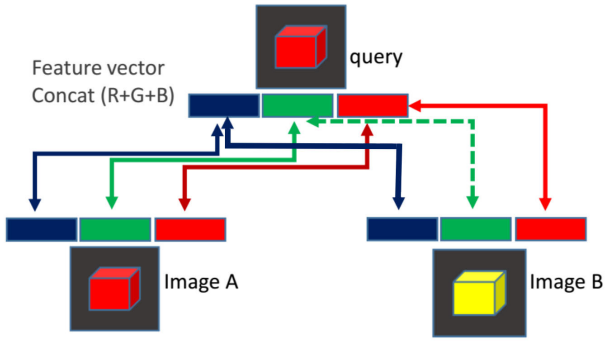

FIGURE 5. This figure gives a visual illustration of a simple case of colour neural descriptors representation and how it affects object description. Both the query image and image A contain a red box whereas image B contains a yellow box. The query image and Image A represent similar objects, and hence they have similar colour neural descriptors with respect to every colour channel. In the case of Image B, the green descriptor differs with respect to the query image. The weak resemblance between colour neural descriptors are represented by dashed lines whereas high resemblance is marked with solid lines.

b: NE-O and NE-O3.

NE-O represents the descriptor obtained using opponent colour space (Eq. 1), which is a combination of DCFs based on the channels of the opponent colour space. In the Eq. 1, the intensity information is represented by channel O3 and the colour information by O1 and O2. Due to the subtraction, the offsets become cancels out, and hence, the descriptor is invariant to changes in light intensity. The *NE-O* descriptor is constructed as the concatenation of O1, O2 and O3. Based on our preliminary tests, in some cases, results obtained with just O3 feature vector as colour neural descriptor outperformed combination of all the three components (O1, O2 and O3). We name this O3 feature vector as the *NE-O3* descriptor.

$$\begin{Bmatrix} O1 \\ O2 \\ O3 \end{Bmatrix} = \begin{Bmatrix} \frac{R^* - G^*}{\sqrt{2}} \\ \frac{R^* + G^* - 2B^*}{\sqrt{6}} \\ \frac{R^* + G^* + B^*}{\sqrt{3}} \end{Bmatrix}. \quad (1)$$

c: NE-TCD (Transformed Colour Distribution.)

In general, *NE-Raw* is not invariant to changes in lighting conditions. However, by normalizing the pixel value distributions (Eq. 2), shift invariance can be achieved with respect to changes in illumination. Since each channel is normalized independently, the descriptor is also robust to changes in colour intensity and arbitrary offsets. In (Eq. 2), μ_C is the mean and σ_C is the standard deviation of the colour distribution in channel C computed over the area under consideration

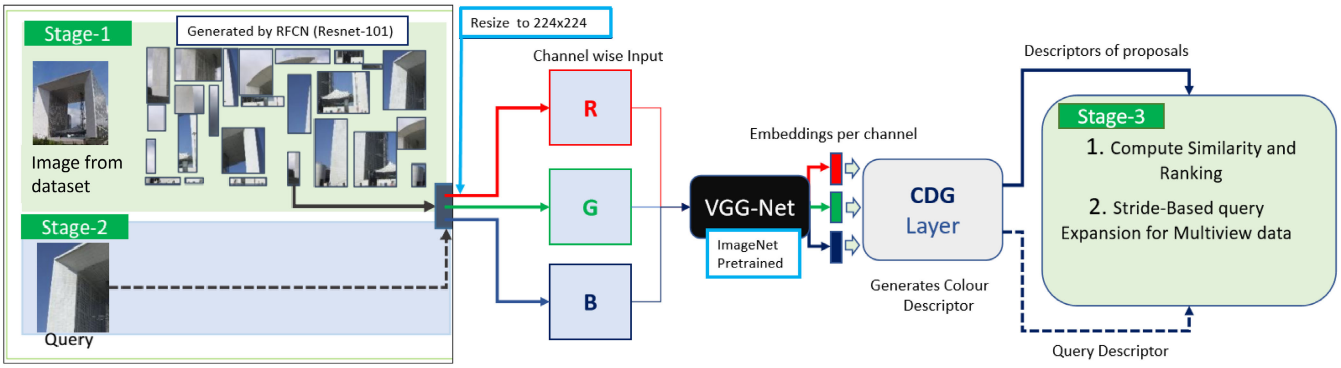

FIGURE 6. An illustration of object proposal generation for a query search. *Stage-1*: Generation of multiple object proposals using RPN. *Stage-2*: Extraction of query features to match with each of the proposals. *Stage-3*: Computing similarity and ranking instances, and for multiview data we employ query expansion technique.

(e.g. a patch or an image). This yields for every channel a distribution where $\mu = 0$ and $\sigma = 1$. At the CDG layer R', G' and B' are computed and concatenated to form *NE-TCD* colour neural descriptor.

$$\begin{Bmatrix} R' \\ G' \\ B' \end{Bmatrix} = \begin{Bmatrix} \frac{R^* - \mu_{R^*}}{\sigma_{R^*}} \\ \frac{G^* - \mu_{G^*}}{\sigma_{G^*}} \\ \frac{B^* - \mu_{B^*}}{\sigma_{B^*}} \end{Bmatrix}. \tag{2}$$

d: ***NE-C.***

We created this descriptor by passing the DCFs R^*, G^* and B^* of the three colour channels through the CDG layer. The resultant descriptor is a concatenation of the neural features corresponding to those colour channels, i.e $R^* + G^* + B^*$.

C. INSTANCE SEARCH AND RETRIEVAL

In this section, we present the proposed instance retrieval method based on colour neural descriptors. Fig. 6 illustrates the three-stage pipeline of our approach: (1) Dataset feature extraction, (2) Query feature extraction, and (3) Retrieving and ranking the top-K instances based on a similarity score.

1) DATASET FEATURE EXTRACTION

First, we process all the images in the dataset to calculate the descriptors, which are necessary for retrieving the images with objects similar to the queried one. Let $H = [H_1, H_2, \ldots.., H_n]$ be the set of images, we process each image H_j and generate M region proposals for each of them. The number of proposals depends on the proposal selection criteria as mentioned in section III-A1, where we select every i^{th} proposal to reduce the computation complexity. Then, we resize the proposals to 224×224 pixels and we extract DCFs with respect to those regions as mentioned in section III-A2. Next, we pass them to the CDG layer to create the colour neural descriptors as explained in section III-B2. In Fig. 6, *stage-1* illustrates how the image proposals are extracted from the dataset.

2) QUERY FEATURE EXTRACTION

Given a query instance H_q, the DCFs for each colour channel are extracted, and the colour neural descriptors are obtained as explained in section III-A2. In Fig. 6, *stage-2* shows the query feature extraction process.

3) RETRIEVING AND RANKING USING COSINE SIMILARITY

We aim at retrieving the images in the dataset that are the most similar to the query instance, sorting the retrieved list in descending order. First, we compute the similarity between the query instance H_q and the proposals of all images H of the dataset. Then, we create a hit list by sorting the images of the dataset in descending order, considering the similarity of every image as the highest similarity of any of its proposals and discarding images whose similarity is lower than an established threshold (Eq. 3).

$$S(H_q, m_i) = \begin{cases} > 0.75, & \text{retrieve } H_i. \\ else, & \text{discard.} \end{cases} \tag{3}$$

In order to retrieve only images with a high probability of being similar to the query, we determined experimentally the selected threshold, t. We used a sample set of images from the Outex dataset [57], which is an image retrieval dataset containing texture patterns. We evaluated four different values, $t = [0.60, 0.75, 0.80, 0.90]$, with 10 queries and selecting $t = 0.75$ because it was the value that returns consistently related images. Other values yielded a much more small or big number of retrievals what we considered less appropriate because a higher number of retrieved images increases the computational time to evaluate possible matches, and a lower value leaves out some potential candidates. We use the cosine similarity, see Eq. 4, to evaluate the similarity between the query image and each object proposal because it is one of the most commonly used metric for image retrieval. If the computed score, *CosSim*, is higher than the threshold t, then we include the proposal in the hit list.

$$CosSim(H_q, m_i) = \frac{\sum_{j=1}^{d} H_{q_j} m_{i_j}}{\sqrt{\sum_{j=1}^{d} H_{q_j}^2} \sqrt{\sum_{j=1}^{d} m_{i_j}^2}}, \tag{4}$$

TABLE 1. Dimensions of each of the descriptors.

NE-Raw	NE-C	NE-O3	NE-O	NE-TCD
1000	3000	1000	3000	3000

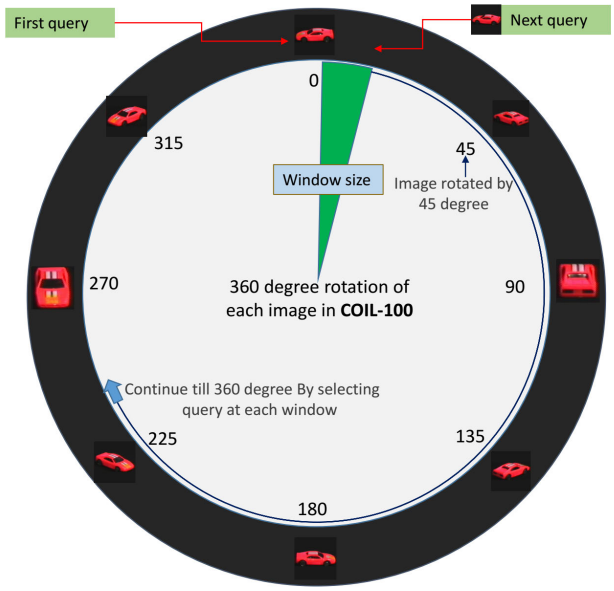

FIGURE 7. Query Expansion applied to COIL dataset. The red cars are rotated from 0 to 360 degrees with an interval of 5 degrees, and the car at degree 0 is the initial selected query.

where H_q is the query image, m_i is the i_{th} proposal, and d is the dimension of the colour neural descriptors, see Table 1.

D. STRIDE-BASED QUERY EXPANSION (SBQE) FOR MULTI-VIEW DATA

Multi-view datasets contain objects captured from various points of view, and hence it is difficult to retrieve all of them using a single query. We implement a query expansion technique to retrieve such multi-view objects in a cascading way. The pseudo-code is shown in Algorithm 1. As an example, let us take a query representing an object with 0-degree rotation, and a dataset of images containing the same object, but with different viewpoints produced by several degrees of rotation. Therefore, using the non rotated object as a query, probably we will only be capable of retrieving images with close rotations, around ± 45 degrees with respect to the original one, which correspond with rotations in the range $[45, 315]$ degrees. The rest of the images related to the query object presumably will be discarded due to high variations in their appearance caused by the rotation.

Hence, if we expand the query by considering, for example, the s^{th} image retrieved in the hit list, let us say the one rotated with 5 degrees as the next query image, then we could retrieve images with the object rotated from 310 to 50 degrees. The maximum number of images retrieved with respect to each query is based on the size of the stride s, and we will select the s^{th} image as the next query to retrieve the next subsequent images. We realised that the s^{th} image could be of any degree

Algorithm 1 Stride-Based Query Expansion (SBQE) to Retrieve Objects From Multi-View Datasets

Input: query image H_q, stride size s and K number of retrievals
Output: top-K instances

```
while length of list (L) < K do
    Extract colour neural descriptor of the query image Hq
    Compute CosSim (CS) score between colour neural descriptors of the query and dataset images
    Select images with CosSim score > 0.75 and sort them in terms of highest similarity with the query
    Append s number of images to a list L by removing the duplicates if present
    if lenght-of-list(L) ==K then
        return top-K instances;
    else
        Hq = last image in the list;
    end if
end while
return top-K instances
```

or even might not belong to the same class as the query image. When selecting the s^{th} image to be the next query, a false retrieval may have a negative cascading effect and we may end up retrieving undesirable images. In order to avoid that, we decided to use a small window with stride $s = 3$.

In Fig. 7, we present the algorithm with a visual explanation. Since we are going to work in COIL-100 with multi-view images, we illustrate how it works using an example taken from this dataset. In COIL, the images have a viewpoint ranging from 0 to 360 degrees, which makes difficult to retrieve all the related images with a single query. In Fig. 7, we can see several cars belonging to the same class. Let the car at 0 degrees be the initial query, and let us consider that we want to retrieve the top-K similar images. In this case, we could select the last instance from the retrieved list, the image rotated by 10 degrees in the window, to be the next query, and we will continue doing the same until the list of retrieved instances contains K images.

IV. EXPERIMENTS AND RESULTS

In this section, we present the experiments and the results obtained by evaluating our approach in four standard datasets.

A. DATASETS

We assessed our methodology using the following datasets:

COIL-100: Columbia Object Image Library (COIL-100) consists of 7,200 colour images of 100 objects class with

72 images per class. The dataset was created by placing objects in a motorized turn against a black background and were rotated from 0-360 degrees in intervals of five degrees to vary the object pose with respect to a fixed camera.

Paris 6K: This dataset consists of 6,412 still images of Paris landmarks or buildings collected from *Flickr*, which includes 55 query images of 11 buildings. Furthermore, it contains a diverse collection of class-specific images, where they differ in terms of illumination, viewpoint, size and resolution.

Revisiting Paris 6K: This dataset is an updated version of the Paris 6k dataset, which is published after correcting some of the annotations mistakes that were present in the original one. There are a total of 6332 database images and 70 query images. The database images contain the same images as present in the original Paris 6k dataset, but the query images are removed.

INSTRE-M: is an instance level object detection and retrieval dataset consisting of 5000 images of 50 classes with 101 images per class. It presents multiple appearances of the same object in each of the 101 images with respect to the class category, and hence it is very suitable for instance level retrieval. A sample of images from the three of them is presented in Fig. 12, where three different queries from each dataset and the top-10 related images retrieved are shown.

B. EVALUATION CRITERIA

We used standard evaluation protocols to evaluate our approach. We calculated the mean Average Precision (mAP) to measure the performance in all the experiments. First, we computed the average precision (AP), and then the APs for all the queries are averaged together to obtain the mAP. Eq. 5 defines AP, where $P(i)$ is the precision at the cut-off value i, N is the total number of retrieved images which are ranked according to their similarity scores, in this equation represented by K, and $IsRelevant(i)$ is an indicator function which equals 1 if the retrieved image at rank i is relevant, and 0 otherwise. N can refer to all retrieved images by the method, and thus it will take a different value for each query image, or it can be set to an established amount of retrieved images of a hit list of size K or top-K retrievals.

$$AP = \frac{\sum_{i=1}^{K}(P(i) \times IsRelevant(i))}{K} \tag{5}$$

Then, we calculated the mAP given by Eq. 6 where Q_N is the total number of queries.

$$mAP = \frac{\sum_{q=1}^{Q_N}(AP(q))}{Q_N} \tag{6}$$

C. EXPERIMENTAL SETUP

1) EXPERIMENTAL SETUP FOR PARIS 6k AND INSTRE DATASETS

For our experiments, we extracted 1000-D feature from the FC8 layer of the VGG-16 Network and we used the RPN to generate object proposals. All the experiments were done using TensorFlow (version-1.14.0) framework in an Nvidia Geforce GTX 1060 GPU machine with 16GB RAM and IntelCore processor (i7-7700HQ-2.80GHz). The programming language used for carrying out all the experiments is Python3.6 with CUDA support. For efficient storage of descriptors and faster retrieval, we used the HDF5 binary file format.

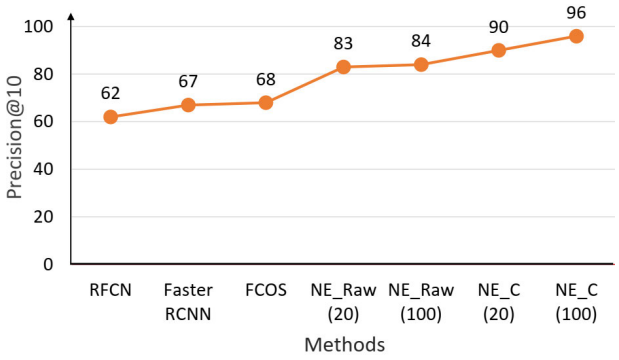

FIGURE 8. Comparison of various state of the art methods with our architecture in the Paris 6k dataset in terms of precision@10. The highest precision of 96 is obtained when we use our proposed architecture.

For determining the effectiveness of our proposed baseline architecture, we compared the performance considering a different number of proposals with the state-of-the-art Region-based CNNs: Fully Convolutional One-Stage object detection (FCOS) [58], Faster R-CNN with VGG-16 and R-FCN with ResNet. In FCOS, the proposal number varies, whereas, in FasterRCNN and RFCN, we extract features corresponding to the 300 proposals generated by them. We measured the performances in terms of *precision@10* given by Eq. 7, where R represents relevance, and is set to 1 if the i^{th} retrieved image contains the query image or 0 in another case. During the evaluation, we found out that the highest precision of 96 was obtained with *NE-C* with 100 proposals as can be seen in (Fig. 8) compared with other approaches. This demonstrates that the proposed architecture can achieve state-of-the-art results even with a lower number of proposals per image.

$$Precision@10 = \frac{\sum_{i=1}^{10} R(i)}{10} \tag{7}$$

After validating our architecture, we measured how *precision*@10 changes depending on the different number of proposals used, to know performance versus relative time trade-off. We define relative time in a range from 0 to 100, which comprises all the steps required, from the extraction of the descriptor up-to retrieval. The value 100 represents the maximum time taken by the descriptor. When the relative time of a descriptor is 50, it means that the descriptor is 2× faster. Whenever the number of proposals increases, the precision obtained is higher but it comes with a cost concerning the computation time. As illustrated in Fig. 9, while we consider 100 proposals per image we obtain an mAP of 96. If we reduce the proposals to 20, we obtain a precision of 90 but 5× faster. This experiment was done to

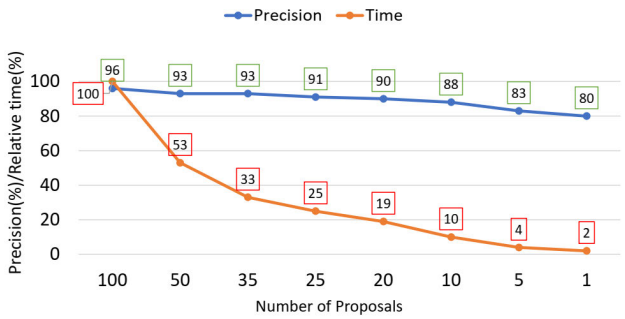

FIGURE 9. Precision@10 vs computational complexity with regard to the number of proposals considered per image in Paris 6k dataset with NE-C colour neural descriptor. Computational cost is shown in terms of relative time from 100 to 1 proposal.

illustrate that with a lower number of proposals we can have a faster retrieval framework when speed is the main concern. In order to carry out the rest of the experiments, we selected 100 proposals per image to ensure good performance at a reasonable computational cost.

2) EXPERIMENTAL SETUP FOR COIL-100

COIL-100 dataset contains multi-view images with single objects on a black background. Thus, in order to address instance retrieval in such dataset, we do not generate object proposals as we do for the Paris 6k and INSTRE dataset. We directly use the VGG-16 *FC8* features to create colour neural descriptors, and we employ the presented query expansion technique to retrieve instances. We used all the 7200 images as queries. Since the dataset consists of multi-view objects on a 360-degree turntable, we employed the query expansion technique to retrieve rotated views. Every image of the dataset contains a single object under a homogeneous background. For this reason, we directly extract the FC8 activation without generating proposals using RPN.

D. EXPERIMENTS AND RESULTS IN PARIS 6k DATASET

In the Paris 6k dataset, the queries are already provided with bounding boxes annotations in the dataset. Following the standard evaluation protocol for the Paris 6k dataset [25], we first cropped the 55 query images using the bounding boxes. We then extracted the query and the dataset features with respect to different colour neural descriptors, where the dataset features are stored in a database. To measure the effectiveness of the proposed descriptors, We first obtained their mAPs considering top-10 retrievals and then compared them. In Table. 2, we present the mAPs for top 10 ($K = 10$ in Eq. 6) retrievals achieved with the different proposed colour neural descriptors. The best performance is yielded using the *NE-C* descriptor with a mAP of **97.4** followed by *NE-TCD* and *NE-O3* with mAP of **96.9** and **95.02**, respectively. In order to compare with the other approaches, we have selected our best performing descriptor NE-C. In Table 3, we present the mAPs reported by various state-of-the-art approaches and compare with them. Among the earlier works,

TABLE 2. mAP for top-10 retrievals obtained with the baseline (*NE-Raw*) –shown in italics– and the proposed colour neural descriptors in Paris 6k dataset. The best result is shown in bold.

Proposed Descriptors	mAP
NE-C	**97.4**
NE-Raw	*92.2*
NE-O	89.4
NE-O3	95.02
NE-TCD	96.9

TABLE 3. Performance comparison with state-of-the-art methods for instance retrieval based on mAPs in the original Paris 6k dataset. We present the dimension of descriptors (dim) and mAP for all methods.

Method	dim	Fine-tuned	mAP
SPOC [59]	512	No	63.52
SPOC [59]	512	Yes	74.09
SPOC(ACK) [60]	256	yes	74.60
MAC [61]	512	No	67.02
MAC [61]	512	Yes	78.73
MAC(ACK) [60]	256	yes	75.69
RMAC [28]	512	No	72.02
RMAC [28]	512	yes	77.94
RMAC(ACK) [60]	256	yes	75.76
CROW [62]	512	No	68.94
CROW [62]	512	yes	77.48
CroW(ACK) [60]	256	yes	75.94
Gem [28]	512	yes	79.67
Gem(ACK) [60]	256	yes	76.26
NE-C(ours)	3000	No	**81.70**

the highest mAP reported was **79.67** by Gem [28]. Using our approach, we obtained an mAP of **81.70** using the NE-C descriptor and thus outperforming state-of-the-art results.

With these experiments, we demonstrate that the proposed colour neural descriptors are very efficient for content-based instance retrieval. Furthermore, due to the low performance of the *NE-O* descriptor, we discard it for the next sets of experiments. In Fig. 12, we show the top-10 retrieved instances for some query image examples using *NE-C*.

E. EXPERIMENTS AND RESULTS IN REVISITING PARIS 6k DATASET

To evaluate our proposal using the revisiting Paris 6k dataset, we followed the Medium-setup and the new evaluation protocol as explained in [26]. In Table 4, we compared the colour neural descriptor *NE-C* –which outperformed the rest– against some recent and relevant state-of-the-art approaches. Among the state-of-the-art methods, the highest mAP of **80.7** was obtained with DELF-GLD [40] method, in comparison to a mAP of **82.02** using *NE-C*. We also present the mean precision at 10 (mp@10), which is the mean of the precision for the top 10 retrievals as reported in the work [40]. While comparing, *NE-C* yielded **97.2**, being a bit lower than some other recent methods. We can notice that for a small number of retrievals –such as 10– the mean precision is saturated since most of the approaches are able to get a high result, however, achieving a high mAP is more challenging as the number of retrievals increases.

TABLE 4. Performance comparison with state-of-the-art methods for instance retrieval based on mAPs and mean precision at 10 (mp@10) in the revisiting-Paris 6K dataset. These methods were presented by [40]. In bold, the results of the proposed method and the best results of the state-of-the-art methods.

Method	mAP	mp@10
HesAff-rSIFT-ASMK+SP [63]	61.4	97.9
HesAff-rSIFT-ASMK [63]	61.2	97.9
ResNet101-R-MAC [64]	78.9	96.9
DELF-ASMK+SP [26], [30]	76.9	99.3
AlexNet-GeM [28]	58.0	91.6
HesAff-HardNet-ASMK+SP [65]	65.2	98.9
VGG16-GeM [28]	69.3	97.9
ResNet101-GeM [28]	77.2	98.1
ResNet101-GeM+DSM [29]	77.4	99.1
DELF-D2R-R-ASMK+SP [40]	78.2	**99.4**
DELF-GLD [40]	**80.7**	99.1
NE-C (ours)	**82.02**	**97.2**

TABLE 5. Performance comparison with state-of-the-art methods for instance retrieval based on mAPs in INSTRE dataset. We present the mAP for all methods.

Method	dim	mAP
CroW [62]	512	41.6
CAM [67]	512	32.5
R-MAC [68]	512	47.7
R-MAC-ResNet [31]	2048	62.6
BLCF [39]	336	63.6
BLCF-Gaussian [39]	336	63.6
BLCF-SalGAN [39]	336	69.8
Lin *et el*. [69]	1024	57.5
NE-C (ours)	3000	**78.8**
NE-TCD (ours)	3000	**77.5**
NE-O3 (ours)	1000	**70.21**

TABLE 6. mAP and per query search time (in seconds) for top-20 retrievals with respect to proposed colour neural descriptors and the Stride-Based Query Expansion (SBQE) in COIL dataset.

Proposed Descriptors	mAP	Time
NE-C	98.8	0.042 secs
NE-O3	96.0	0.022 secs
NE-TCD	98.7	0.043 secs
SBQE(NE-C)	**99.8**	0.77 secs

F. EXPERIMENTS AND RESULTS IN INSTRE DATASET

Next, we evaluated the retrieval performance in INSTRE-M, which constitutes a similar scenario to the Paris 6k dataset. To evaluate the performance, we computed mAP following the protocol described by [66], which uses 1250 query images. In Table 5 we present the achieved results, and it can observed that concatenation of channel-specific DCFs, *NE-C*, yielded the best performance with a mAP of **78.8** followed by *NE-TCD* with mAP **77.5**. In Fig. 12, we show the top-10 retrieved instances for some query examples using *NE-C*.

G. EXPERIMENTS AND RESULTS IN COIL-100

In the first experiment, we employed the method described in Section III with the proposed colour neural descriptors for top-20 retrievals to determine the best descriptor. Since we are addressing a different dataset, we compared all the descriptors again for K = 20 retrievals. Table. 6 presents the achieved results, where the highest mAP (**99.8**) was obtained by using *SBQE* (query expansion of NE-C) followed by *NE-C* and *NE-TCD* with mAPs **98.8** and **98.7**, respectively. Also, we registered and present the time required for per-query instance search with respect to each of the descriptors. Furthermore, in table. 7, we even compared the descriptors with some of the works that were reported in [70] for top-20 retrievals based on *precision*. We observe that all the descriptors achieve comparable mAPs, and we select the best one (*NE-C*) to compare with the state-of-the-art approaches.

a: RETRIEVING MULTI-VIEW INSTANCES

Additionally to the previous experiments, we evaluated a top-72 retrieval system. In this way, we are allowing the retrieval of all the 72 objects per class in the dataset to verify if all the rotated or multiviewed objects are retrieved correctly. We used *NE-C* descriptors with and without query expansion, and compared the results with the baseline *NE-Raw* descriptor. Fig. 10 shows the results obtained by plotting mAP versus K retrievals, where K goes from 1 to 72. The best results were obtained using *NE-C* with SBQE as compared to *NE-C* and *NE-Raw*. At $K = 72$, the mAP obtained using *NE-C* with SBQE is 98.3, whereas using *NE-C* and *NE-Raw* are 91.7 and 87.9, respectively. The performance is boosted while query expansion is applied to *NE-C*, but it has high computational cost as compared to *NE-C* without query expansion.

Based on the previous experiment, we compare *NE-C* and SBQE against the state-of-the-art methods, presenting the results in Table 8. The mAP obtained using *NE-C* and SBQE is **97.9**, which is superior to the best mAP of **95.4** as presented in [71]. The query expansion approach is computationally more expensive, but it is very useful when it is important to boost the performance of the colour neural descriptors for image retrieval in multi-view datasets.

In Fig. 11, we illustrate the retrieval of multi-view images. The initial query image corresponds to the orange mug rotated 315 degrees. Then, a window of stride 3 selects the next query, which in this case is rotated by 325 degrees. The bottom two rows show the top-11 retrieved images when the initial queries are positioned at 315 and 145 degrees, respectively. We set the size of the stride as $s = 3$ and we retrieved top-s similar images. The s_{th} retrieved instance in the hit list is set as the next query, and the process continues until K instances are retrieved. We chose the orange cup rotated 315 degrees as the initial query image. It can be observed that each of the top-s retrievals are in close vicinity to 315 degrees, and are slightly rotated (*clockwise or anticlockwise*) with respect to their queries.

In Fig. 12, we show the top-10 retrieved instances for some query examples of the COIL-dataset using *NE-C* descriptors.

V. DISCUSSION

In this work, we proposed colour neural descriptors for instance retrieval, and we evaluated them in four datasets

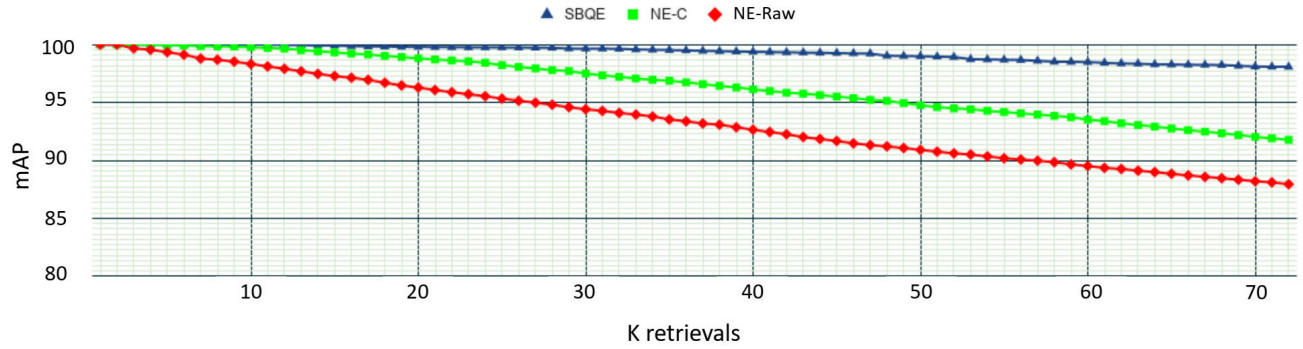

FIGURE 10. mAPs at top-*K* instance retrievals in COIL dataset using descriptors NE-Raw and NE-C (with and without query expansion), where SBQE represents NE-C with query expansion.

FIGURE 11. In this figure, we illustrate the retrieval of an image sample in COIL dataset using the proposed query expansion approach.

TABLE 7. Precision of classical image descriptors with 20 returns in Coil-100 dataset which were reported in [70].

Method	Precision
LBP	74.30
MSD	97.72
CDH	92.48
HSV	96.73
PUD	99.11
NE-C (ours)	**99.56**
SBQE(NE-C) (ours)	**99.99**

TABLE 8. Performance comparison with state-of-the-art methods for instance retrieval based on mAPs in COIL dataset. The best results are shown in bold, and the second best value is italisized.

Proposed Descriptors	mAP
Ahmed *et al.* [72]	93.0
BoVW [71]	78.5
txx [73]	61.5
fuzzy weights [74]	80.0
vwa [75]	86.0
BoCIDVW [71]	*95.4*
NE-C (ours)	**92.3**
SBQE (NE-C) (ours)	**97.9**

to assess its performance in terms of mAP. We wanted to provide a solution that employs colour models with deep convolutional neural networks. In situations in which it is needed to retrieve an object in datasets containing objects with very similar appearances, the colour becomes a very discriminative feature. To retrieve specific instances from a particular dataset, most of the proposed works based on deep learning methods to date adopt fine-tuning approaches. However, the raw neural activation obtained directly by passing an image through the CNN may not be feasible for similarity search. This is due to the fact that, under different illumination conditions, the appearance of most objects varies, and hence, we may have different neural activations for the same object.

FIGURE 12. Top-10 retrieved images of COIL-100 (rows 1-3), INSTRE-M (rows 4-6) and Paris 6K (rows 7-9) datasets. The above results were obtained by using the colour neural descriptors that achieved best results in each dataset. The queries are framed with blue rectangles, correctly retrieved images with green ones and incorrectly retrieved images with red rectangles.

As a result, we may not be able to retrieve all instances related to the query object, and as a consequence, other approaches opt for fine-tuning. Therefore, in order to make the descriptors more discriminative, we enhanced the DCFs by using colour models, and later on, we evaluated some combinations of the DCFs to obtain more discriminative feature vectors. The hypothesis behind our proposal for creating colour neural descriptors is that, if we separate the three channels of an image and consider CNN features specific to each of them, then the vector obtained by its combination is more discriminative than the one obtained by passing through the CNN the complete RGB image.

The main advantage of our proposed approach is the usage of a hybrid architecture in which we combined two state-of-the-art networks for instance search without explicitly training or fine-tuning. Besides, we expanded our solution looking for retrieving multi-view images by introducing a stride based query expansion technique. In most of the cases, to extract such instances, fine-tuning an algorithm specific to the dataset is the cue. This is because, when an object is rotated, for example in a turntable, the appearance may significantly vary, and as result, the neural activations of the object change as well. By applying the proposed stride-based query expansion, we were able to successfully retrieve such rotated images with high mAP, but with the drawback of increasing the computational time. Furthermore, the descriptor obtained proved to be competitive for instance retrieval, outperforming state-of-the-art results in four datasets: COIL-100, INSTRE-M, Paris 6K and Revisiting-Paris 6k. However, the feature extraction is complex since they are required to be extracted from three different channels. Nevertheless, the complexity and the retrieval time can be reduced significantly by parallelizing the extraction process in a GPU machine.

A. PERFORMANCE TRADE-OFFS AND TIME CONSUMPTION

We observed that the precision and mAPs obtained by our proposal are superior to the state-of-the-art approaches. However, apart from exhibiting high performance, it can be noticed that, depending on the chosen colour neural descriptor, there is a trade-off between mAP and computational cost. In Table 6, we can observe that *NE-O3* is approximately 1.3× faster than *NE-C* and *NE-TCD*, and 35× than *SBQE(NE-C)*, but it has the minimum mAP of 96 as compared to the highest

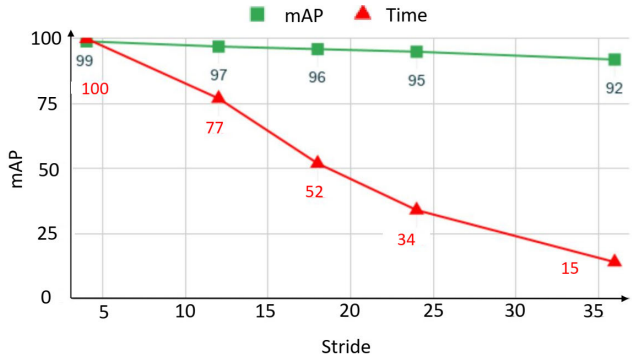

FIGURE 13. mAP vs relative time complexity of the stride base query expansion approach with respect to stride size for top-72 retrievals in COIL dataset using *NE-C* colour neural descriptor. Time complexity is shown in terms of relative time for a stride of 3.

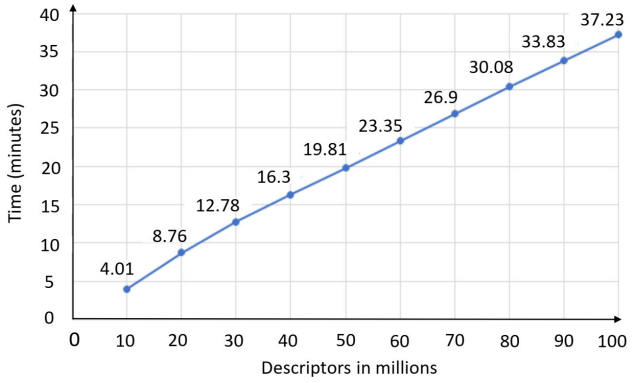

FIGURE 14. Computation time (in seconds) vs number of descriptors in millions with respect to similarity score computation between feature vector and database vectors.

mAP of 99.8 obtained by *SBQE(NE-C)*. In Fig. 13, we show the *relative time* vs *mAP* trade-offs of *NE-C* descriptor with query expansion. We define relative time in a range from 0 to 100, where 100 represents the maximum time taken by the descriptor.

In Fig. 13, we show relative time with respect to a stride of 3. We compute the results in COIL dataset using *NE-C* and *SBQE* approach for top-72 retrievals. As we had seen, for a stride of 3, the mAP is 99.8 and we consider it as the scenario with the maximum time, that is 100%. We can observe that as the size of the stride increases the computational complexity reduces significantly along with a slight decay in the mAP. Therefore, for instance, retrieval systems where time is a prime concern, they can increase the stride size up to 10 if an mAP close to 80 suffices, or they can use *NE-O3* or *NE-Raw* descriptors to speed up the approach. Whereas, *NE-C* is the one to select when precision is of utmost importance.

In general, the computational time required to process a single image for creating the colour neural descriptor is approximately 0.25 seconds, and for generating all the 100 proposals is around 0.20 seconds. We also observed that, as compared to *NE-Raw* descriptor, the colour neural descriptors are computationally expensive since a feature extraction of three individual channels is required. But, due to the availability of recently advanced parallel computing resources with powerful GPUs, we significantly reduced the descriptor creation time by simultaneously extracting deep features corresponding to three colour channels. As a consequence, the time required to compute *NE-Raw* and *NE-C* descriptors are approximately equal, which is 0.08 seconds.

B. SCALABILITY OF PROPOSED APPROACH

During the last few years, as the number of images has increased exponentially in the order of millions and billions, optimizing the matching and sorting tasks in databases of feature vectors in terms of time is critical for a quick instance retrieval. Besides, the storage of billions of feature vectors can result in the need for huge RAM memory. To deal with these two issues, we save the descriptors in HDF5 binary format, which allows storing vast amounts of numerical data in a single file.

In order to compute the retrieval time and to establish a time complexity order for big databases, we generated 100 million random vectors of dimension 3000 representing our colour descriptors *NE-C*. Since the memory of our RAM is limited with 16 GB, we created 1000 HDF5 files each one containing 10^5 vectors. To compute the similarities of a vector with respect to all 100 million vectors, we loaded an HDF5 file into the RAM, and once compared, we removed it and a newer one was loaded subsequently. In Fig. 14, we show the computational time with respect to the number of descriptors. The average computational time taken for loading 10^5 vectors and computing the similarity scores is approximately 2.3 seconds. Then, to compare with 10 million descriptors, the computational time was 241 seconds (about four minutes) and required 100 HDF5 files to be loaded and unloaded. At last, the computational time to compute the similarity scores of 100 million descriptors took just 37 minutes, where approximately 41000 comparisons are made per second, which is reasonably fast for many applications that do not require real-time performance. However, in the case of computers with different specification or feature vector dimensions, the number of HDF5 files and the vectors stored in each file could be adapted.

In addition, we can observe in Fig. 14 that the computation time is linearly dependent on the number of descriptors present in a database, and hence, we have a linear order of complexity $O(n)$. Once the similarity scores are obtained, we also need to sort the hit list to find the most relevant retrievals. We performed the sorting by means of the quick-sort algorithm, which has an average complexity of $O(nlogn)$. Thus, the overall complexity of instance search is given by $O(n) + O(nlogn)$. In fact, due to this inherent linearity, we can further scale up this approach for billions of descriptors by storing them in more HDF5 binary format files. Moreover, based on the observed experimental results in which the time varies almost linearly with respect to the number of descriptors analyzed, the computation time is expected to be approximately 373 minutes (6 hours) for

one billion descriptors comparison. Typically, in a recent high-end machine, say with 128 GB RAM, 10 million colour descriptors can easily fit in. For instance, if there are one billion descriptors, then 100 HDF5 batch files can be used to store 10 million descriptors each. Moreover, the loading and unloading overhead can be further reduced in a machine with high RAM, and thus, the instance retrieval can be speeded up. In this way, the retrieval step can be scaled.

Similarly, the feature extraction step can be also scaled up while maintaining a linear order of complexity. For n images on the dataset, there are n forward passes of the network, and in each pass, we obtain M (a constant) number of object proposals at the same time (as mentioned in section III-A1). As n grows more and more, the number of proposals generated for each image will become trivial as compared to n. Therefore, the order of complexity for feature extraction is $O(n)$, which is linear. However, since feature extraction using any CNN is computationally expensive while dealing with large scale datasets, it is ideal to run parallel instances of the feature extractor (CNN) in mutually exclusive image batches. In our case, for faster feature extraction, we selected a batch size of 200, and it takes approximately 1.85 seconds to extract features of 200 images and to save them as HDF5 files.

Moreover, the computation time can be further reduced if the descriptors are converted into binary codes using deep hashing, which has emerged as an important technique for image retrieval in large scale datasets. However, in this work, our prime focus was only to create colour descriptors for the instance retrieval without fine-tuning. Since we have obtained a new deep feature representation to define colour descriptor, we aimed at proving the full efficacy of this original representation without applying any techniques on top of it. Besides, we were able to do a fast retrieval where a query descriptor can be compared with 40,000 descriptors in less than just two seconds. In addition, we can efficiently store more than one million descriptors into the RAM using HDF5 binary format.

VI. CONCLUSION

In this work, we have presented colour neural descriptors for instance-based retrieval using CNN feature maps and colour models. First, we modified the input part of the network to generate DCFs for the different colour channels. Next, the extracted activations were passed through the colour neural descriptor Generation (CDG) layer to construct the colour neural descriptors. For developing a query-based retrieval system, we first created object proposals for each of the images in a given dataset, and then we calculated the corresponding colour neural descriptor for each proposal. After that, we computed the cosine similarity between the query descriptor and each object proposal descriptor, retrieving those images where the similarity scored higher than a specified threshold. Additionally, we introduced a stride-based-query-expansion technique, especially appropriate to retrieve images from a multi-view dataset. We selected an initial query to retrieve the top-K similar instances, and then we used a stride of size s to select the s^{th} retrieved instance to be used as the next subsequent query. In contrast to prior works, which relied on fine-tuning a network, we enhance the DCFs to increase the discriminative power concerning colour variations.

We evaluated the proposed method using standard protocols. We experimentally showed that our approach significantly boosts the retrieval performance in terms of precision without fine-tuning techniques applied. Besides, to address multi-view image retrieval, we use a query expansion technique based on stride. The descriptor obtained proves to be competitive for instance retrieval, outperforming state-of-the-art results in four datasets: COIL-100, INSTRE-M, Paris 6K and Revisiting-Paris 6k.

In the future, we will train a network to directly generate colour descriptors along with the proposals in order to decrease the feature extraction time, and also we would consider multi-label retrieval using the proposed colour descriptors. In addition, other colour models such as HSL, HSV, CMYK can also be used for obtaining discriminative colour descriptors by applying the same formulation. However, since we used a model previously trained with RGB images, the model can detect intrinsic features such as edge, shapes and other key points as long as the provided image consists of the RGB colour channels. Besides, if the colour space changes, the model will not be able to detect discriminative features since other colour spaces would have different channel values than the RGB scale. Due to this reason, we opted for RGB colour space, and we consider experimenting with different colour spaces in the future.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of Nvidia Corporation for their kind donation of GPUs (GeForce GTX Titan X and K-40) that were used in this work.

REFERENCES

[1] Y. Xu, X. Zhao, and J. Gong, "A large-scale secure image retrieval method in cloud environment," *IEEE Access*, vol. 7, pp. 160082–160090, 2019.
[2] P. Chhabra, N. K. Garg, and M. Kumar, "Content-based image retrieval system using ORB and SIFT features," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 2725–2733, Apr. 2020.
[3] J. Wen, X. Zhou, M. Li, P. Zhong, and Y. Xue, "A novel natural language steganographic framework based on image description neural network," *J. Vis. Commun. Image Represent.*, vol. 61, pp. 157–169, May 2019.
[4] O. García-Olalla, E. Alegre, L. Fernández-Robles, E. Fidalgo, and S. Saikia, "Textile retrieval based on image content from CDC and webcam cameras in indoor environments," *Sensors*, vol. 18, no. 5, p. 1329, Apr. 2018.
[5] S. Saikia, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Object detection for crime scene evidence analysis using deep learning," in *Proc. Int. Conf. Image Anal. Process.* Springer, 2017, pp. 14–24. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-68548-9_2
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[7] Y. Djenouri, G. Srivastava, and J. C.-W. Lin, "Fast and accurate convolution neural network for detecting manufacturing data," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2947–2955, Apr. 2021.
[8] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.

[9] S. Cao, G. An, Z. Zheng, and Q. Ruan, "Interactions guided generative adversarial network for unsupervised image captioning," *Neurocomputing*, vol. 417, pp. 419–431, Dec. 2020.
[10] D. Połap, "An adaptive genetic algorithm as a supporting mechanism for microscopy image analysis in a cascade of convolution neural networks," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106824.
[11] D. Polap, "Analysis of skin marks through the use of intelligent things," *IEEE Access*, vol. 7, pp. 149355–149363, 2019.
[12] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and Á. García-Martín, "Semantic-aware scene recognition," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107256.
[13] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.
[14] H. Han, J. Li, A. K. Jain, S. Shan, and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2333–2348, Oct. 2019.
[15] S. Saikia, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Query based object retrieval using neural codes," in *Proc. Int. Joint Conf. SOCO-CISIS-ICEUTE*. León, Spain: Springer, Sep. 2017, pp. 513–523.
[16] S. Maji and S. Bose, "CBIR using features derived by deep learning," 2020, *arXiv:2002.07877*. [Online]. Available: http://arxiv.org/abs/2002.07877
[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[18] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.
[19] H. Li, Y. Huang, and Z. Zhang, "An improved faster R-CNN for same object retrieval," *IEEE Access*, vol. 5, pp. 13665–13676, 2017.
[20] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[23] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005-96, 1996.
[24] 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. 1–21, Feb. 2015.
[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[26] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting Oxford and Paris: Large-scale image retrieval benchmarking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5706–5715.
[27] J. Zhu, J. Wang, S. Pang, W. Guan, Z. Li, Y. Li, and X. Qian, "Co-weighting semantic convolutional features for object retrieval," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 368–380, Jul. 2019.
[28] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.
[29] O. Simeoni, Y. Avrithis, and O. Chum, "Local features and visual words emerge in activations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11651–11660.
[30] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3456–3465.
[31] A. Gordo, J. Almazán, 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, Sep. 2017.
[32] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107148.
[33] Y. Cai, Y. Li, C. Qiu, J. Ma, and X. Gao, "Medical image retrieval based on convolutional neural network and supervised hashing," *IEEE Access*, vol. 7, pp. 51877–51885, 2019.
[34] S. R. Dubey, S. K. Roy, S. Chakraborty, S. Mukherjee, and B. B. Chaudhuri, "Local bit-plane decoded convolutional neural network features for biomedical image retrieval," *Neural Comput. Appl.*, vol. 32, pp. 7539–7551, Jun. 2019.
[35] W. Zhou, Z. Shao, C. Diao, and Q. Cheng, "High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder," *Remote Sens. Lett.*, vol. 6, no. 10, pp. 775–783, 2015.
[36] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, p. 489, May 2017.
[37] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
[38] A. Salvador, X. Giro-i-Nieto, F. Marques, and S. Satoh, "Faster R-CNN features for instance search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 394–401.
[39] E. Mohedano, K. McGuinness, X. Giró-i-Nieto, and N. E. O'Connor, "Saliency weighted convolutional features for instance search," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2018, pp. 1–6.
[40] M. Teichmann, A. Araujo, M. Zhu, and J. Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5109–5118.
[41] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2475–2483.
[42] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3270–3278.
[43] X. Lu, Y. Chen, and X. Li, "Discrete deep hashing with ranking optimization for image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2052–2063, Jun. 2020.
[44] Y. Ding, W. K. Wong, Z. Lai, and Z. Zhang, "Discriminative dual-stream deep hashing for large-scale image retrieval," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102288.
[45] H. Lu, M. Zhang, X. Xu, Y. Li, and H. T. Shen, "Deep fuzzy hashing network for efficient image retrieval," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 1, pp. 166–176, Jan. 2021.
[46] Y. Chen, X. Lu, and X. Li, "Supervised deep hashing with a joint deep network," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107368.
[47] Y. Cai, K. Huang, and T. Tan, "Matching tracking sequences across widely separated cameras," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 765–768.
[48] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt, "Cascade of descriptors to detect and track objects across any network of cameras," *Comput. Vis. Image Understand.*, vol. 114, no. 6, pp. 624–640, Jun. 2010.
[49] D.-N. Truong Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray, "People re-identification by spectral classification of silhouettes," *Signal Process.*, vol. 90, no. 8, pp. 2362–2374, Aug. 2010.
[50] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *J. Appl. Remote Sens.*, vol. 8, no. 1, Jul. 2014, Art. no. 083584.
[51] J. Pujari, S. N. Pushpalatha, and D. Padmashree, "Content-based image retrieval using color and shape descriptors," in *Proc. Int. Conf. Signal Image Process.*, Dec. 2010, pp. 239–242.
[52] A. Alzu'bi, A. Amira, N. Ramzan, and T. Jaber, "Robust fusion of color and local descriptors for image retrieval and classification," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Sep. 2015, pp. 253–256.
[53] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2911–2918.
[54] D. Cortes, G. Calderón, A. Arista, K. Toscano, and M. Nakano, "Aerial image classification using texture and color-based descriptors," in *Proc. IEEE 1st Congreso Nacional Ciencias Geoespaciales (CNCG)*, Dec. 2016, pp. 1–4.
[55] R. Ashraf, M. Ahmed, S. Jabbar, S. Khalid, A. Ahmad, S. Din, and G. Jeon, "Content based image retrieval by using color descriptor and discrete wavelet transform," *J. Med. Syst.*, vol. 42, no. 3, pp. 42–44, Mar. 2018.
[56] M. Sotoodeh, M. R. Moosavi, and R. Boostani, "A novel adaptive LBP-based descriptor for color image retrieval," *Expert Syst. Appl.*, vol. 127, pp. 342–352, Aug. 2019.
[57] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[58] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," 2019, *arXiv:1904.01355*. [Online]. Available: http://arxiv.org/abs/1904.01355
[59] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," 2015, *arXiv:1510.07493*. [Online]. Available: http://arxiv.org/abs/1510.07493
[60] Q. Wang, J. Lai, L. Claesen, Z. Yang, L. Lei, and W. Liu, "A novel feature representation: Aggregating convolution kernels for image retrieval," *Neural Netw.*, vol. 130, pp. 1–10, Oct. 2020.
[61] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 3, pp. 251–258, 2016.
[62] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 685–701. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46604-0_48
[63] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 247–261, Feb. 2016.
[64] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 241–257. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46466-4_15
[65] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: Learning affine regions via discriminability," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 284–300.
[66] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 926–935.
[67] A. Jimenez, J. M. Alvarez, and X. Giro-i-Nieto, "Class-weighted convolutional features for visual instance search," Jul. 2017, *arXiv:1707.02581*. [Online]. Available: https://arxiv.org/abs/1707.02581
[68] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 3–20. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46448-0_1
[69] J. Lin, Y. Zhan, and W.-L. Zhao, "Instance search based on weakly supervised feature learning," *Neurocomputing*, vol. 424, pp. 117–124, Feb. 2021.
[70] S. Liu, J. Wu, L. Feng, H. Qiao, Y. Liu, W. Luo, and W. Wang, "Perceptual uniform descriptor and ranking on manifold for image retrieval," *Inf. Sci.*, vol. 424, pp. 235–249, Jan. 2018.
[71] A. Mukherjee, J. Sil, A. Sahu, and A. S. Chowdhury, "A bag of constrained informative deep visual words for image retrieval," *Pattern Recognit. Lett.*, vol. 129, pp. 158–165, Jan. 2020.
[72] K. T. Ahmed, S. Ummesafi, and A. Iqbal, "Content based image retrieval using image features information fusion," *Inf. Fusion*, vol. 51, pp. 76–99, Nov. 2019.
[73] S. Newsam and Y. Yang, "Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery," in *Proc. 15th Annu. ACM Int. Symp. Adv. Geograph. Inf. Syst. (GIS)*, 2007, pp. 1–8.
[74] W. Bouachir, M. Kardouchi, and N. Belacel, "Improving bag of visual words image retrieval: A fuzzy weighting scheme for efficient indexation," in *Proc. 5th Int. Conf. Signal Image Technol. Internet Based Syst.*, Nov. 2009, pp. 215–220.
[75] A. Mukherjee, S. Chakraborty, J. Sil, and A. S. Chowdhury, "A novel visual word assignment model for content-based image retrieval," in *Proc. Int. Conf. Comput. Vis. Image Process.* Springer, 2017, pp. 79–87. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-10-2104-6_8

SURAJIT SAIKIA received the M.Sc. degree in mathematics with specialization in computer science and the M.Tech. degree from the Sri Sathya Sai Institute of Higher Learning, India, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree in artificial intelligence with the University of León, with a focus on deep learning and computer vision. His research interest mainly includes object detection, image retrieval, reinforcement learning, and scene understanding.

LAURA FERNÁNDEZ ROBLES received the M.Sc. degrees in industrial engineering and in intelligent systems in engineering from the University of León, in 2009 and 2011, respectively, and the Ph.D. degree from the University of Groningen, The Netherlands, and University of León, Spain, in 2016. She is currently an Assistant Lecturer and a Researcher with the University of León. She has participated in three European projects and four Spanish projects, among other smaller projects. She is coauthor in 25 indexed journal publications, 18 contributions in lecture notes and several proceedings of international conferences. Moreover, she is co-inventor of seven patents, four of which are licensed to companies. She has also been the supervisor of two Ph.D. Thesis. Her current research interests include computer vision, pattern recognition, and data science applied to industrial, cyber-security, medical, and animal ethology problems.

EDUARDO FIDALGO FERNÁNDEZ is currently an Industrial Engineer and Computer Science Doctor with the University of León. He is the coordinator of the GVIS – former VARP – research group. His objective is the research and development of solutions to problems related to cybersecurity and cybercrime with the Spanish National Cybersecurity Institute (INCIBE). He is involved in Artificial Intelligence projects, mainly related to natural language processing, artificial vision and both machine/deep learning techniques. He is an active member of the GVIS Group at the University of León. He has participated in competitive research projects, intellectual property registers and patents that are being exploited. He is also involved as an author and reviewer of journal and conference papers. From 2008 to 2016, he worked as a full-time Engineer at Indra Sistemas, where he acquired experience as a developer, software analyst, test leader, and project management. His primary expertise is the technical management of international projects.

ENRIQUE ALEGRE received the B.Sc. degree in industrial engineering from the University of Cantabria and the Ph.D. degree in computer science from the University of León.

He is currently the Director of the Research Group for Vision and Intelligent Systems, University of León, and he has participated in 20 research projects in public and competitive calls and even a bigger number with companies, has been the Principal Investigator, among others, of three European Projects. He has been co-inventor of 12 patents, five of which are licensed to companies, and 15 intellectual property registries. He has also been the coauthor in more than 140 papers, 45 of which have been published in indexed journals and 33 of them in top ones. In the almost 25 years working in Academia, he has done 11 international research visits: one of them a sabbatical year at the University of California and other eight in Groningen (The Netherlands), DCU (Ireland), Surrey (UK), Malta or Edimburgo (UK) universities, amongst others. He has been the supervisor of 11 Ph.D. Thesis and is or has been editor in five journals and reviewer in 35 International Journals or Conferences.