ACTNET: end-to-end learning of feature activations and multi-stream aggregation for effective instance image retrieval

Syed Sameed Husain, Eng-Jon Ong and Miroslaw Bober, *Member, IEEE*

Abstract—We propose a novel CNN architecture called ACTNET for robust instance image retrieval from large-scale datasets. Our key innovation is a learnable activation layer designed to improve the signal-to-noise ratio (SNR) of deep convolutional feature maps. Further, we introduce a controlled multi-stream aggregation, where complementary deep features from different convolutional layers are optimally transformed and balanced using our novel activation layers, before aggregation into a global descriptor. Importantly, the learnable parameters of our activation blocks are explicitly trained, together with the CNN parameters, in an end-to-end manner minimising triplet loss. This means that our network jointly learns the CNN filters and their optimal activation and aggregation for retrieval tasks. To our knowledge, this is the first time parametric functions have been used to control and learn optimal aggregation. We conduct an in-depth experimental study on three non-linear activation functions: Sine-Hyperbolic, Exponential and modified Weibull, showing that while all bring significant gains the Weibull function performs best thanks to its ability to equalise strong activations. The results clearly demonstrate that our ACTNET architecture significantly enhances the discriminative power of deep features, improving significantly over the state-of-the-art retrieval results on all datasets.

Index Terms—Image retrieval, global image descriptor, Convolutional Neural Network, deep features, activation functions, compact image signature

1 INTRODUCTION

THE last decade has seen an explosive growth in the usage of multimedia. Conventional solutions to the management of huge volumes of multimedia fell below expectations, stimulating active research in areas such as instance image retrieval (IIR) and object recognition. Visual recognition capabilities are crucial for many applications including mobile visual search, augmented reality, robotic vision, automotive navigation, mobile commerce, surveillance & security, content management and medical imaging, and they will only grow in importance as AI systems proliferate. In IIR, the aim is to retrieve images depicting instances of a user-specified object in a large unordered collection of images. The task is challenging as the objects to be retrieved are often surrounded by background clutter or partially occluded. Additionally, variations in the object's appearance exist due to non-linear transformations such as view-point and illumination variations, ageing and weather changes, etc. Consequently, robust techniques that can cope with significant variability of local image measurements and numerous outliers are required. Moreover, modern systems must be scalable due to the overwhelming volumes of multimedia data.

In order to overcome these challenges, a compact mathematical representation of image content that supports fast and robust object retrieval is required. While Convolutional Neural Networks (CNN) have contributed significantly to the best performing methods in many computer vision tasks, including image classification, they are yet to fully meet the demanding user expectations in image retrieval tasks, especially on industrial-scale datasets. This is evidenced by the best prior art result reported on a extremely challenging $\mathcal{R}$Oxford-hard+1Million dataset at 19.9% mAP [1]. (which is improved by the presented design to 29.9%).

To bridge the gap between retrieval performance and user's expectations, we develop a novel deep neural network architecture called the Activation Network (ACTNET). In our design, we focus on areas where relatively little innovation happened; specifically on a novel and learnable aggregation framework capable of combining multi-layer dense convolutional features into a powerful global descriptor. Our architecture brings very significant performance gains, which translate into world leading retrieval rates (50% relative gain in mAP on $\mathcal{R}$Oxford-hard+1Million) or enable significant reduction in computational complexity (5-fold). Our main contributions are:

- we propose a novel trainable activation layer that improves the signal-to-noise ratio (SNR) of deep convolutional feature maps. We conduct an in-depth experimental study to illustrate the effects of applying three non-linear activation functions: Sine-Hyperbolic, Exponential and modified Weibull. The results show that these activation functions significantly enhance the discriminative power of convolutional features, leading to world-class retrieval results.
- we design a multi-layer CNN architecture (ACTNET) where deep features from different convolutional layers are transformed using our novel activation layers and optimally aggregated into a compact global descriptor. Crucially, the trainable parameters

S. Husain, E.Ong and M. Bober are with University of Surrey, Guildford, Surrey, UK. E-mails: {sameed.husain, e.ong, m.bober}@surrey.ac.uk

from multiple CNN layers including the parameters of activation blocks are explicitly trained in an end-to-end manner minimising the triplet loss.
- we develop a low-latency, small-memory and low-power ACTNET model for real world applications. Detailed experimental results show that our low-complexity ACTNET achieves comparable performance to the state-of-the-art RMAC [2] and GEM [3] CNNs while providing five times faster extraction speed.
- we perform extensive experiments on several datasets and show that ACTNET outperforms the best CNN-based methods: MAC [4], RMAC [2] and GEM [3].

The ACTNET architecture generates a robust and compact global descriptor that works extremely well for instance image retrieval. At its base, it uses any CNN model (here we apply it to two diverse models: ResNext [5] and MobileNet [6]) as a powerful convolutional feature extractor. These deep features are transformed and aggregated using novel activation layers, average pooling layers and a PCA+Whitening layer. ACTNET is trained in an end-to-end manner using Stochastic Gradient Descent (SGD) with a triplet loss function to jointly optimise multi-layer features and their aggregation. The proposed method achieves retrieval performances of 72.7%, 52.2%, 80.7% and 62.7% on $\mathcal{R}$Oxf$M$, $\mathcal{R}$Oxf$H$, $\mathcal{R}$Par$M$ and $\mathcal{R}$Par$H$ datasets respectively, outperforming the latest state-of-the-art methods including those based on computationally costly local descriptor indexing and spatial verification.

The paper is organised as follows: Section 2 presents a literature review of existing methods. Section 3 presents in detail our novel ACTNET architecture. The experimental setup and an extensive evaluation of ACTNET is presented in Section 4. In Section 5 we compare our results with the state-of-the-art, demonstrating significant improvement over recent global descriptors on all retrieval datasets. Finally, Section 6 concludes the paper.

2 Related Work on Image Retrieval

This section presents an overview of systems that have contributed significantly to instance image retrieval.

2.1 Classical image retrieval

Classical techniques for IIR involve aggregating scale-invariant hand-crafted descriptors such as SIFT [7] or SURF [8] into a single global image descriptor for fast matching. The most popular global representations that encode distributions of local descriptors in an image are Bag of Features (BoF) [9], Fisher Vectors (FV) [10], Vector of Locally Aggregated Descriptors (VLAD) [11], Triangulation Embedding (TEmb) [12] and Robust Visual Descriptor (RVD) [13]. In BoF, each image is represented as a histogram of visual word occurrences. Fisher Vectors encode descriptors based on a Fisher-Kernel framework; VLAD is a simplified version of FV, computed by aggregating the residual vectors between local descriptors and their corresponding visual words. The TEmb representation encodes descriptors using democratic aggregation while RVD uses rank-based multi-assignment and direction preserving mapping to aggregate descriptors. Several advancements have been made to improve the performance of global representations such as incorporating a large visual vocabulary [14], [15], query expansion [16], [17] and spatial verification [14].

Methods based on hand-crafted descriptors offer limited retrieval performance due to poor robustness of local features detectors and descriptors, to complex image transformations such significant view point and illumination changes.

2.2 Deep CNN image retrieval

Recent approaches for IIR use the deep convolutional features of a CNN, typically trained for ImageNet classification, to compute global image representations. Azizpour et al. [18] perform max-pooling while Babenko et al. [19] apply sum-pooling to the last convolutional features of VGG16 [20]. One step improvement is the cross-dimensional weighted sum-pooling of Kalantidis et al. [21]. Popular aggregation methods such as Fisher Vectors, VLAD and RVD are adapted to encode deep features in the work of Ong et al. [22], Arandjelovic et al. [23] and Husain et al. [13]. Tolias et al. [4] develop a hybrid scheme called Regional Maximum Activations of Convolutions (RMAC), where last convolutional layer features are first max-pooled over regions using a fixed grid. The regional features are PCA+whitened, L2-normalised and sum-aggregated to form an image signature. A modified version of RMAC that combines multi-scale and multi-layer feature extraction is introduced by Seddati et al. [24]. The method of Jimenez et al. [25] uses VGG16 and class activation maps to compute a global descriptor. Xu et al. [26] propose a semantic-based aggregation approach where probabilistic proposals corresponding to special semantic content in an image are used to aggregate deep features. A hybrid deep feature aggregation method that unifies sum and weighted pooling to compute an image representation is presented by Pang et al. [27].

The methods that use convolutional features from a CNN, trained for ImageNet classification, perform sub-optimally for IIR due to the CNN being tuned to optimise intra-class generalisation. Current methods aim to solve this limitation by finetuning the trained networks for retrieval tasks.

2.3 Finetuned deep CNN image retrieval

Arandjelovic et al. [23] introduce NetVLAD, where the last convolutional layer of VGG16 is followed by a generalised VLAD layer and the whole architecture is trained using weakly supervised triplet loss. Gordo et al. [2] build on the core RMAC network and train the deep image retrieval (DIR) architecture on the Landmarks dataset with triplet loss. Radenovic et al. [3] propose to aggregate deep features using a generalised mean pooling (GEM) layer. More precisely, the GEM layer is added to the last convolutional layer of ResNet101 and the resultant architecture is trained on the Landmarks dataset to minimise contrastive loss. Two variants of GEM pooling are weighted generalised mean pooling (WGEM) [28] and attention-aware generalised mean pooling (AGEM) [29]. Xu et al [30] develop an

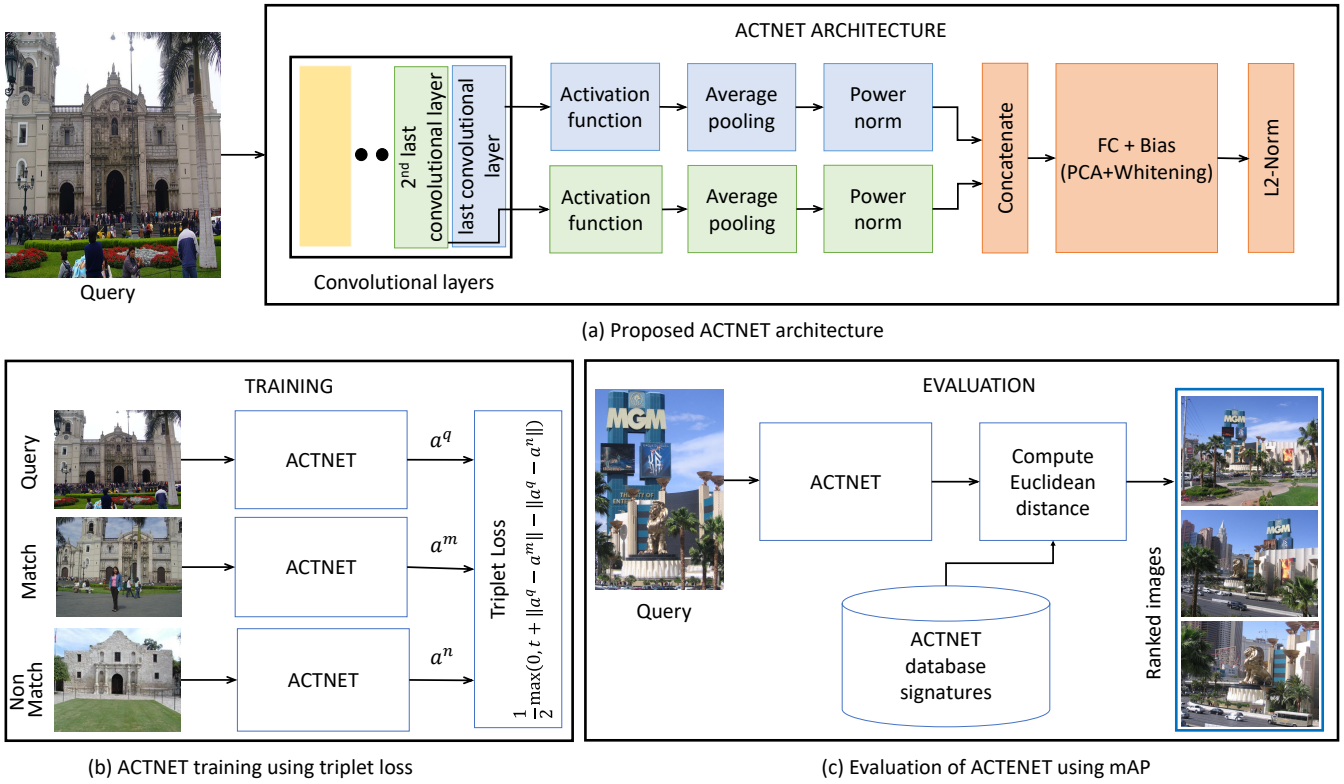


Fig. 1. (a) Proposed ACTNET architecture with multi-layer activation based aggregation, (b) training of ACTNET using triplet loss and (c) evaluation of ACTNET on state-of-the-art datasets

adversarial soft-detection-based aggregation (ASDA) network which pools deep features using adversarial detector and soft region proposal layer. Teichmann et al. [31] propose to aggregate and match deep local features (DELF) based on regional aggregated selective match kernels (R-ASMK). Husain et al. [32] propose a deep architecture, Region-Entropy based Multi-layer Abstraction Pooling (REMAP), where multi-layer regional features are aggregated based on entropy-guided pooling.

In summary, while very significant progress has recently been achieved, the retrieval performance of state-of-the-art methods is still underwhelming, especially on challenging large-scale datasets where many distractor images resembling the query image, but depicting a different object, exist. We attribute this weakness to three main causes, which our present work address, namely: (1) CNN architectures, and in particular the design of the aggregation stage which is responsible for robust global descriptor generation, (2) training protocols resulting in sub-optimal learning of extreme cases, and (3) mapping of the deep representations into compact signatures, which may lead to the reduction of its discriminatory power.

3 ACTNET ARCHITECTURE

In this section, we present the motivation, introduce details of our novel ACTNET architecture, describe its components and conduct an in-depth experimental evaluation of various elements of ACTNET.

First, we present a novel learnable aggregation for deep descriptors. We observe that distinctive features (i.e. represented by strong CNN responses) are typically contaminated by the background noise, i.e. the signals (CNN responses) from weaker but numerous features. In the max-pooling aggregation, only the strongest response is retained in each channel, however there is no guarantee that the strongest response corresponds to the object of interest. In the average pooling, on the other hand, all responses are averaged within channels, leading to the contamination effect. Our aim is to maximise the signal-to-noise ratio (SNR) of the convolutional feature map before aggregation, i.e. we would like to amplify the signals originating from the actual object of interest in an image, relative to the background noise. In our design, this is achieved by passing filter responses from convolutional layer through a learnable, non-linear "amplifier" function, implemented as a parametric activation layer. We propose three differentiable amplification functions, and perform an in-depth experimental study to compare their behaviour. The results show that activation functions significantly enhance the robustness and discriminative power of deep features, leading to the state-of-the art retrieval performance.

Second, based on our parametric activation layer, we design a deep multi-layer architecture, called Activation Network (ACTNET), to optimise retrieval tasks. We note that various convolutional layers of a deep CNN act as powerful signal generators that extract dense local features at various levels of semantic abstraction and with varying extent of region (spatial) support. However, conventional aggregation methods, such as max- or average-pooling fail to combine such features effectively. In ACTNET, multiple

parametric activations deliver adaptive multi-stream aggregation, where a hierarchy of deep features from different convolutional layers are amplified using activation functions and optimally aggregated into a discriminative global signature. The ACTNET architecture is trained in an end-to-end manner, where the convolutional weights, activation parameters and PCA+Whitening weights are jointly trained with triplet loss. The key aspect is that ACTNET learns the activations across multiple layers during training to select and balance features reflecting complementary and distinct levels of visual abstractions, thus boosting retrieval performance.

While current image retrieval systems exhibit good extraction speed on GPU-accelerated computers, our aim is to deliver an efficient and effective system in the hands of mobile users. To this end, we develop a low power, low latency and small memory footprint ACTNET suitable for running on mobile devices. The low-complexity Mobile-ACTNET achieves performance comparable to computationally expensive methods such as GEM [3] or RMAC [2] with a five times faster descriptor extraction speed.

ACTNET, presented in Figure 1a, consists of a baseline CNN followed by our aggregation network. Any CNN can serve as the base; we show the results on two distinct networks: the high-performance ResNext101 [5] and low-complexity MobileNetV2 [6]. For these CNNs, only the convolutional layers are retained, with the fully connected layers for classification discarded. The outputs of the final convolutional blocks are each fed to separate "aggregation streams" that utilise our proposed activation layers. Details of the multi-stream activation layer are presented in Section 3.1. In each stream, the activation layer is followed by average pooling and power normalisation layers. The outputs of all streams are then combined with a concatenation layer. This is followed by a PCA+Whitening layer before the final L2-normalisation layer, producing a discriminative global descriptor for an image. The training procedure (shown in Figure 1b) adopts a three stream siamese architecture where the image signatures extracted by each of the streams are jointly examined by the triplet loss function. The training procedure for the ACTNET is described in Section 3.4. In the evaluation phase (Figure 1c), an ACTNET signature is computed from a query image and compared against a database of pre-computed signatures using Euclidean distance. Based on the similarity scores, the database images are ranked and mean Average Precision (mAP) is computed.

3.1 Learnable non-linear activation layer

In this section, we present a novel CNN layer which we call the 'activation layer' that provides the ability to learn and tune non-linear activation functions. The aim is to give greater emphasis to important responses from the object in an image and reduce the impact of background noise by non-linearly scaling values of the input tensors. Crucially, this non-linear scaling is tune-able during the learning process.

The fundamental difference between our work and state-of-the-art on developing adaptive activation functions is that we aim to improve the SNR of the final convolutional feature map before aggregation into a global descriptor while current work focuses on improving the speed and stability of the network. Our activation function non-linearly amplifies convolutional layer values, thereby improving the robustness and discriminative power of deep features. Recent works on using activation functions to help fast training of CNNs are [33], [34], [35], [36], [37]. The rectified linear activation (ReLU) function [34] has made it easier to effectively train a CNN compared to activation functions such as sigmoid or tanh, by addressing the problem of vanishing gradients. Agostinelli et al. [37] propose to use an adaptive linear piecewise model for the activation function. Farhadi et al. [37] propose an adaptive ReLU function that provides a smoother scaling increase of activation values. Qian et al. [36] explore methods to linearly combine basic activation functions.

We denote the non-linear activation function as $\phi(\mathbf{X}; \theta_\phi, \mathbf{w}_\theta)$, where $\phi : \mathbb{R}^{W\times H\times D} \rightarrow \mathbb{R}^{W\times H\times D}$ is an operation that takes as input a tensor $\mathbf{X}$ of size $W \times H \times D$ and applies an associated real-valued function $\theta_\phi : \mathbb{R} \rightarrow \mathbb{R}$ to each element of the input tensor. Importantly, we consider parameterised functions θ_ϕ, where their respective parameters are denoted as a vector of real values: $\mathbf{w}_\theta$. More specifically, we have:

$$\phi(\mathbf{X}; \theta_\phi, \mathbf{w}_\theta) = (\theta_\phi(x_{ijk}; \mathbf{w}_\theta))_{i,j,k=1}^{W,H,D} \quad (1)$$

In this work, we have considered three different activation functions for θ_ϕ: Sine-Hyperbolic, Exponential and modified Weibull. These three are selected thanks to the following properties that they have in common:

- Each activation function non-linearly amplifies the input values within a real-valued interval. This allows us to place greater emphasis on larger values (important activations), compared to a linear scaling.
- The non-linearity of the amplification is adjustable for learning purposes. As such, all the activation functions considered are parameterised.
- Finally, for use in a stochastic gradient descent framework, we require all the functions to be differentiable with respect to all their parameters.

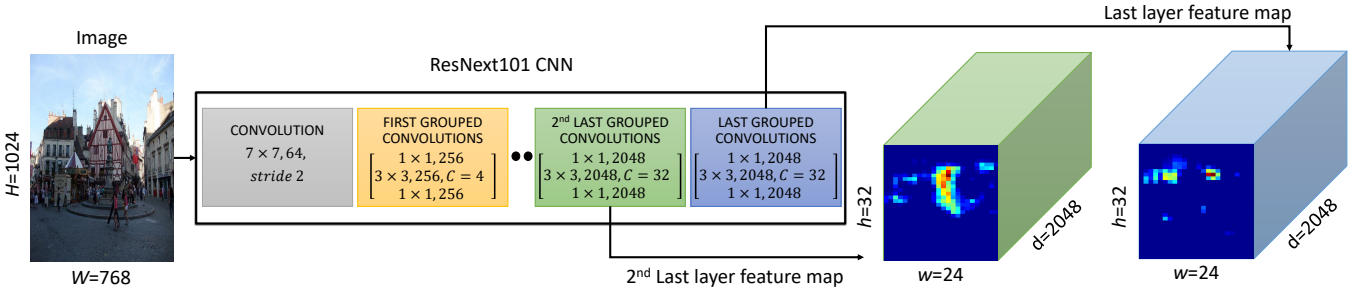
We next give the details of the three different activation functions used in ACTNET. For clarity and convenience, we will omit writing the vector of parameters, $\mathbf{w}_\theta$, when describing the function, writing it as $\theta_\phi(x)$ instead of $\theta_\phi(x; \mathbf{w}_\theta)$.

Sine-Hyperbolic function (SinH)

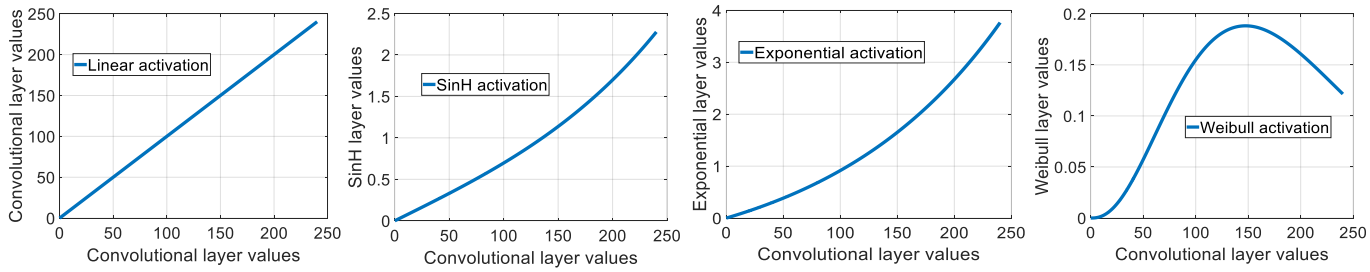
The first activation function we consider is the monotonically increasing Sine-Hyperbolic.

$$\theta_\phi(x) = \alpha \sinh(\beta x) \quad (2)$$

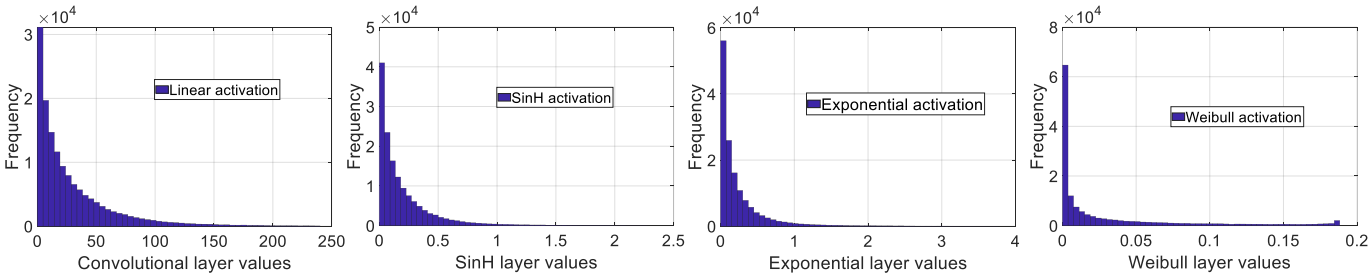
where α and β are the scaling parameters. The β parameter determines the strength of the non-linear re-weighting of the original activation values. As such, it effectively controls the influence that higher valued features have on the average pooling operation performed later on. The α parameter has two important roles: (1) normalising the values of the scaled activations, (2) weighting the multi-layer aggregation streams to compute the optimal global descriptor. The set



(a) An illustration of how the inputs to the activation functions are generated. The output tensors from the final convolutional layers are passed to the activation functions shown below.



(b) The profile curves of the activation functions considered in this paper: Sine-Hyperbolic, Exponential and Weibull.



(c) histograms of feature map values before and after passing through activation functions.

Fig. 2. Process of transforming convolutional layer values using the activation functions

of parameters associated with this activation function is: $\mathbf{w}_\theta = (\alpha, \beta)$.

Since the Sine-Hyperbolic function is differentiable, the chain-rule can be readily applied to obtain the partial derivatives required for back-propagation:

$$\frac{\partial \theta_\phi}{\partial x} = \alpha\beta\cosh(\beta x) \quad (3)$$

$$\frac{\partial \theta_\phi}{\partial \alpha} = \sinh(\beta x) \quad (4)$$

$$\frac{\partial \theta_\phi}{\partial \beta} = \alpha x\cosh(\beta x) \quad (5)$$

Exponential function (Exp)

The second activation function proposed to improve the SNR of feature maps is based on the exponential function:

$$\theta_\phi(x) = \alpha(\exp(\beta x) - 1) \quad (6)$$

where α and β are the scaling parameters, which play analogous roles as the parameters in the SinH activation function. Thus, the set of parameters associated with this function is given as: $\mathbf{w}_\theta = (\alpha, \beta)$. The partial derivatives used for learning for the exponential function are as follows:

$$\frac{\partial \theta_\phi}{\partial x} = \alpha\beta\exp(\beta x) \quad (7)$$

$$\frac{\partial \theta_\phi}{\partial \alpha} = \exp(\beta x) - 1 \quad (8)$$

$$\frac{\partial \theta_\phi}{\partial \beta} = \alpha x\exp(\beta x) \quad (9)$$

Modified Weibull function (WB)

Finally, we propose to transform the input values by applying a modified Weibull function:

$$\theta_\phi(x) = \left(\frac{x}{\alpha}\right)^{\beta-1}\exp(-(x/\gamma)^\zeta) \quad (10)$$

with parameters: $\mathbf{w}_\theta = (\beta, \alpha, \gamma, \zeta)$. Here β is known as the shape parameter and determines where the peak of the activation function will be. The Weibull function is a product of two terms, one term increasing at polynomial rate $(x/\alpha)^{\beta-1}$, another term decreasing at an exponential rate $\exp(-(x/\gamma)^\zeta)$. For small enough values, the polynomial term dominates, thus the effect of this activation function is to non-linearly increase these values. However, eventually

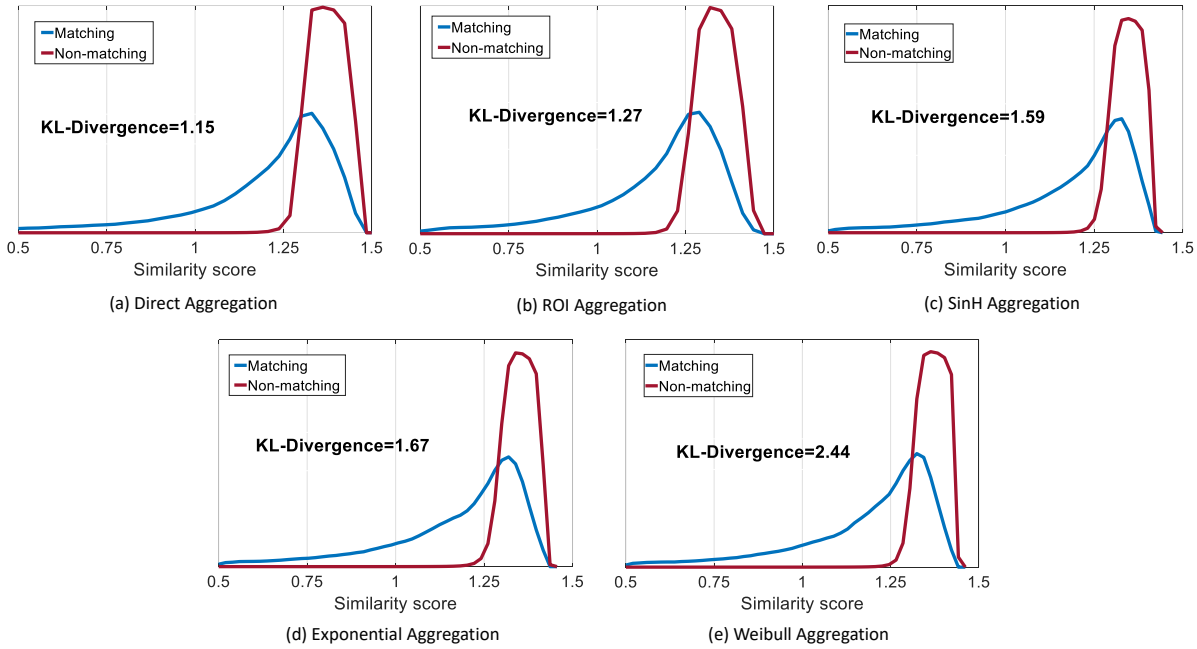


Fig. 3. Histograms of Euclidean distances between matching and non-matching descriptors, for different aggregation methods

the inverse exponential term starts to dominate, thereby reducing the output value as the input value further increases. The point x_0 where activation values change from being increased to being decreased can be found by solving $\partial\theta_\phi/\partial x = 0$ for x:

$$x_0 = \gamma\left(\frac{\beta-1}{\zeta}\right)^{1/\zeta} \tag{11}$$

When we consider how the Weibull function changes the result of the average pooling operation later, we find that two fundamentally different forms of re-weighting occur: 1) tensor values before x_0 will have a non-linearly greater influence (polynomial rate), the closer they are to x_0; 2) tensor values after x_0 are effectively "reversed", where they exert an exponentially decreasing influence as they get larger.

The partial derivatives used for back propagation are:

$$\frac{\partial\theta_\phi}{\partial x} = \left(\frac{\beta-1}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-2}\exp(-(x/\gamma)^\zeta) - \tag{12}$$

$$\frac{\zeta}{\gamma^\zeta}\left(\frac{x}{\alpha}\right)^{\beta-1}x^{\zeta-1}\exp(-(x/\gamma)^\zeta) \tag{13}$$

$$\frac{\partial\theta_\phi}{\partial\beta} = \log\left(\frac{x}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1}\exp(-(x/\gamma)^\zeta) \tag{14}$$

$$\frac{\partial\theta_\phi}{\partial\alpha} = (1-\beta)\left(\frac{x^{\beta-1}}{\alpha^\beta}\right)\exp(-(x/\gamma)^\zeta) \tag{15}$$

$$\frac{\partial\theta_\phi}{\partial\gamma} = \zeta\left(\frac{x^{\beta+\zeta-1}}{\alpha^{\beta-1}\gamma^{\zeta+1}}\right)\exp(-(x/\gamma)^\zeta) \tag{16}$$

$$\frac{\partial\theta_\phi}{\partial\zeta} = \left(\frac{x}{\alpha}\right)^{\beta-1}\left(-\frac{x}{\gamma}\right)^\zeta\log\left(\frac{x}{\zeta}\right)\exp(-(x/\gamma)^\zeta) \tag{17}$$

The above three activation functions are integrated into the ACTNET architecture in Section 3.3. However, we will first analyse how different activation functions affect the image features in the next section.

3.2 Analysis of activation functions

Figure 2 demonstrates the process of applying activation functions to the features. Firstly, an input tensor for the activation function is obtained by extracting selected convolutional layer output from a base CNN (here ResNext101). The elements of this tensor fall in some non-negative range (Figure 2a). We now apply different functions for non-linearly scaling the values of tensor. Transformations of input values by the activation functions considered are shown in Figure 2b. Here, it can be seen that the three activation functions have different shapes, providing different degrees of non-linear amplification depending on the signal level and the associated parameters. Importantly, the addition of the activation layer before aggregation gives us a significant improvement over direct aggregation where the deep features are linearly scaled and aggregated (refer Section 4). The SinH function provides a non-linear increase that is in between the linear curve and exponential curve. The Exponential function scales up the input values more steeply than SinH. Finally, we see that the Weibull function provides us with a robust amplification, increasing values up to a certain point, before its activation values starts decreasing; the rationale here is that we aim to equalise the power of few strong activations corresponding to prominent and distinctive image features. This equalisation is crucial as we cant know - at the extraction and aggregation stage - which features correspond to foreground and which to background. This depends on the query and the other image being matched, and for some queries the strong activations may correspond to the background clutter, hence we do not want any activation to significantly dominate the aggregation process. In summary the Weibull function gives us two benefits - attenuation of the weaker activations and equalisation of the stronger ones.

The distributions of tensor values before and after pass-

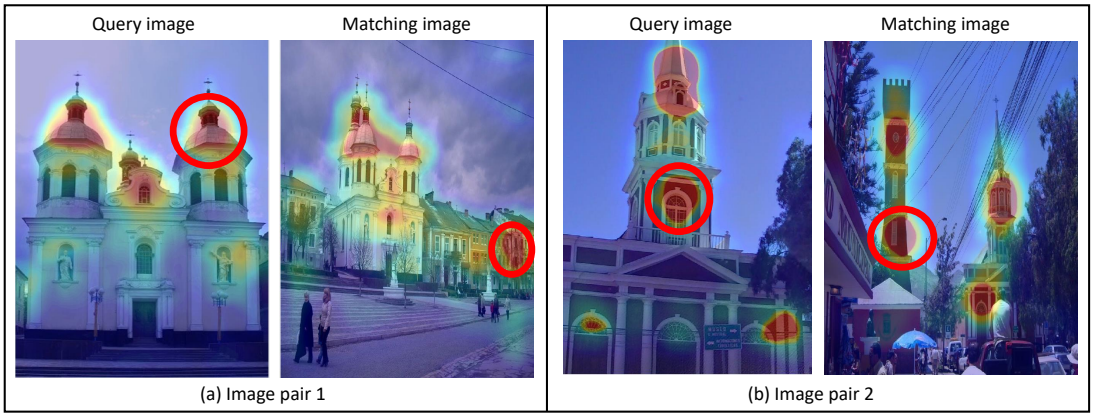


Fig. 4. Visualisation of feature maps of the last convolutional layer for two pairs of images: each comprising of a query image and a matching image.

ing through the activation functions for a sample image are presented in Figure 2c. It can be observed that activation functions significantly improve the SNR of feature maps by squeezing the majority of noisy responses to insignificantly small values. Importantly, the Weibull function results in a distribution with a large peak within the first bin showing the noise reduction phenomenon as well as a small peak in the final bin of the distribution showing the ability to equalise strong activations.

The application of activation functions to the feature map values modifies the underlying distributions of similarity scores for matching and non-matching pairs of images. To demonstrate the benefits of the proposed activation layer, we compute the class-separability between histograms of distances for matching and non-matching image pairs. The descriptors are extracted using the following aggregation approaches:

1) Direct aggregation (DA) [19]: the features from the last convolution layers are aggregated using average pooling
2) Region of Interest based aggregation (ROIA) [2]: the features are first max-pooled across several multi-scale overlapping regions. The regional descriptors are aggregated using sum pooling
3) SinH aggregation (SinHA): the features are transformed using the Sine-Hyperbolic activation layer before aggregation using average pooling
4) Exponential aggregation (ExpA): the features are passed through the Exponential activation layer before average pooling.
5) Weibull aggregation (WBA): the features are transformed using the Weibull activation layer before average pooling.

In all the aforementioned methods, the aggregated signatures are then PCA+Whitened and L2-normalised to form a global descriptor. We then compute $P(s/m)$ and $P(s/n)$ as the probability of observing a Euclidean distance s for a matching and a non-matching descriptor pair m and n respectively. The separability between $P(s/m)$ and $P(s/n)$ across values of s is computed using KL-Divergence (KLD). It can be observed from Figure 3 that the aggregation methods based on activation functions, SinHA (KLD=1.59), ExpA (KLD=1.67), WBA (KLD=2.44), provide much better separability between matching and non-matching distributions when compared to DA (KLD=1.15) and ROIA (KLD=1.27). It can also be observed that the Weibull aggregation achieves the highest KL-Divergence of 2.44. Importantly, we find that an increase in KL-divergence correlates strongly with the increase in retrieval accuracy, as shown in Section 4.

In Figure 4, we visualise why it is important to apply activation function before aggregation. We show the last convolutional layer responses for two pairs of images: each comprising of a query image (left) and a matching image (right). The strongest responses are indicated by red circles. The max-pooling method of [18] generates global descriptors from a query image and matching image that have a very low similarity score (Euclidean distance 1.42); this is because the maximum activation for the matching image (right) is from the background and not the actual object to be retrieved. The average-pooling [19] method alleviates the problem of max-pooling, however the matching descriptor is now contaminated with significant amount of background responses resulting in a low similarity score (Euclidean distance 1.39). The application of SinH or Exponential functions to the input tensor significantly improves the signal-to-noise ratio thereby increasing the similarity score between global descriptors (Euclidean distance 1.27 and 1.21). The additional benefit of the Weibull activation function is that it balances the influence of the strong activations, thus providing the best similarity score (Euclidean distance 1.08).

3.3 Image retrieval ACTNET architecture

We will now formally describe the ACTNET architecture. First, let $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$ denote an RGB input image of resolution $W \times H$. Next, for the selected base CNN, given an input image $\mathbf{x}$, suppose we are interested in using K of its convolutional layer outputs. Let us denote these K convolutional layers as functions: $f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_K(\mathbf{x})$. Associated with each of these functions are their own set of learnable parameters, which are optimised via back-propagation using the partial derivatives $\partial\theta_\phi/\partial x$ given in Section 3.1. These convolutional outputs are then be passed as inputs to their own "aggregation streams" as described next.

Aggregation stream

The aim of the aggregation stream is to perform non-linear scaling and aggregation of convolutional features. First, the input tensor $\mathbf{X} \in \mathbb{R}^{W \times H \times D}$ is transformed using the activation function ϕ (Eq. 1) resulting in the output tensor $\mathbf{Y} \in \mathbb{R}^{W \times H \times D}$

$$\mathbf{Y} = \phi(\mathbf{X}; \theta_\phi, \mathbf{w}_\theta) \tag{18}$$

We next apply global average pooling denoted by $P(\mathbf{Y})$:

$$P(\mathbf{Y}) = \left(\frac{1}{WH} \sum_{i}^{W} \sum_{j}^{H} y_{ijk} \right)_{k=1}^{D} \tag{19}$$

The output of the global average pooling operation is the vector $\mathbf{z} = P(\mathbf{Y})$, which we normalise to balance the severity of increase in values due to the non-linear scaling process of the activation layer. In our work, we have found that the choice of fractional power achieves best results. The power normalisation function is denoted as $\psi : \mathbb{R}^D \rightarrow \mathbb{R}^D$, with the rule:

$$\psi(\mathbf{z}) = \lambda z_1^p, \lambda z_2^p, ..., \lambda z_D^p \tag{20}$$

where λ, p are learnable scaling parameters.

To bring it all together, we can now define the full aggregation stream as the function: $\Phi : \mathbb{R}^{W \times H \times D} \rightarrow \mathbb{R}^D$, with the rule:

$$\Phi(\mathbf{X}) = \psi(P(\phi(\mathbf{X}; \theta_\phi, \mathbf{w}_\theta))) \tag{21}$$

To aid the following discussion when multiple streams are considered, we gather all the learnable parameters described above into a set denoted as:

$$\eta = \{\mathbf{w}_\theta, p, \lambda\} \tag{22}$$

Global descriptor generation

We now describe how the global descriptor of an image $\mathbf{x}$ will be generated using convolutional features from the base CNN.

As noted previously, we are interested in using K convolutional layer outputs from the base CNN. Each such convolutional layer gives rise to a separate aggregation stream of form Eq. 21. We denote each of the K separate streams as: $\Phi_1(f_1(\mathbf{x})), \Phi_2(f_2(\mathbf{x})), ..., \Phi_K(f_K(\mathbf{x}))$. Their respective learnable parameters (Eq. 22) are denoted as: $\eta_1, \eta_2, ..., \eta_K$. The global descriptor is then constructed by concatenation of the outputs of all aggregation streams:

$$\mathbf{b} = [\Phi_1(f_1(\mathbf{x})), \Phi_2(f_2(\mathbf{x})), ..., \Phi_K(f_K(\mathbf{x}))]$$

The concatenated feature vector $\mathbf{b}$ is then passed to a PCA and whitening layer followed by a L2-normalisation layer, which produces the final global descriptor used for image retrieval.

3.4 End-to-end training of ACTNET

A vital aspect of the ACTNET is that all its components are considered to represent differentiable operations. Our novel activation layer is differentiable with parameters that can be optimised during training. The average pooling layer and the normalisation layer are also differentiable. The PCA+Whitening projection is implemented as a Fully Connected layer (for the projection with whitened eigenvectors) with bias (for mean subtraction), with weights that can be freely optimised. In conclusion, ACTNET is an end-to-end architecture which can learn the convolutional filter weights, activation parameters and PCA+Whitening parameters, using Stochastic Gradient Descent (SGD) on triplet loss function during training.

We will now show how to build and train ACTNET using ResNext101 [5] as the base CNN. The first phase of training starts by fine-tuning ResNext101 (trained on ImageNet) on the GLRD dataset (refer Section 4) using classification loss. In the second phase, we remove the last three layers of the tuned ResNext101 and add the activation layers, average pooling layers and power normalisation layers to the last two convolutional blocks. The outputs from the normalisation blocks are concatenated and a PCA+Whitening layer is added to form the final ACTNET. We then adopt a three stream siamese architecture to train the warmed-up ACTNET using triplet loss. For siamese network training, a dataset of R triplets, each comprising of a query image, a matching image and a non-matching image is considered. More precisely, let a^q be an ACTNET representation of a query image, a^m be a representation of a matching image and a^n be a representation of a non-matching image. The triplet loss can be defined as:

$$L = \frac{1}{2} \max\left(0, t + ||a^q - a^m||^2 - ||a^q - a^n||^2\right), \tag{23}$$

where t controls the margin. During training, the aim of the triplet loss is to adjust the ACTNET parameters such that the distance between a^q and a^m reduces and distance between a^q and a^n increases.

Given an query image at test time, ACTNET produces a robust and discriminative $D = 4096$ dimensional image representation, well-suited for image retrieval.

3.5 Low complexity ACTNET

The ACTNET systems based on high-performance CNNs such as VGG, ResNet101, ResNext101 and DenseNet require significant amount of computational resources beyond the capabilities of mobile devices. In this section, we develop a low complexity ACTNET to effectively maximise retrieval performance while being mindful of limited resource scenarios. The overall architecture consists of a MobileNetV2 base CNN followed by our aggregation system. The procedure to create the low complexity ACTNET (Mobile-ACTNET) is as follows: we remove the last pooling layer, prediction layer and loss layer of MobileNetV2 (trained on ImageNet). The outputs of the final two convolutional layers are then passed to activation layers, average pooling layers, power normalisation layers and a concatenation layer, generating a global descriptor. Finally, a siamese network is adopted to train Mobile-ACTNET on the GLDR dataset using triplet loss.

Given a query image at test time, Mobile-ACTNET generates a 2560-dimensional image signature. The dimensionality of the final signature is reduced using PCA+Whitening transformation, thereby reducing memory requirements and increasing the matching speed.

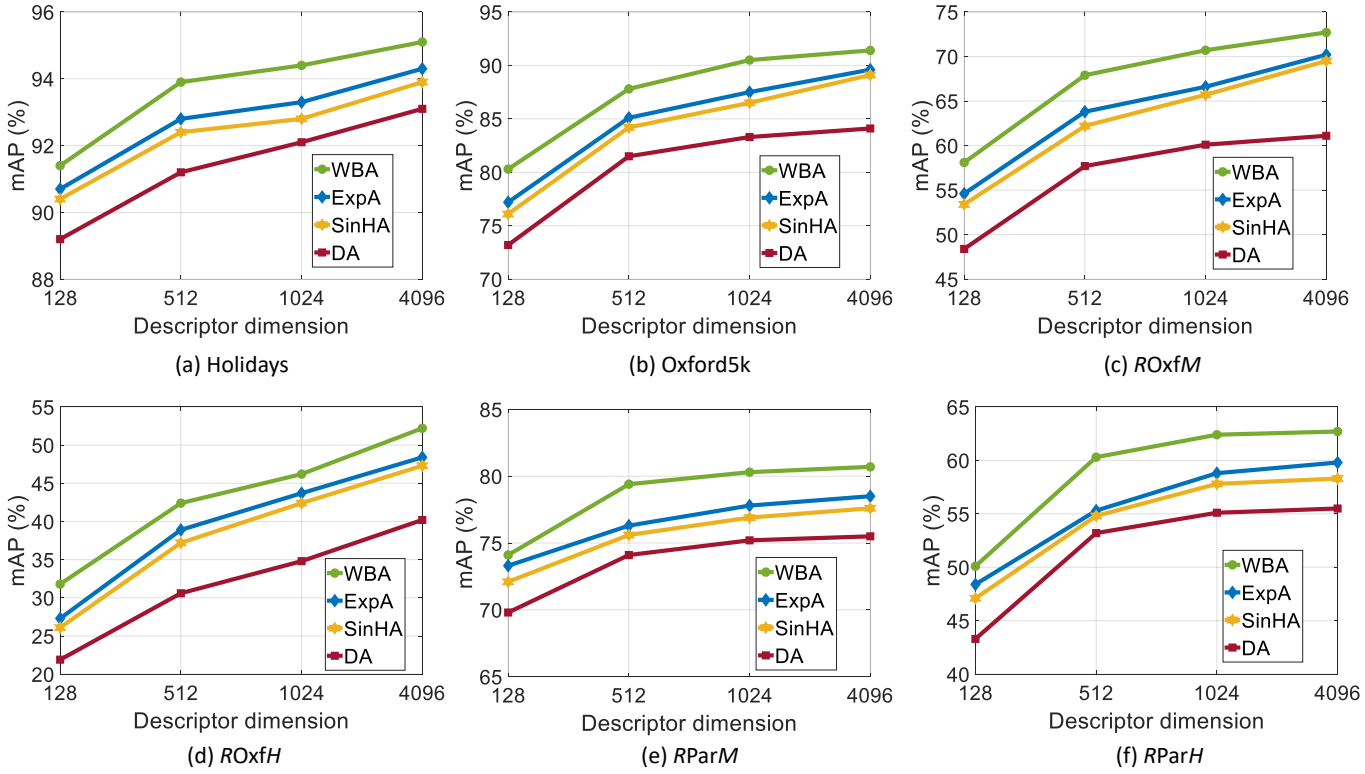


Fig. 5. Comparison of retrieval performance with different aggregation methods: Direct aggregation (DA), SinH aggregation (SinHA), Exponential aggregation (ExpA) and Weibull aggregation (WBA) on (a) Holidays, (b) Oxford5k, (c) ROxf$M$, (d) ROxf$H$, (e)RPar$M$, (f)RPar$H$, (all results in mAP(%));

In all the following experiments, we refer to ResNext-based ACTNET as "ACTNET" and MobileNetV2-based ACTNET as "Mobile-ACTNET".

4 Experimental Results

In this section we detail the implementation details of our architecture and give qualitative and quantitative results to validate our method. We first describe the experimental setup which includes the training parameters, training and test datasets and evaluation protocols. We then discuss the importance of the novel components of ACTNET, namely the robust activation function and multi-layer aggregation. Finally, we compare the proposed method to the state-of-the-art, showing significant improvements in retrieval performance across the board.

4.1 Training setup

We train ACTNET on a subset of Google Landmarks dataset (GLD) [38], which contains 1 million images depicting 15K unique landmarks (classes). In GLD the number of images per class is highly unbalanced, some classes contain thousands of images, while for around 8K classes only 20 or fewer images are present. Furthermore, the GLD contains a non-negligible fraction of images unrelated to the landmarks. For this reason, we preprocess the GLD using the RVDW global descriptor and SIFT based RANSAC [32], to obtain a clean dataset containing 200K images belonging to 2K classes. Finally, we remove all images that overlap with the test datasets. We refer this clean dataset as Google Landmarks Retrieval dataset (GLRD).

We adopt a siamese architecture and train ACTNET on the GLDR dataset using triplet loss. The objective function is optimised by Stochastic Gradient Descent (SGD) with learning rate 1e-03, weight decay of 5e-04, momentum 0.9 and triplet loss margin 0.1. Each triplet contains a query, a matching and a hardest non-matching image and the size of each image is fixed to 1024×768. To ensure that the sampled triplets incur loss and help in the training process, we extract the global descriptors from GLRD using the current model, and sample 5K triplets. All the triplets that cause a loss (the distance between query descriptor and non-matching descriptor is within margin 0.1 of the distance between query descriptor and matching descriptor) are sent to the network for training. After each epoch (training on 5K triplets), our algorithm generates a new set of 5K triplets using the current model for the following epoch. The training process is performed for at most 20 epochs. We overcome the TITAN X GPU memory limitation of 11 GB by processing one triplet at a time and updating the gradients after every 64 triplets.

4.2 Test datasets and evaluation protocol

The original Oxford5k dataset contains 5063 high-resolution images, with a subset of 55 images used as queries. In [1], Radenovic et al. revisited the original Oxford dataset to correct annotation errors, add 15 new challenging queries and introduce new evaluation protocols. The revisited Oxford

TABLE 1
Importance of Multi-layer aggregation of convolutional features

	Holidays	Oxford5k	$\mathcal{R}$Oxf$M$	$\mathcal{R}$Oxf$H$	$\mathcal{R}$Par$M$	$\mathcal{R}$Par$H$	MPEG
SLA	94.1	88.9	68.9	45.5	78.9	59.8	78.9
MLA	95.1	91.4	72.7	52.2	80.7	62.7	82.1

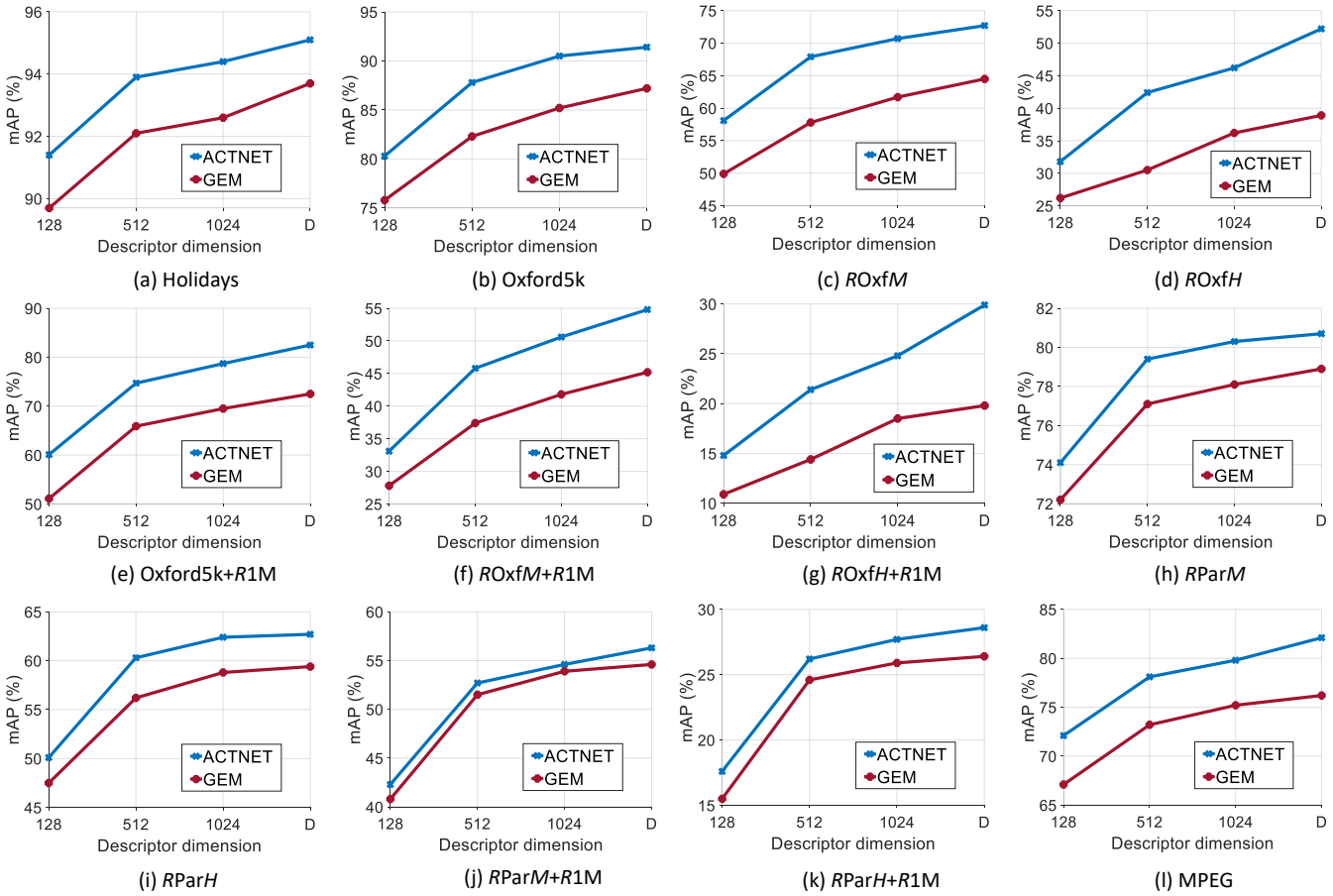


Fig. 6. ACTNET comparison with state-of-the-art GEM network (a) Holidays, (b) Oxford5k, (c) $\mathcal{R}$Oxf$M$, (d) $\mathcal{R}$Oxf$H$, (e) Oxford5k+$\mathcal{R}$1M, (f) $\mathcal{R}$Oxf$M$+$\mathcal{R}$1M, (g) $\mathcal{R}$Oxf$H$+$\mathcal{R}$1M, (h) $\mathcal{R}$Par$M$, (i) $\mathcal{R}$Par$H$, (j) $\mathcal{R}$Par$M$+$\mathcal{R}$1M, (k) $\mathcal{R}$Par$H$+$\mathcal{R}$1M, (l) MPEG, (all results in mAP(%));

dataset ($\mathcal{R}$Oxf) comprises of 4993 images with 70 queries used for evaluation.

The revisited Paris dataset ($\mathcal{R}$Par) is a modified version of Parsi6K dataset: there are 70 query images with 6322 database images.

The Holidays dataset [39] consists of 1491 Holidays pictures, 500 of which are queries. Following standard procedure [2], we manually correct the orientation of the images.

The Motion Picture Experts Group (MPEG) dataset is a collection of 35K images from five image categories: (1) Graphics, (2) Paintings, (3) Video frames, (4) Landmarks and (5) Common objects. A total of 8313 queries are used to measure retrieval performance.

To evaluate system performance in a more challenging scenario, the Holidays, Oxford5k, $\mathcal{R}$Oxf and $\mathcal{R}$Par datasets are augmented with a distractor set containing 1 milion non-matching images ($\mathcal{R}$1M). The retrieval performance for all datasets is computed using mean Average Precision (mAP). We follow the standard protocol for $\mathcal{R}$Oxf and $\mathcal{R}$Par datasets and report results using medium and hard setups referred as $\mathcal{R}$Oxf$M$, $\mathcal{R}$Par$M$, $\mathcal{R}$Oxf$H$ and $\mathcal{R}$Par$H$.

4.3 Impact of activation layer

This experiment investigates the importance of applying activation functions to deep features. The baseline is the Direct aggregation (DA) method where the convolutional layer features are simply aggregated using average pooling. We compare the baseline performance with the proposed SinH aggregation (SinHA), Exponential aggregation (ExpA) and Weibull aggregation (WBA) methods. It can be observed from Figure 5 that it is essential to apply activation functions to the features before aggregation to obtain significant increase in retrieval accuracy. The Weibull aggregation provides average gains of 10.5%, 11.3%, 4.9% and 7.1% on $\mathcal{R}$Oxf$M$, $\mathcal{R}$Oxf$H$, $\mathcal{R}$Par$M$ and $\mathcal{R}$Par$H$, versus Direct aggregation. Furthermore, the WBA network also outperforms ExpA and SinHA on all datasets.

In all the following experiments and comparison with the state-of-the-art, we will use Weibull aggregation with ACTNET.

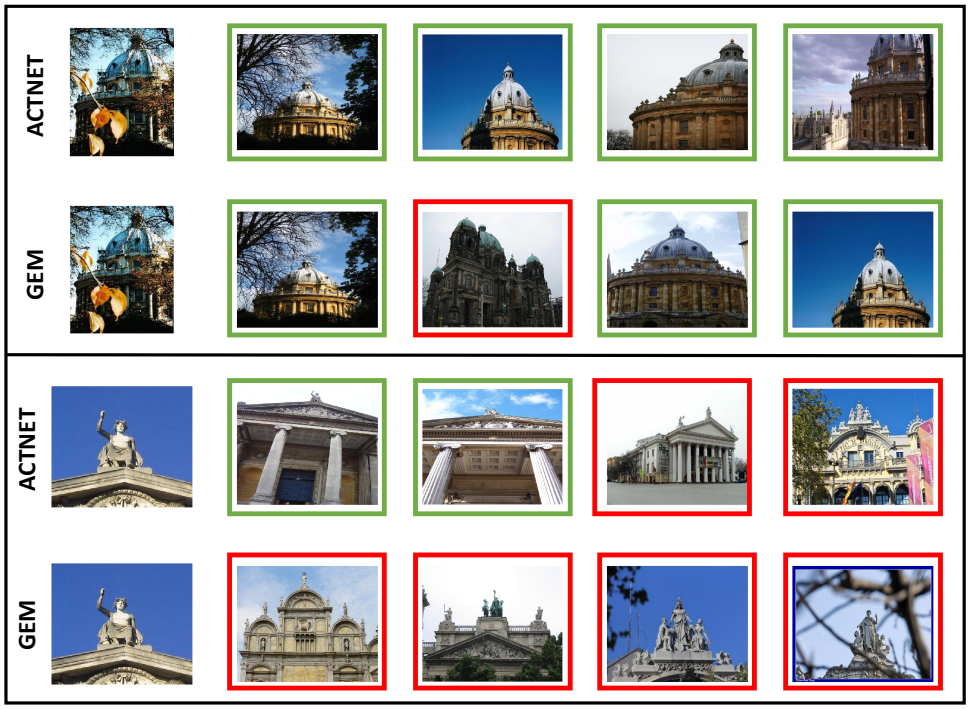


Fig. 7. Top four retrieved images for ACTNET and GEM on $\mathcal{R}$Oxf dataset. On the plot, the correctly retrieved images for a particular query are marked with green border and falsely retrieved images are marked with red border.

4.4 Impact of multi-layer aggregation

This section demonstrates the advantage of aggregating a hierarchy of convolutional features from multiple CNN layers. In ACTNET, deep features from the last two convolutional layers of ResNext101 are transformed using activation layers, aggregated via average pooling and concatenated to form image representation. Importantly, the trainable parameters of the two aggregation streams are optimised explicitly to achieve significant retrieval improvement. It can be observed from Table 1 that multi-layer aggregation (MLA) brings an improvement of +1.0%, +2.5% and 3.1% on Holidays, Oxford5k and MPEG datasets respectively compared to single layer aggregation (SLA), where only the last layer features are used to compute the global descriptor. The gain in retrieval accuracy is even more pronounced on difficult datasets: 6.7% and 2.9% on $\mathcal{R}$Oxf$H$ and $\mathcal{R}$Par$H$ datasets.

4.5 Comparison of GEM and ACTNET

In this section we compare ACTNET with the state-of-the-art GEM network. For a fair assessment, we train both ACTNET and GEM in an end-to-end manner on GLDR dataset with triplet loss. In GEM, the generalised mean pooling layer [3] is added to the last convolutional layer of the ResNext101. The pooling layer is followed by L2-normalisation, PCA+Whitening and L2-normalisation layers.

After training the networks, we extract D-dimensional ACTNET and GEM representations from the test datasets and compute the retrieval performance (Figure 6), as a function of descriptor dimensionality. Compared to GEM, ACTNET offers an average gain of +1.7%, 4.8% and 5.1% on Holidays, Oxford5k and MPEG datasets respectively. The difference in retrieval performance is even more significant on challenging datasets $\mathcal{R}$Oxf$M$ (+8.8%), $\mathcal{R}$Oxf$H$ (+10.2%) and $\mathcal{R}$Par$H$ (+3.4%). On large scale datasets $\mathcal{R}$Oxf$M$+R1M, $\mathcal{R}$Oxf$H$+R1M, $\mathcal{R}$Par$M$+R1M and $\mathcal{R}$Par$H$+R1M dataset, ACTNET achieves 54.8%, 29.9%, 56.3% and 28.6%, outperforming any results published to date. It is also important to note that ACTNET compact signature (128 dimensional) consistently achieves a higher mAP than GEM compact signature on all datasets.

Figure 7 illustrates the top four retrieved images by our ACTNET and baseline GEM on few $\mathcal{R}$Oxf queries. For every query, the correctly retrieved images are marked by a green frame. It can be observed that in both the cases ACTNET is able to retrieve difficult matching images compared to GEM which places them far down in the ranking.

4.6 Performance of low complexity ACTNET

In this section, the retrieval performance of low depth, low latency and small memory Mobile-ACTNET is evaluated. It can be seen from Figure 8 that the performance of Mobile-ACTNET is on average 8% less than ACTNET. However, the descriptor extraction speed of Mobile-ACTNET is approximately five times faster than ACTNET.

We also compare the performance of Mobile-ACTNET with the computationally expensive state-of-the-art ResNet101-GEM [3] and ResNet101-RMAC [2]. We compute single-scale representations of ResNet101-GEM and ResNet101-RMAC using the software provided by the authors. It is interesting to observe from Table 2 that Mobile-ACTNET achieves comparable performance than the aforementioned methods while having five times faster extraction speed.

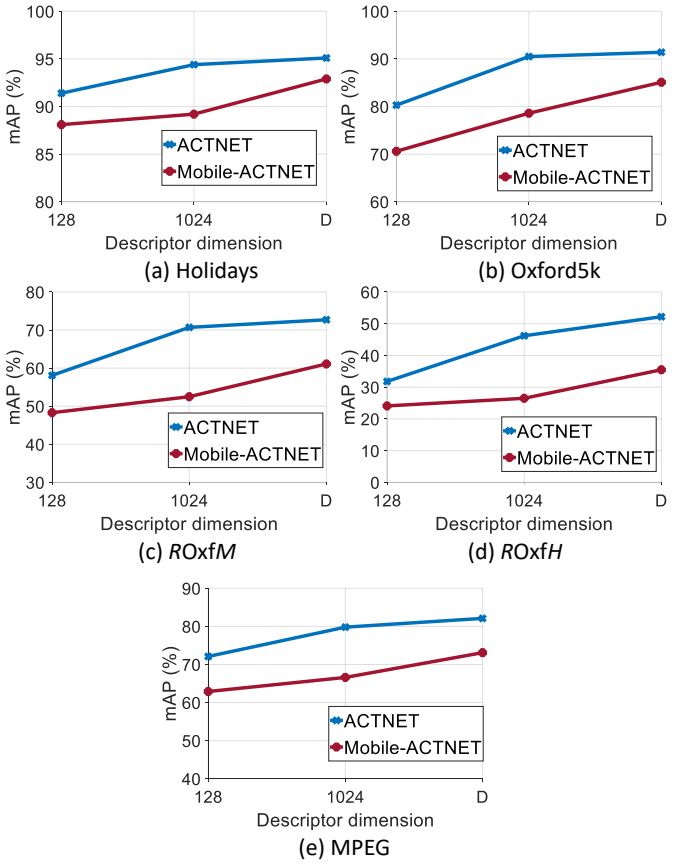


Fig. 8. Low complexity Mobile-ACTNET comparison with ACTNET (a) Holidays, (b) Oxford5k, (c) $\mathcal{R}$Oxf$M$, (d) $\mathcal{R}$Oxf$H$, (e) MPEG (all results in mAP(%)). The dimensionality D of ACTNET and Mobile-ACTNET is 4096 and 2560 respectively.

TABLE 2
Mobile-ACTNET comparison with state-of-the-art ResNet101-GEM and ResNet101-RMAC representations

	Hol	Oxf5k	$\mathcal{R}$Oxf$M$	$\mathcal{R}$Oxf$H$	MPEG
Mobile-ACTNET	92.9	85.1	61.1	35.5	73.1
ResNet101-GEM	92.4	85.7	61.7	35.1	74.1
ResNet101-RMAC	94.0	84.1	60.9	32.4	72.2

5 Comparison with the State-of-the-Art

We extensively compare the performance of ACTNET with the state-of-the-art on compact image signatures and on methods that use query expansion.

The performance of the full dimensional image signatures (1K-4K Dimensions) is summarised in Table 3. In real world scenarios containing billions of images, the use of image representations with dimensionality greater than 1K is prohibitive due to memory requirements and search time; however the results are useful in understanding the maximum capabilities of each signatures.

The proposed ACTNET consistently outperforms the state-of-the-art on all datasets. Compared to RMAC [2], ACTNET provides a significant gain of +5.3%, +11.8%, 1.8% and 9.9% in mAP on Oxford5k, $\mathcal{R}$Oxf$M$, $\mathcal{R}$Par$M$ and MPEG datasets respectively. The difference in mAP points between ACTNET and RMAC is even higher on challenging datasets (+19.8% $\mathcal{R}$Oxf$H$, +15.5% $\mathcal{R}$Oxf$M$+R1M, +17.4% $\mathcal{R}$Oxf$M$+R1M, +3.3% $\mathcal{R}$Par$H$ and +1.5% $\mathcal{R}$Par$M$+R1M). Compared to the best competitor GEM representation [3], ACTNET is more than 13.7, 9.6, 10, 6.4 and 3.9 mAP points ahead on $\mathcal{R}$Oxf$H$, $\mathcal{R}$Oxf$M$+R1M, $\mathcal{R}$Oxf$H$+R1M, $\mathcal{R}$Par$H$ and $\mathcal{R}$Par$H$+R1M datasets respectively. It is important to note that our ACTNET representation is computed from single image scale compared to RMAC [2] and GEM [3] representations where 3 image scales are used. We also compare ACTNET with our implementation of ResNext+GEM and the retrieval performance shows that ACTNET creates significantly more robust and discriminative image signature.

It has recently become common practice to apply Query Expansion (QE) on top of global image signatures to further improve retrieval accuracy. More precisely, we use α query expansion [3] and diffusion (DFS) [40]. It can be observed from Table 4 that ACTNET+αQE achieves significantly better mAP compared to RMAC+αQE and GEM+αQE. Furthermore, ACTNET+DFS achieves 70.7%, 49.1%, 88.9% and 78.4% on $\mathcal{R}$Oxf$M$+R1M, $\mathcal{R}$Oxf$H$+R1M, $\mathcal{R}$Par$M$+R1M and $\mathcal{R}$Par$H$+R1M datasets respectively, outperforming the best published results to date. It is also interesting to note that ACTNET+DFS outperforms the computationally complex DELF-D2R-R-ASMK+SP [31] approach which uses CNN based local features, a codebook of 65k visual words and Spatial Verification (SP).

Finally, we compare the retrieval performance of compact (128 dimensional) image signatures which are practical in real world applications. To compute a compact signature, we forward pass an image through ACTNET to obtain 4096 dimensional representation. The top 128 values from 4096 forms the compact signature. The results presented in Table 5 show that ACTNET significantly outperforms state-of-the-art RMAC and GEM methods. On large scale datasets of $\mathcal{R}$Oxf$M$+R1M, $\mathcal{R}$Oxf$H$+R1M, $\mathcal{R}$Par$M$+R1M and $\mathcal{R}$Par$H$+R1M, ACTNET achieves the best ever performances of 33.1%, 14.8%, 42.3% and 17.6% respectively.

6 Conclusion

In this paper we introduced a novel CNN network architecture called ACTNET. The key innovation is a novel trainable activation layer designed to improve the signal-to-noise ratio in the aggregation stage of deep convolutional features. Our activation layer amplifies a few selected strong filter responses (corresponding to prominent features) thus reducing the impact of weaker features, which effectively constitute background noise. In ACTNET, deep features from final convolutional layers are transformed using the activation function and optimally aggregated into a global descriptor.

The parameters of activation blocks are trained jointly with the CNN filter parameters in an end-to-end manner by minimising the triplet loss. We proposed and evaluated three parametric activation functions: Sine-Hyperbolic, Exponential and modified Weibull, showing that while all bring significant gains, the Weibull achieved the best performance thanks to its equalising properties.

We provided a thorough evaluation on all key benchmarks demonstrating that ACTNET architecture with learnable aggregation generates global representations that significantly outperform the latest state-of-the-art methods,

TABLE 3
Comparison with the state-of-the-art using full dimensional descriptor on Oxford5k, ROxford Medium protocol (ROxf$M$), ROxford Hard protocol (ROxf$H$), RParis Medium protocol (RPar$M$), RParis Hard protocol (RPar$H$), Holidays and MPEG without and with one Million distractors (R1M). All results are computed in terms of mAP(%). The results for MPEG datasets are computed using the software provided by the authors

	Oxford 5k	ROxf$M$	ROxf$H$	ROxf$M$+ R1M	ROxf$H$+ R1M	RPar$M$	RPar$H$	RPar$M$+ R1M	RPar$H$+ R1M	Hol	MPEG
NetVLAD [23]	71.6	37.1	13.8	20.7	6.0	59.8	35.0	31.8	11.5	87.5	65.1
SPoC [41]	68.1	39.8	12.4	21.5	2.8	69.2	44.7	41.6	15.3	83.9	63.2
CroW [21]	70.8	42.4	13.3	21.2	3.3	70.4	47.2	42.7	16.3	85.1	68.2
MAC [1]	80.0	41.7	18.0	24.2	5.7	66.2	44.1	40.8	18.2	85.5	70.6
AGEM [29]	-	67.0	40.7	-	-	78.1	57.3	-	-	-	-
FS.EGM [42]	-	63.0	34.5	-	-	68.7	43.9	-	-	-	-
OS.EGM [42]	-	64.2	35.9	-	-	69.9	46.1	-	-	-	-
ASDA [30]	87.7	66.4	38.5	-	-	71.6	47.9	-	-	-	-
GEM [1]	87.8	64.7	38.5	45.2	19.9	77.2	56.3	52.3	24.7	93.9	74.1
RMAC [1]	86.1	60.9	32.4	39.3	12.5	78.9	59.4	54.8	28.0	94.8	72.2
ResNext+GEM	87.2	64.5	38.9	45.2	19.8	78.8	59.5	54.6	26.4	93.7	76.2
ACTNET	**91.4**	**72.7**	**52.2**	**54.8**	**29.9**	**80.7**	**62.7**	**56.3**	**28.7**	**95.1**	**82.1**

TABLE 4
Performance evaluation of full dimensional image signatures using Query Expansion

	ROxf$M$	ROxf$H$	ROxf$M$+ R1M	ROxf$H$+ R1M	RPar$M$	RPar$H$	RPar$M$+ R1M	RPar$H$+ R1M
DELF-D2R-R-ASMK+SP [31]	71.9	48.5	-	-	78.0	54.0	-	-
DELF-GLD-ASMK+SP [31]	76.0	52.4	-	-	80.2	58.6	-	-
GEM+αQE [1]	67.2	40.8	49.0	24.2	80.7	61.8	58.0	31.0
RMAC+αQE [1]	64.8	36.8	45.7	19.5	82.7	65.7	61.0	35.0
ResNext-GEM+αQE	70.4	43.5	51.5	25.3	82.6	65.5	62.1	35.5
ACTNET+αQE	77.2	56.6	61.5	36.3	84.6	68.6	65.6	40.9
GEM+DFS [1]	69.8	40.5	61.5	33.1	88.9	78.5	84.9	71.6
RMAC+DFS [1]	69.0	44.7	56.6	28.4	89.5	80.0	83.2	70.4
ACTNET+DFS	**78.3**	**57.1**	**70.7**	**49.1**	**91.1**	**83.1**	**88.9**	**78.4**

TABLE 5
Comparison with the state-of-the-art using small dimensional descriptor (128-Dim). The results for RMAC and GEM are computed using the software provided by authors

	Dim	Holidays	Oxford5k	ROxf$M$	ROxf$H$	ROxf$M$+ R1M	ROxf$H$+ R1M	RPar$M$	RPar$H$	RPar$M$+ R1M	RPar$H$+ R1M
RMAC [2]	128	88.5	77.9	46.2	21.1	21.9	7.4	72.3	48.1	40.9	15.6
GEM [3]	128	85.9	79.5	49.9	26.2	27.8	10.9	72.2	47.5	40.8	15.5
ACTNET	128	**91.4**	**80.3**	**58.1**	**31.8**	**33.1**	**14.8**	**74.1**	**50.1**	**42.3**	**17.6**

including those based on computationally costly local descriptor indexing and spatial verification. We also show that ACTNET retains its leading performance when using short codes of 128 bytes and when applied to low-complexity CNN.

REFERENCES

[1] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 5706–5715.

[2] J. R. J. Gordo, Albertand Almazán and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, Sep 2017.

[3] F. Radenovi, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[4] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *CoRR*, 2015.

[5] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 4510–4520.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, pp. 91–110, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, pp. 346–359, 2008.

[9] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 1470–1477.

[10] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors." in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3384–3391.

[11] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1704–1716, Sep 2012.

[12] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[13] S. S. Husain and M. Bober, "Improving large-scale image retrieval through robust aggregation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1783–1796, Sept 2017.

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

[15] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[16] O. Chum, A. Mikulk, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR 2011*, June 2011, pp. 889–896.

[17] G. Tolias and H. Jégou, "Visual query expansion with or without geometry: Refining local descriptors by feature aggregation," *Pattern Recognition*, vol. 47, pp. 3466–3476, 2014.

[18] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1790–1802, Sept 2016.

[19] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky, "Neural codes for image retrieval," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 2014, pp. 584–599.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014.

[21] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *ECCV Workshops*, 2016.

[22] E. Ong, S. Husain, and M. Bober, "Siamese network of deep fisher-vector descriptors for image retrieval," *CoRR*, vol. abs/1702.00338, 2017. [Online]. Available: http://arxiv.org/abs/1702.00338

[23] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, June 2018.

[24] O. Seddati, S. Dupont, S. Mahmoudi, and M. Parian, "Towards good practices for image retrieval based on cnn features," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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

[26] J. Xu, C. Wang, C. Qi, C. Shi, and B. Xiao, "Unsupervised semantic-based aggregation of deep convolutional features," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 601–611, 2019.

[27] S. Pang, J. Xue, J. Zhu, L. Zhu, and Q. Tian, "Unifying sum and weighted aggregations for efficient yet effective image representation computation," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 841–852, 2019.

[28] X. Wu, G. Irie, K. Hiramatsu, and K. Kashino, "Weighted generalized mean pooling for deep image retrieval," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 495–499.

[29] Y. Gu, C. Li, and J. Xie, "Attention-aware generalized mean pooling for image retrieval," *CoRR*, 2018.

[30] J. Xu, C. Wang, C. Shi, and B. Xiao, "Weakly supervised soft-detection-based aggregation method for image retrieval," *CoRR*, 2018.

[31] M. Teichmann, A. Araujo, M. Zhu, and J. Sim, "Detect-to-retrieve: Efficient regional aggregation for image search," *CoRR*, 2018.

[32] S. S. Husain and M. Bober, "Remap: Multi-layer entropy-guided pooling of dense cnn features for image retrieval," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.

[33] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 2146–2153.

[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, 13–15 May 2010, pp. 249–256.

[35] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3, 2013, pp. 1319–1327.

[36] S. Qian, H. Liu, C. Liu, S. Wu, and H. S. Wong, "Adaptive activation functions in convolutional neural networks," *Neurocomputing*, vol. 272, pp. 204 – 212, 2018.

[37] F. Agostinelli, M. D. Hoffman, P. J. Sadowski, and P. Baldi, "Learning activation functions to improve deep neural networks," *CoRR*, vol. abs/1412.6830, 2014.

[38] H. Noh, A. R. T. S. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3476–3485, 2017.

[39] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, feb 2010.

[40] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations," in *CVPR*, 2017.

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

[42] O. Siméoni, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Graph-based particular object discovery," *Machine Vision and Applications*, vol. 30, pp. 243–254, Mar 2019.

Syed Sameed Husain is a Research Fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom. He received MSc and PhD degrees from University of Surrey, in 2011 and 2016, respectively. His research interests include machine learning, computer vision, deep learning and image retrieval. Sameed's team has recently won the prestigious Google Landmark Retrieval Challenge.

Eng-Jon Ong received the computer science degree in 1997 and the PhD degree in computer vision in 2001 from Queen Mary, University of London. Following that, he joined the Centre for Vision, Speech and Signal Processing at the University of Surrey as a researcher. His main interests are in visual feature tracking, data mining, pattern recognition, and machine learning methods.

Miroslaw Bober is a Professor of Video Processing at the University of Surrey, U.K. In 2011 he co-founded Visual Atoms Ltd, a company specializing in visual analysis and search technologies. Between 1997 and 2011 he headed Mitsubishi Electric Corporate R&D Center Europe (MERCE-UK). He received BSc degree from AGH University of Science and Technology, and MSc and PhD degrees from University of Surrey. His research interests include computer vision, machine learning and AI, with a focus on analysis and understanding of visual and multimodal data, and efficient representation of its semantic content. Miroslaw led the development of ISO MPEG standards for over 20 years, chairing the MPEG-7, CDVS and CVDA groups. He is an inventor of over 80 patents, many deployed in products. His publication record includes over 100 refereed publications, including three books and book chapters, and his visual search technologies recently won the Google Landmark Retrieval Challenge on Kaggle.