Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval

Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, Member, IEEE, Zhi-Hua Zhou, Fellow, IEEE

Abstract—Deep convolutional neural network models pre-trained for the ImageNet classification task have been successfully adopted to tasks in other domains, such as texture description and object proposal generation, but these tasks require annotations for images in the new domain. In this paper, we focus on a novel and challenging task in the pure unsupervised setting: fine-grained image retrieval. Even with image labels, fine-grained images are difficult to classify, let alone the unsupervised retrieval task. We propose the Selective Convolutional Descriptor Aggregation (SCDA) method. SCDA firstly localizes the main object in fine-grained images, a step that discards the noisy background and keeps useful deep descriptors. The selected descriptors are then aggregated and dimensionality reduced into a short feature vector using the best practices we found. SCDA is unsupervised, using no image label or bounding box annotation. Experiments on six fine-grained datasets confirm the effectiveness of SCDA for fine-grained image retrieval. Besides, visualization of the SCDA features shows that they correspond to visual attributes (even subtle ones), which might explain SCDA’s high mean average precision in fine-grained retrieval. Moreover, on general image retrieval datasets, SCDA achieves comparable retrieval results with state-of-the-art general image retrieval approaches.

Index Terms—Fine-grained image retrieval, selection and aggregation, unsupervised object localization.

I. INTRODUCTION

AFTER the breakthrough in image classification using Convolutional Neural Networks (CNN) [1], pre-trained CNN models trained for one task (e.g., recognition or detection) have also been applied to domains different from their original purposes (e.g., for describing texture [2] or finding object proposals [3]). Such adaptations of pre-trained CNN models, however, still require further annotations in the new domain (e.g., image labels). In this paper, we show that for fine-grained images which contain only subtle differences among categories (e.g., varieties of dogs), pre-trained CNN models can both localize the main object and find images in the same variety. Since no supervision is used, we call this novel and challenging task fine-grained image retrieval.

In fine-grained image classification [4], [5], [6], [7], [8], [9], categories correspond to varieties in the same species. The categories are all similar to each other, only distinguished by slight and subtle differences. Therefore, an accurate system usually requires strong annotations, e.g., bounding boxes for object or even object parts. Such annotations are expensive and unrealistic in many real applications. In answer to this difficulty, there are attempts to categorize fine-grained images with only image-level labels, e.g., [6], [7], [8], [9].

In this paper, we handle a more challenging but more realistic task, i.e., Fine-Grained Image Retrieval (FGIR). In FGIR, given database images of the same species (e.g., birds, flowers or dogs) and a query, we should return images which are in the same variety as the query, without resorting to any other supervision signal. FGIR is useful in applications such as biological research and bio-diversity protection. As illustrated in Fig. 1, FGIR is also different from general-purpose image retrieval. General image retrieval focuses on retrieving near-duplicate images based on similarities in their contents (e.g., textures, colors and shapes), while FGIR focuses on retrieving the images of the same types (e.g., the same species for the animals and the same model for the cars). Meanwhile, objects in fine-grained images have only subtle differences, and vary in poses, scales and rotations.

To meet these challenges, we propose the Selective Convolutional Descriptor Aggregation (SCDA) method, which
automatically localizes the main object in fine-grained images and extracts discriminative representations for them. In SCDA, only a pre-trained CNN model (from ImageNet which is not fine-grained) is used and we use absolutely no supervision. As shown in Fig. 2, the pre-trained CNN model first extracts convolution activations for an input image. We propose a novel approach to determine which part of the activations are useful (i.e., to localize the object). These useful descriptors are then aggregated and dimensionality reduced to form a vector representation using practices we propose in SCDA. Finally, a nearest neighbor search ends the FGIR process.

We conducted extensive experiments on six popular fine-grained datasets (CUB200-2011 [10], Stanford Dogs [13], Oxford Flowers 102 [14], Oxford-IIIT Pets [15], Aircrafts [16] and Cars [11]) for image retrieval. Moreover, we also tested the proposed SCDA method on standard general-purpose retrieval datasets (INRIA Holiday [17] and Oxford Building 5K [12]). In addition, we report the classification accuracy of the SCDA method, which only uses the image labels. Both retrieval and classification experiments verify the effectiveness of SCDA. The key advantages and major contributions of our method are:

- We propose a simple yet effective approach to localize the main object. This localization is unsupervised, without utilizing bounding boxes, image labels, object proposals, or additional learning. SCDA selects only useful deep descriptors and removes background or noise, which benefits the retrieval task.
- With the ensemble of multiple CNN layers and the proposed dimensionality reduction practice, SCDA has shorter but more accurate representation than existing deep learning based methods (cf. Sec. IV). For fine-grained images, as presented in Table III, SCDA achieves the best retrieval results. Furthermore, SCDA also has accurate results on general-purpose image retrieval datasets, cf. Table V.
- As shown in Fig. 8, the compressed SCDA feature has stronger correspondence to visual attributes (even subtle ones) than the deep activations, which might explain the success of SCDA for fine-grained tasks.

Moreover, beyond the specific fine-grained image retrieval task, our proposed method could be treated as one kind of transfer learning, i.e., a model trained for one task (image classification on ImageNet) is used to solve another different task (fine-grained image retrieval). It indeed reveals the reusability of deep convolutional neural networks.

The rest of this paper is organized as follows. Sec. II introduces the related work about general deep image retrieval and fine-grained image tasks. The details of the proposed SCDA method are presented in Sec. III. In Sec. IV, for fine-grained image retrieval, we compare our method with several baseline approaches and three state-of-the-art general deep image retrieval approaches. Moreover, discussion on the quality of the SCDA feature is illustrated. Sec. V concludes the paper.

II. RELATED WORK

We will briefly review two lines of related work: deep learning approaches for image retrieval and research on fine-grained images.

A. Deep Learning for Image Retrieval

Until recently, most image retrieval approaches were based on local features (with SIFT being a typical example) and feature aggregation strategies on top of these local features. Vector of Locally Aggregated Descriptors (VLAD) [18] and Fisher Vector (FV) [19] are two typical feature aggregation strategies. After the success of CNN [1], image retrieval also embraced deep learning. Out-of-the-box features from pre-trained deep networks were shown to achieve state-of-the-art results in many vision related tasks, including image retrieval [20].

Some efforts (e.g., [21], [22], [23], [24], [25], [26], [27]) studied what deep descriptors can be used and how to use them in image retrieval, and have achieved satisfactory results. In [21], to improve the invariance of CNN activations without degrading their discriminative ability, they proposed the multi-scale orderless pooling (MOP-CNN) method. MOP-CNN firstly extracts CNN activations from the fully connected layers for local patches at multiple scale levels, and performed orderless VLAD pooling of these activations at each level separately, and finally concatenated the features. After that, [22] has extensively evaluated the performance of such features with and without fine-tuning on related dataset. This work has shown that PCA-compressed deep features can outperform compact descriptors computed on traditional SIFT-like features. Later, [23] found that using sum-pooling to aggregate deep features on the last convolutional layer leads to better performance, and proposed the sum-pooled convolutional (SPOC) features. Based on that, [25] applied weighting both spatially and per channel before sum-pooling to create a final aggregation. [27] proposed a compact image representation derived from the convolutional layer activations that encodes multiple image regions without the need to refeed multiple inputs to the network. Very recently, the authors of [26] investigated several effective usages of CNN activations on both image retrieval and classification. In particular, they aggregated activations of each layer and concatenated them into the final representation, which achieved satisfactory results.

However, these approaches directly used the CNN activations/descriptors and encoded them into a single representation, without evaluating the usefulness of the obtained deep descriptors. In contrast, our proposed SCDA method can select only useful deep descriptors and remove background or noise by localizing the main object unsupervisedly. Meanwhile, we have also proposed several good practices of SCDA for retrieval tasks. In addition, the previous deep learning based image retrieval approaches were all designed for general image retrieval, which is quite different from fine-grained image retrieval. As will be shown by our experiments, state-of-the-art general image retrieval approaches do not work well for the fine-grained image retrieval task.
Additionally, several variants of image retrieval were studied in the past few years, e.g., multi-label image retrieval [28], sketch-based image retrieval [29] and medical CT image retrieval [30]. In this paper, we will focus on the novel and challenging fine-grained image retrieval task.

B. Fine-Grained Image Tasks

Fine-grained classification has been popular in the past few years, and a number of effective fine-grained recognition methods have been developed in the literature [4], [5], [6], [7], [8], [9].

We can roughly categorize these methods into three groups. The first group, e.g., [31], [8], attempted to learn a more discriminative feature representation by developing powerful deep models for classifying fine-grained images. The second group aligned the objects in fine-grained images to eliminate pose variations and the influence of camera position, e.g., [5]. The last group focused on part-based representations. However, because it is not realistic to obtain strong annotations (object bounding boxes and/or part annotations) for a large number of images, more algorithms attempted to classify fine-grained images using only image-level labels, e.g., [6], [7], [8], [9].

All the previous fine-grained classification methods needed image-level labels (others even needed part annotations) to train their deep networks. Few works have touched unsupervised retrieval of fine-grained images. Wang et al. [32] proposed Deep Ranking to learn similarity between fine-grained images. However, it requires image-level labels to build a set of triplets, which is not unsupervised and cannot scale well for large scale image retrieval tasks.

One related research to FGIR is [33]. The authors of [33] proposed the fine-grained image search problem. [33] used the bag-of-word model with SIFT features, while we use pre-trained CNN models. Beyond this difference, a more important difference is how the database is constructed.

[33] constructed a hierarchical database by merging several existing image retrieval datasets, including fine-grained datasets (e.g., CUB200-2011 and Stanford Dogs) and general image retrieval datasets (e.g., Oxford Buildings and Paris). Given a query, [33] first determines its meta class, and then does a fine-grained image search if the query belongs to the fine-grained meta category. In FGIR, the database contains images of one single species, which is more suitable in fine-grained applications. For example, a bird protection project may not want to find dog images given a bird query. To our best knowledge, this is the first attempt to fine-grained image retrieval using deep learning.

III. SELECTIVE CONVOLUTIONAL DESCRIPTOR AGGREGATION

In this section, we propose the Selective Convolutional Descriptor Aggregation (SCDA) method. Firstly, we will introduce the notations used in this paper. Then, we present the descriptor selection process, and finally, the feature aggregation details will be described.

A. Preliminary

The following notations are used in the rest of this paper. The term “feature map” indicates the convolutional result of one channel; the term “activations” indicates feature maps of all channels in a convolution layer; and the term “descriptor” indicates the d-dimensional component vector of activations. “pool” refers to the activations of the max-pooled last convolution layer, and “fc” refers to the activations of the last fully connected layer.

Given an input image $I$ of size $H \times W$, the activations of a convolution layer are formulated as an order-3 tensor $T$ with $h \times w \times d$ elements, which include a set of 2-D feature maps $S = \{S_n\} (n = 1, \ldots, d)$. $S_n$ of size $h \times w$ is the $n$-th feature map of the corresponding channel (the $n$-th channel). From another point of view, $T$ can be also considered as having $h \times w$ cells and each cell contains one $d$-dimensional deep descriptor. We denote the deep descriptors as $X = \{x_{(i,j)}\}$, where $(i,j)$ is a particular cell ($i \in \{1, \ldots, h\}, j \in \{1, \ldots, w\}, x_{(i,j)} \in \mathbb{R}^d$). For instance, by employing the popular pre-trained VGG-16 model [34] to extract deep descriptors, we can get a $7 \times 7 \times 512$ activation tensor in pool5 if the input image is $224 \times 224$. Thus, on one hand, for this image, we have 512 feature maps (i.e., $S_n$) of size $T \times 7$; on the other hand, 49 deep descriptors of 512-d are also obtained.
B. Selecting Convolutional Descriptors

What distinguishes SCDA from existing deep learning-based image retrieval methods is: using only the pre-trained model, SCDA is able to find useful deep convolutional features, which in effect localizes the main object in the image and discards irrelevant and noisy image regions. Note that the pre-trained model is not fine-tuned using the target fine-grained dataset. In the following, we propose our descriptor selection method, and then present quantitative and qualitative localization results.

1) Descriptor Selection: After obtaining the pool$_5$ activations, the input image $I$ is represented by an order-3 tensor $T$, which is a sparse and distributed representation [35], [36]. The distributed representation argument claims that concepts are encoded by a distributed pattern of activities spread across multiple neurons [37]. In deep neural networks, a distributed representation means a many-to-many relationship between two types of representations (i.e., concepts and neurons): Each concept is represented by a pattern of activity distributed over many neurons, and each neuron participates in the representation of many concepts [35], [36].

In Fig. 3, we show some images taken from five fine-grained datasets, CUB200-2011 [10], Stanford Dogs [13], Oxford Flowers 102 [14], Aircrafts [16] and Cars [11]. We randomly sample several feature maps from the 512 feature maps in pool$_5$ and overlay them to original images for better visualization. As can be seen from Fig. 3, the activated regions of the sampled feature map (highlighted in warm color) may indicate semantically meaningful parts of birds/dogs/flowers/aircrafts/cars, but can also indicate some background or noisy parts in these fine-grained images.

In addition, the semantic meanings of the activated regions are quite different even for the same channel. For example, in the 464th feature map for birds on the right side, the activated region in the first image indicates the Pine Warbler’s tail and the second does the Black-capped Vireo’s head. In the 274th feature map for dogs, the first indicates the German Shepherd’s head, while the second even has no activated region for the Cockapoo, except for a part of noisy background. The other examples of flowers, aircrafts and cars have the same characteristics. In addition, there are also some activated regions representing the background, e.g., the 19th feature map for Pine Warbler and the 418th one for German Shepherd. Fig. 3 conveys that not all deep descriptors are useful, and one single channel contains at best weak semantic information due to the distributed nature of this representation. Therefore, selecting and using only useful deep descriptors (and removing noise) is necessary. However, in order to decide which deep descriptor is useful (i.e., containing the object we want to retrieve), we cannot count on any single channel individually.

We propose a simple yet effective method (shown in Fig. 2), and its quantitative and qualitative evaluation will be demonstrated in the next section. Although one single channel is not very useful, if many channels fire at the same region, we could expect this region to be an object rather than the background. Therefore, in the proposed method, we add up the obtained pool$_5$ activation tensor through the depth direction. Thus, the $h \times w \times d$ 3-D tensor becomes an $h \times w$ 2-D tensor, which we call the “aggregation map”, i.e., $A = \sum_{n=1}^{d} S_n$ (where $S_n$ is the $n$-th feature map in pool$_5$). For the aggregation map $A$, there are $h \times w$ summed activation responses, corresponding to $h \times w$ positions. Based on the aforementioned observation, it is straightforward to say that the higher activation response of a particular position $(i,j)$, the more possibility of its corresponding region being part of the object. Additionally, fine-grained image retrieval is an unsupervised problem, in which we have no prior knowledge of how to deal with it. Consequently, we calculate the mean value $\bar{a}$ of all the positions in $A$ as the threshold to decide which positions localize objects: the position $(i,j)$ whose activation response is higher than $\bar{a}$ indicates the main object, e.g., birds, dogs or aircrafts, might appear in that position. A mask map $M$ of the same size as $A$ can be obtained as:

$$M_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} > \bar{a} \\ 0 & \text{otherwise} \end{cases},$$

(1)

where $(i,j)$ is a particular position in these $h \times w$ positions.

In Fig. 3, the figures in the second last column for each fine-grained datasets show some examples of the mask maps for birds, dogs, flowers, aircrafts and cars, respectively. For these figures, we first resize the mask map $M$ using the bicubic interpolation, such that its size is the same as the input image. We then overlay the corresponding mask map (highlighted in red) onto the original images. Even though the proposed method does not train on these datasets, the main objects (e.g., birds, dogs, aircrafts or cars) can be roughly detected. However, as can be seen from these figures, there are still several small noisy parts activated on a complicated background. Fortunately, because the noisy parts are usually smaller than the main object, we employ Algorithm 1 to collect the largest connected component of $M$, which is denoted as $\hat{M}$, to get rid of the interference caused by noisy parts. In the last column, the main objects are kept by $\hat{M}$, while the noisy parts are discarded, e.g., the plant, the cloud and the grass.

Therefore, we use $\hat{M}$ to select useful and meaningful deep convolutional descriptors. The descriptor $x_{(i,j)}$ should be kept when $\hat{M}_{i,j} = 1$, while $\hat{M}_{i,j} = 0$ means the position $(i,j)$ might have background or noisy parts:

$$F = \left\{ x_{(i,j)} | \hat{M}_{i,j} = 1 \right\},$$

(2)

<table>
<thead>
<tr>
<th>Algorithm 1 Finding connected components in binary images</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong> A binary image $I$;</td>
</tr>
<tr>
<td>1: Select one pixel $p$ as the starting point;</td>
</tr>
<tr>
<td>2: while True do</td>
</tr>
<tr>
<td>3: Use a flood-fill algorithm to label all the pixels in</td>
</tr>
<tr>
<td>the connected component containing $p$;</td>
</tr>
<tr>
<td>4: if All the pixels are labeled then</td>
</tr>
<tr>
<td>5: Break;</td>
</tr>
<tr>
<td>6: end if</td>
</tr>
<tr>
<td>7: Search for the next unlabeled pixel as $p$;</td>
</tr>
<tr>
<td>8: end while</td>
</tr>
<tr>
<td>9: return Connectivity of the connected components, and</td>
</tr>
<tr>
<td>their corresponding size (pixel numbers).</td>
</tr>
</tbody>
</table>


Figure 3. Sampled feature maps of fine-grained images from five fine-grained datasets (CUB200-2011, Stanford Dogs, Oxford Flowers, Aircrafts and Cars). Although we resize the images for better visualization, our method can deal with images of any resolution. The first column of each subfigure are the input images, and the randomly sampled feature maps are the following four columns. The last two columns are the mask maps $M$ and the corresponding largest connected component $\tilde{M}$. The selected regions are highlighted in red with the black boundary. (The figure is best viewed in color.)

where $F$ stands for the selected descriptor set, which will be aggregated into the final representation for retrieving fine-grained images. The whole convolutional descriptor selection process is illustrated in Fig. 2b-2e.

2) Qualitative Evaluation: In this section, we give the qualitative evaluation of the proposed descriptor selection process. Because four fine-grained datasets (i.e., CUB200-2011, Stanford Dogs, Aircrafts and Cars) supply the ground-truth bounding box for each image, it is desirable to evaluate the proposed method for object localization. However, as seen in Fig. 3, the detected regions are irregularly shaped. So, the minimum rectangle bounding boxes which contain the detected regions are returned as our object localization predictions.

We evaluate the proposed method to localize the whole-object (birds, dogs, aircrafts or cars) on their test sets. Example predictions can be seen in Fig. 4. From these figures, the predicted bounding boxes approximate the ground-truth bounding boxes fairly accurately, and even some results are better than the ground truth. For instance, in the first dog image shown in Fig. 4, the predicted bounding box can cover both dogs; and in the third one, the predicted box contains less background, which is beneficial to retrieval performance. Moreover, the predicted boxes of Aircrafts and Cars are almost identical to the ground-truth bounding boxes in many cases. However, since we utilize no supervision, some details of the fine-grained objects, e.g., birds’ tails, cannot be contained accurately by the predicted bounding boxes.

3) Quantitative Evaluation: We also report the results in terms of the Percentage of Correctly Localized Parts (PCP) metric for object localization in Table I. The reported metrics are the percentage of whole-object boxes that are correctly localized with a $>50\%$ IOU with the ground-truth bounding boxes. In this table, for CUB200-2011, we show the PCP results of two fine-grained parts (i.e., head and torso) reported in some previous part localization based fine-grained classification algorithms [38], [4], [5]. Here, we first compare the whole-object localization rates with that of fine-grained parts for a rough comparison. In fact, the torso bounding box is highly similar to that of the whole-object in CUB200-2011.
By comparing the results of PCP for torso and our whole-object, we find that, even though our method is unsupervised, the localization performance is just slightly lower or even comparable to that of these algorithms using strong supervisions, e.g., ground-truth bounding box and parts annotations (even in the test phase). For Stanford Dogs, our method can get 78.86% object localization accuracy. Moreover, the results of Aircrafts and Cars are 94.91% and 90.96%, which validates the effectiveness of the proposed unsupervised object localization method.

Additionally, in our proposed method, the largest connected component \( \tilde{M} \) of the obtained mask map is kept. We further investigate how this filtering step affects object localization performance by removing this processing. Then, based on \( M \), the object localization results based on these datasets are: 45.18%, 68.67%, 59.83% and 79.36% for CUB200-2011, Stanford Dogs, Aircrafts and Cars, respectively. The localization accuracy based on \( M \) is much lower than the accuracy based on \( \tilde{M} \), which proves the effectiveness of obtaining the largest connected component. Besides, we also consider these drops through the relation to the size of the ground truth bounding boxes. From this point of view, Fig. 5 shows the percentage of the whole images covered by the ground truth bounding boxes on four fine-grained datasets, respectively. It is obvious that most ground truth bounding boxes of CUB200-2011 and Aircrafts are less than 50% size of the whole images. Thus, for the two datasets, the drops are large. However, for Cars, as shown in Fig. 5d, the percentage’s distribution approaches a normal distribution. For Stanford Dogs, a few ground truth bounding boxes cover less than 20% image size or covering more than 80% image size. Therefore, for these two datasets, the effect of removing the largest connected component processing could be small.

What’s more, because our method does not require any supervision, a state-of-the-art unsupervised object localization method, i.e., [39], is conducted as the baseline. [39] uses off-the-shelf region proposals to form a set of candidate bounding boxes for objects. Then, these regions are matched across images using a probabilistic Hough transform that evaluates the confidence for each candidate correspondence considering both appearance and spatial consistency. After that, dominant objects are discovered and localized by comparing the scores of candidate regions and selecting those that stand out over other regions containing them. As [39] is not a deep learning based method, most of its localization results on these fine-grained datasets are not satisfactory, which are reported in Table 1. Specifically, for many images of Aircrafts, [39] returns the whole images as the corresponding bounding boxes predictions. While, as shown in Fig. 5c, only a small percentage of ground truth bounding boxes approach the whole images, which could explain why the unsupervised localization accuracy on Aircrafts of [39] is much worse than ours.
### C. Aggregating Convolutional Descriptors

After the selection process, the selected descriptor set
\( F = \{ x_{(i,j)} | M_{i,j} = 1 \} \) is obtained. In the following, we compare several encoding or pooling approaches to aggregate these convolutional features, and then give our proposal.

- **Vector of Locally Aggregated Descriptors (VLAD)** [18] is a popular encoding approach in computer vision. VLAD uses \( k\)-means to find a codebook of \( K \) centroids \( \{ c_1, \ldots, c_K \} \) and maps \( x_{(i,j)} \) into a single vector \( v_{(i,j)} = [0 \ldots 0 x_{(i,j)} - c_k \ldots 0] \in \mathbb{R}^{K \times d} \), where \( c_k \) is the closest centroid to \( x_{(i,j)} \). The final representation of \( F \) is
\[
\sum_{i,j} v_{(i,j)}. \]

- **Fisher Vector** (FV) [19]. The encoding process of FV is similar to VLAD. But it uses a soft assignment (i.e., Gaussian Mixture Model) instead of using \( k\)-means for pre-computing the codebook. Moreover, FV also includes second-order statistics.

- **Pooling approaches**. We also try two traditional pooling approaches, i.e., global average-pooling and max-pooling, to aggregate the deep descriptors, i.e.,
\[
\begin{align*}
    p_{\text{avg}} &= \frac{1}{N} \sum_{i,j} x_{(i,j)}, \\
    p_{\text{max}} &= \max_{i,j} x_{(i,j)},
\end{align*}
\]
where \( p_{\text{avg}} \) and \( p_{\text{max}} \) are both \( 1 \times d \) dimensional. \( N \) is the number of the selected descriptors.

After encoding or pooling the selected descriptor set \( F \) into a single vector, for VLAD and FV, the square root normalization and \( \ell_2 \)-normalization are followed; for max- and average-pooling methods, we do \( \ell_2 \)-normalization (the square root normalization did not work well). Finally, the cosine similarity is used for nearest neighbor search. We use two datasets to demonstrate which type of aggregation method is optimal for fine-grained image retrieval. The original training and testing splits provided in the datasets are used. Each image in the testing set is treated as a query, and the training images are regarded as the gallery. The top-\( k \) MAP retrieval performance is reported in Table II.

For the parameter choice of VLAD/FV, we follow the suggestions reported in [40]. The number of clusters in VLAD and the number of Gaussian components in FV are both set to 128. As shown in the table, larger values lead to lower accuracy. Moreover, we find the simpler aggregation methods such as global max- and average-pooling achieve better retrieval performance compared with the high-dimensional encoding approaches. These observations are also consistent with the findings in [25] for general image retrieval. The reason why VLAD and FV do not work well in this case is related to the rather small number of deep descriptors that need to be aggregated. The average number of deep descriptors selected per image for CUB200-2011 and Stanford Dogs is 40.12 and 46.74, respectively. Then, we propose to concatenate the global max-pooling and average-pooling representations, “avg&maxPool”, as our aggregation scheme. Its performance is significantly and consistently higher than the others. We use the “avg&maxPool” aggregation as “SCDA feature” to represent the whole fine-grained image.

### D. Multiple Layer Ensemble

As studied in [41], [42], the ensemble of multiple layers boosts the final performance. Thus, we also incorporate another SCDA feature produced from the relu5-2 layer which is three layers in front of pool5 in the VGG-16 model [34].

Following pool5, we get the mask map \( M_{\text{relu5-2}} \) from relu5-2. Its activations are less related to the semantic meaning than those of pool5. As shown in Fig. 6 (c), there are many noisy parts. However, the bird is more accurately detected than pool5. Therefore, we combine \( M_{\text{pool5}} \) and \( M_{\text{relu5-2}} \) together to get the final mask map of relu5-2, \( \tilde{M}_{\text{pool5}} \) is firstly upsampled to the size of \( M_{\text{relu5-2}} \). We keep the descriptors when their...
The ground truth bounding boxes to replace the whole images. As shown in Table III, because the ground truth bounding boxes of these fine-grained images just contain the main objects, “SIFT FV gtBBBox” achieves significantly better performance than that of the whole images.

For the \( f \) baseline, because it requires the input images at a fixed size, the original images are resized to \( 224 \times 224 \) and then fed into VGG-16. Similar to the SIFT baseline, we also feed the ground truth bounding boxes to replace the whole images. The \( f \) feature of the ground truth bounding box achieves better performance. Moreover, the retrieval results of the \( f \) feature using the bounding boxes predicted by our method are also shown in Table III, which are slightly lower than the ground-truth ones. This observation validates the effectiveness of our method’s object localization once again.

For the \( p \) baseline, the \( p \) descriptors are extracted directly without any selection process. We pool them by both average- and max-pooling, and concatenate them into a 1,024-d representation. As shown in Table III, the performance of \( p \) is better than “fc8_im”, but much worse than the proposed SCDA feature. In addition, VLAD and FV is employed to encode the selected deep descriptors, and we denote the two methods as “selectVLAD” and “selectFV” in Table III. The features of selectVLAD and selectFV have larger dimensionality, but lower mAP in the retrieval task.

State-of-the-art general image retrieval approaches, e.g., SPoC, CroW and R-MAC, can not get satisfactory results for fine-grained images. Hence, general deep learning image retrieval methods could not be directly applied to FGIR.

We also report the results of SCDA\(^{+}\) and SCDA\(_{flip}\)\(^{+}\) on these six fine-grained datasets in Table III. In general,
SCDA\textsubscript{flip}\textsuperscript{+} is the best amongst the compared methods. Comparing these results with the ones of SCDA, we find the multiple layer ensemble strategy (cf. Sec. III-D) could improve the retrieval performance, and furthermore horizontal flip boosts the performance significantly. Therefore, if your retrieval tasks prefer a low dimensional feature representation, SCDA is the optimal choice, or, the post-processing on SCDA\textsubscript{flip}\textsuperscript{+} features is recommended.

1) Post-Processing: In the following, we compare several feature compression methods on the SCDA\textsubscript{flip}\textsuperscript{+} feature: (a) Singular Value Decomposition (SVD); (b) Principal Component Analysis (PCA); (c) PCA whitening (its results were much worse than other methods and are omitted) and (d) SVD whitening. We compress the SCDA\textsubscript{flip}\textsuperscript{+} feature to 256-d and 512-d, respectively, and report the compressed results in Table IV. Comparing the results shown in Table III and Table IV, the compressed methods can reduce the dimensionality without hurting the retrieval performance. SVD (which does not remove the mean vector) has slightly higher rates than PCA (which removes the mean vector). The “512-d SVD+whitening” feature can achieve better retrieval performance: 2%-~4% higher than the original SCDA\textsubscript{flip}\textsuperscript{+} feature on CUB200-2011 and Oxford Flowers, and significantly 7%-~13% on Aircrafts and Cars. Moreover, “512-d SVD+whitening” with less dimensions generally achieves better performance than other compressed SCDA features. Therefore, we take it as our optimal choice for FGIR. In the following, we present some retrieval examples based on “512-d SVD+whitening”.

In Fig. 7, we show two successful retrieval results and two failure cases for each fine-grained dataset, respectively. As shown in the successful cases, our method can work well when the same kind of birds, animals, flowers, aircrafts or cars appear in different kinds of background. In addition, for these failure cases, there exist only tiny differences between the query image and the returned ones, which can not be accurately detected in this pure unsupervised setting. We can also find some interesting observations, e.g., the last failure case of the flowers and pets. For the flowers, there are two correct predictions in the top-5 returned images. Even though the flowers in the correct predictions have different colors with the query, our method can still find them. For the pets’ failure cases, the dogs in the returned images have the same pose as the query image.

C. Quality and Insight of the SCDA Feature

In this section, we discuss the quality of the proposed SCDA feature. After SVD and whitening, the former distributed dimensions of SCDA have more discriminative ability, i.e., directly correspond to semantic visual properties that are useful for retrieval. We use five datasets (CUB200-2011, Stanford Dogs, Oxford Flowers, Aircrafts and Cars) as examples to illustrate the quality. We first select one dimension of “512-d SVD+whitening”, and then sort the value of that dimension in the descending order. Then, we visualize images in the same order, which is shown in Fig. 8.

Images of each column have some similar “attributes”, e.g., living in water and opening wings for birds; brown and white heads and similar looking faces for dogs; similar shaped inflorescence and petals with tiny spots for flowers; similar poses and propellers for aircrafts; similar point of views and motorcycle types for cars. Obviously, the SCDA feature has the ability to describe the main objects’ attributes (even subtle attributes). Thus, it can produce human-understandable interpretation manuals for fine-grained images, which might explain its success in fine-grained image retrieval. In addition, because the values of the compressed SCDA features might be positive, negative and zero, it is meaningful to sort these values either in the descending order (shown in Fig. 8), or ascending order. The images returned in ascending also exhibit some similar visual attributes.

D. General Image Retrieval Results

For further investigation of the effectiveness of the proposed SCDA method, we compare it with three state-of-the-art general image retrieval approaches (SPoC [23], CroW [25] and R-MAC [27]) on the INRIA Holiday [17] and Oxford Building 5K [12] datasets. Following the protocol in [22], [23], for the Holiday dataset, we manually fix images in the wrong orientation by rotating them by ±90 degrees, and report the mean average precision (mAP) over 500 and 55 queries for Holiday and Oxford Building 5K, respectively.

In the experiments of these two general image retrieval datasets, we use the SCDA and SCDA\textsubscript{flip} features, and compress them by SVD whitening. As shown by the results presented in Table V, the compressed SCDA\textsubscript{flip} (512-d) achieves the highest mAP among the proposed ones. In addition, comparing with state-of-the-arts, the compressed SCDA\textsubscript{flip} (512-d) is significantly better than SPoC [23], CroW [25], and comparable with the R-MAC approach [27]. Therefore, the proposed SCDA not only significantly outperforms the general image retrieval state-of-the-art approaches for fine-grained image retrieval, but can also obtain comparable results for general-purpose image retrieval tasks.

E. Fine-Grained Classification Results

In the end, we compare with several state-of-the-art fine-grained classification algorithms to validate the effectiveness of SCDA from the classification perspective.

In the classification experiments, we adopt two strategies to fine-tune the VGG-16 model with only the image-level labels. One strategy is directly fine-tuning the pre-trained model of the original VGG-16 architecture by adding the horizontal flips of the original images as data augmentation. After obtaining the fine-tuned model, we extract the SCDA\textsubscript{flip} as the whole image representations and feed them into a linear SVM [44] to train a classifier.

The other strategy is to build an end-to-end SCDA architecture. Before each epoch, the masks $\tilde{M}$ of pool5 and relu5.3 are extracted first. Then, we implement the selection process as an element-wise product operation between the convolutional activation tensor $T$ and the mask matrix $\tilde{M}$. Therefore, the descriptors located in the object region will remain, while the other descriptors will become zero vectors. In the forward pass of the end-to-end SCDA, we select the descriptors of
pool_3 and relu_5_2 as aforementioned, and then, both max- and average-pool (followed by $\ell_2$ normalization) the selected descriptors into the corresponding SCDA feature. After that, the SCDA features of pool_3 and relu_5_2 are concatenated, which is the so-called “SCDA^++”, as the final representation of the end-to-end SCDA model. Then, a classification (fc-softmax) layer is added for end-to-end training. Because the partial derivative of the mask is zero, it will not affect the backward processing of the end-to-end SCDA. After each epoch, the masks will be updated based on the learned SCDA model in the last epoch. When end-to-end SCDA converges, the SCDA^++ is also extracted.

For both strategies, the coefficient $\alpha$ of SCDA^++ is set to 1 to let the classifier to learn and then select important dimensions automatically. The classification accuracy comparison is listed in Table VI.

For the first fine-tuning strategy, the classification accuracy of our method (“SCDA (e.f.)”) is comparable or even better than the algorithms trained with strong supervised annotations, e.g., [4], [5]. For these algorithms using only image-level labels, our classification accuracy is comparable with the algorithms using similar fine-tuning strategies ([6], [7], [9]), but still does not perform as well as those using more powerful deep architectures and more complicated data augmentations [8], [31]. For the second fine-tuning strategy, even though “SCDA (end-to-end)” obtains a slightly lower classification accuracy than “SCDA (e.f.)”, the end-to-end SCDA model contains the least number of parameters (i.e., only 15.53M), which attributes to no fully connected layers in the architecture.

Thus, our method has less dimensions and is simple to implement, which makes SCDA more scalable for large-scale datasets without strong annotations and is easier to generalize. In addition, the CroW [25] paper presented the classification accuracy on CUB200-2011 without any fine-tuning (56.5% by VGG-16). We also experiment on the 512-d SCDA feature (only contains the max-pooling part this time for fair comparison) without any fine-tuning. The classification accuracy on that dataset is 73.7%, which outperforms their performance by a large margin.

### F. Additional Experiments on Completely Disjoint Classes

For further investigating the generalization ability of the proposed SCDA method, we additionally conduct experiments on a recently released fine-grained dataset for biodiversity analysis, i.e., the Moth dataset [45]. This dataset includes 2120 moth images of 675 highly similar classes, which are
completely disjoint with the images of ImageNet. In Table VII, we present the retrieval results of SCDA and other baseline methods. Because there are several classes in Moth have less than five images per class, we only report the top-1 mAP results. Consistent with the observations in Sec. IV-B, SCDA still outperforms other baseline methods, which proves the proposed method could generalize well.

G. Computational Time Comparisons

In this section, we compare the inference speeds of our SCDA with other methods. Because the methods listed in Table VIII can handle arbitrary image resolutions, different fine-grained data sets have different speeds. Specifically, much larger image will cause the GPUs out of memory. Thus, according to the original image scaling, we resize the images until \( \min(\text{im_height}, \text{im_width}) = 700 \) pixels, when \( \min(\text{im_height}, \text{im_width}) > 700 \). As the speeds reported in Table VIII, it is understandable that the speed of SCDA is lower than that of pool5. In general, SCDA has the comparable computational speeds with CroW, and is significantly faster than R-MAC. But, its speed is slightly lower (about 1 frame/sec) than SPoC. In practice, if your retrieval tasks prefer high accuracy, SCDA \( \text{flip}^+ \) is recommended. While, if you prefer efficiency, SCDA is scalable enough for handling large scale fine-grained datasets. Meanwhile, SCDA will bring good retrieval accuracy (cf. Table III).

H. Summary of Experimental Results

In the following, we summarize several empirical observations of the proposed selective convolutional descriptor aggregation method for FGIR.

- Simple aggregation methods such as max- and average-pooling achieved better retrieval performance than high-dimensional encoding approaches. The proposed SCDA representation concatenated both the max- and average-pooled features, which achieved the best retrieval performance as reported in Table II and Table III.
- Convolutional descriptors performed better than the representations of the fully connected layer for FGIR. In Table III, the representations of “pool5”, “selectFV”
and “selectVLAD” are all based on the convolutional descriptors. No matter what kind of aggregation methods they used, their top-\(k\) retrieval results are (significantly) better than the fully connected features.

- Selecting descriptors is beneficial to both fine-grained image retrieval and general-purposed image retrieval. As the results reported in Table III and Table V, the proposed SCDA method achieved the best results for

### Table VI

Comparison of classification accuracy on six fine-grained datasets. The “SCDA (f.t.)” denotes the SCDA features are extracted from the directly fine-tuned VGG-16 model. The “SCDA (end-to-end)” represents the SCDA features are from the fine-tuned end-to-end SCDA model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train phase</th>
<th>Test phase</th>
<th>Model</th>
<th>(\gamma) Para.</th>
<th>Dim.</th>
<th>Birds</th>
<th>Dogs</th>
<th>Flowers</th>
<th>Pets</th>
<th>Aircrafts</th>
<th>Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCDA (f.t.)</td>
<td></td>
<td></td>
<td>VGG-16×1</td>
<td>135.07M</td>
<td>4,096</td>
<td>80.5</td>
<td>92.1</td>
<td>91.0</td>
<td>79.5</td>
<td>85.9</td>
<td></td>
</tr>
<tr>
<td>SCDA (end-to-end)</td>
<td></td>
<td></td>
<td>VGG-16 (w/o FCs)×1</td>
<td>15.53M</td>
<td>4,096</td>
<td>80.1</td>
<td>77.4</td>
<td>90.3</td>
<td>78.6</td>
<td>85.1</td>
<td></td>
</tr>
</tbody>
</table>

Table VII

Comparison of retrieval performance on the Moth dataset [45]. The best result is marked in bold. Note that, because there are several fine-grained categories of Moth containing less than five images for each category, we here only report the Top-1 MAP results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimension</th>
<th>Top-1 mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>fc5</td>
<td>4,096</td>
<td>42.52</td>
</tr>
<tr>
<td>pool5</td>
<td>1,024</td>
<td>42.67</td>
</tr>
<tr>
<td>selectFV</td>
<td>2,048</td>
<td>40.33</td>
</tr>
<tr>
<td>selectNLAD</td>
<td>1,024</td>
<td>42.41</td>
</tr>
<tr>
<td>SPoC (with cen.)</td>
<td>256</td>
<td>42.96</td>
</tr>
<tr>
<td>CroW</td>
<td>256</td>
<td>50.78</td>
</tr>
<tr>
<td>R-MAC</td>
<td>512</td>
<td>45.38</td>
</tr>
<tr>
<td>SCDA</td>
<td>1,024</td>
<td>47.48</td>
</tr>
<tr>
<td>SCDA +</td>
<td>2,048</td>
<td>49.78</td>
</tr>
<tr>
<td>SCDA_flip +</td>
<td>4,096</td>
<td>50.52</td>
</tr>
<tr>
<td>SCDA_flip + (SVD whitening)</td>
<td>256</td>
<td>54.96</td>
</tr>
<tr>
<td>SCDA_flip + (SVD whitening)</td>
<td>512</td>
<td>57.19</td>
</tr>
</tbody>
</table>

Table VIII

Comparisons of inference speeds (frames/sec) on six fine-grained image datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Birds</th>
<th>Dogs</th>
<th>Flowers</th>
<th>Pets</th>
<th>Aircrafts</th>
<th>Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>pool5</td>
<td>9.54</td>
<td>9.01</td>
<td>6.15</td>
<td>10.3</td>
<td>2.92</td>
<td>5.81</td>
</tr>
<tr>
<td>SPoC (w/o cen.)</td>
<td>8.70</td>
<td>8.77</td>
<td>5.92</td>
<td>10.1</td>
<td>2.76</td>
<td>5.46</td>
</tr>
<tr>
<td>SPoC (with cen.)</td>
<td>8.40</td>
<td>8.62</td>
<td>5.78</td>
<td>10.1</td>
<td>2.72</td>
<td>5.49</td>
</tr>
<tr>
<td>CroW</td>
<td>7.81</td>
<td>7.04</td>
<td>5.26</td>
<td>7.75</td>
<td>2.60</td>
<td>4.72</td>
</tr>
<tr>
<td>R-MAC</td>
<td>4.22</td>
<td>4.52</td>
<td>3.00</td>
<td>5.05</td>
<td>1.93</td>
<td>3.62</td>
</tr>
<tr>
<td>SCDA</td>
<td>9.09</td>
<td>7.81</td>
<td>4.85</td>
<td>9.61</td>
<td>2.05</td>
<td>4.16</td>
</tr>
<tr>
<td>SCDA +</td>
<td>7.46</td>
<td>6.66</td>
<td>3.34</td>
<td>7.14</td>
<td>1.11</td>
<td>2.35</td>
</tr>
<tr>
<td>SCDA_flip +</td>
<td>3.80</td>
<td>3.48</td>
<td>1.81</td>
<td>3.83</td>
<td>0.55</td>
<td>1.19</td>
</tr>
</tbody>
</table>

FGIR, meanwhile was comparable with general image retrieval state-of-the-art approaches.

- The SVD whitening compression method can not only reduce the dimensions of the SCDA feature, but also improve the retrieval performance, even by a large margin (cf. the results of Aircrafts and Cars in Table IV). Moreover, the compressed SCDA feature had the ability to describe the main objects’ subtle attributes, which is shown in Fig. 8.

V. Conclusions

In this paper, we proposed to solely use a CNN model pre-trained on non-fine-grained tasks to tackle the novel
and difficult fine-grained image retrieval task. We proposed the Selective Convolutional Descriptor Aggregation (SCDA) method, which is unsupervised and does not require additional learning. SCDA first localized the main object in fine-grained image unsupervised with high accuracy. The selected (localized) deep descriptors were then aggregated using the best practices we found to produce a short feature vector for a fine-grained image. Experimental results showed that, for fine-grained image retrieval, SCDA outperformed all the baseline methods including general image retrieval state-of-the-arts. Moreover, these features of SCDA exhibited well-defined semantic visual attributes, which may explain why it has high retrieval accuracy for fine-grained images. Meanwhile, SCDA had the comparable retrieval performance on standard general image retrieval datasets. The satisfactory results of both fine-grained and general-purpose image retrieval datasets validated the benefits of selecting convolutional descriptors.

In the future, we consider including the selected deep descriptors’ weights to find object parts. Another interesting direction is to explore the possibility of pre-trained models for more complicated vision tasks such as unsupervised object segmentation. Indeed, enabling models trained for one task to be reusable for another different task, particularly without additional training, is an important step toward the development of learnware [46].

REFERENCES


Jian-Hao Luo received his BS degree in the College of Computer Science and Technology from Jilin University, China, in 2015. He is currently working toward the PhD degree in the Department of Computer Science and Technology, Nanjing University, China. His research interests are computer vision and machine learning.

Zhi-Hua Zhou (S’00-M’01-SM’06-F’13) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as an Assistant Professor in 2001, and is currently Chair Professor and Standing Deputy Director of the National Key Laboratory for Novel Software Technology, Nanjing. He is also the Founding Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning, data mining and pattern recognition. He has authored the book “Ensemble Methods: Foundations and Algorithms”, and published more than 100 papers in top-tier international journals or conference proceedings. He has received various awards/honors including the National Natural Science Award of China, the IEEE CIS Outstanding Early Career Award, the Microsoft Professorship Award, the Fok Ying Tung Young Professorship Award, and twelve international journals/conferences papers or competitions awards. He also holds 15 patents. He is an Executive Editor-in-Chief of the Frontiers of Computer Science, Associate Editor-in-Chief of the Science China: Information Sciences, Associate Editor of the ACM Transactions on Intelligent Systems and Technology, IEEE Transactions on Neural Networks and Learning Systems, etc. He served as Associate Editor-in-Chief for Chinese Science Bulletin (2008-2014), Associate Editor for IEEE Transactions on Knowledge and Data Engineering (2008-2012) and Knowledge and Information Systems (2003-2008). He founded AcmCL (Asian Conference on Machine Learning), served as Advisory Committee member for IJCAI, Steering Committee member for PAKDD and PIII, and Chair of various conferences such as General co-chair of ACML 2012, PCM 2013, PAKDD 2014, Program co-chair of SDM 2013, ICDM 2015, IJCAI 2015 Machine Learning Track and AAAI 2019, Workshop co-chair of KDD 2012, ICDM 2014, Tutorial co-chair of CIKM 2014, KDD 2015, and Area chair of ICML, NIPS, etc. He is the Chair of the IEEE CIS Data Mining and Big Data Analytics Technical Committee, the Chair of the CAAI (Chinese Association of Artificial Intelligence) Machine Learning Society, the Chair of the CCF (China Computer Federation) Artificial Intelligence & Pattern Recognition Society, and the Chair of the IEEE Computer Society Nanjing Chapter. He is a Fellow of the ACM, AAAS, AAAI, IEEE, IAPR, IET/IEEE and CCF.

Xiu-Shen Wei received the BS degree in Computer Science and Technology in 2012. He is currently a PhD candidate in the Department of Computer Science and Technology at Nanjing University, China. He achieved the first place in the Apparent Personality Analysis competition (in association with ECCV 2016) and the first runner-up in the Cultural Event Recognition competition (in association with ICCV 2015) as the team director. His research interests are computer vision and machine learning.

Jianxin Wu (M’09) received his BS and MS degrees in computer science from Nanjing University, and his PhD degree in computer science from the Georgia Institute of Technology. He is currently a professor in the Department of Computer Science and Technology at Nanjing University, China, and is associated with the National Key Laboratory for Novel Software Technology, China. He was an assistant professor in the Nanyang Technological University, Singapore, and has served as an area chair for CVPR 2017, ICCV 2015, senior PC member for AAAI 2017, AAAI 2016 and an associate editor for Pattern Recognition Journal. His research interests are computer vision and machine learning. He is a member of the IEEE.