Multi-Grained Selection and Fusion for Fine-Grained Image Representation

Jianrong Jiang, Hongxing Wang*
Key Laboratory of Dependable Service Computing in Cyber Physical Society (Chongqing University)
Ministry of Education, China
School of Big Data & Software Engineering
Chongqing University, Chongqing, China
Email: {jrjiang, ihxwang}@cqu.edu.cn

Abstract—How to learn a good fine-grained image representation is a key problem for fine-grained tasks. Most previous supervised methods suffer from insufficient training data, which require laborious annotations of fine-grained objects. In this paper, we propose an annotation-free method for fine-grained image representation, dubbed Multi-Grained Selection and Fusion (MGSF). The proposed MGSF extracts two types of visual features, i.e., fine-grained discriminative features that highlight informative convolutional parts by spatial selection and channel selection, and coarse-grained scene-level features that provide context information for fine-grained objects. Extensive experiments in fine-grained image retrieval demonstrate the superiority of our proposed representation compared with the state-of-the-art approaches on several fine-grained datasets.

Index Terms—fine-grained image retrieval, feature fusion, feature selection, channel selection

I. Introduction

Convolutional Neural Networks (CNNs) have been shown to be exceptionally effective at visual representation in computer vision, e.g., image recognition [1], [2], object detection [3], and image retrieval [4]. However, learning a good CNN model usually requires labeling a large number of training data, which is extremely difficult for fine-grained tasks [5]–[7]. To avoid fine-grained annotations, methods with weak-supervision [8]–[11] or even no supervision [12]–[14] are gaining increased attention.

Owing to the availability of pre-trained CNN models [1], [2], [15]–[17] on the large-scale ImageNet dataset [18], we can easily transfer these models to other image data for visual feature extraction [19]. However, it still challenges those directly extracted CNN features to represent the large intra-class variance and small inter-class variance in fine-grained images. To further capture fine-grained information, existing efforts focus on selecting salient areas from convolutional feature maps [9], [11], [12], [20]. In such cases, the selected areas are kept the same in all convolutional channels, where the differences between channels are not considered. Moreover, the fine-grained area selection ignores coarse-grained scene context for fine-grained objects.

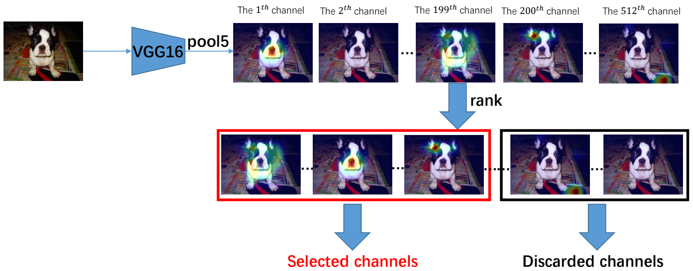


Fig. 1. An illustration of channel selection. In the last pooling layer of VGG16 [2], we select convolutional channels with more activated areas.

To address the above issues, we propose an unsupervised Multi-Grained Selection and Fusion (MGSF) method for fine-grained image feature extraction by aid of a pre-trained CNN model. The proposed MGSF consists of a fine-grained module and a coarse-grained module. Specifically, the fine-grained module aims at selecting discriminative parts for fine-grained objects from convolutional feature maps. Different from previous works that perform the selection in spatial units of feature maps, we also introduce channel selection beyond that. When exploiting spatial information to discard the backgrounds and localize object parts for fine-grained images, we observe that frequent patterns of channels in feature maps provide key clues for the presence of objects. As shown in Fig. 1, not all channels are equally activated. Most activated locations are with semantic information of objects. Therefore, we sort the channels according to the sizes of their activated areas, and select those channels with more activations. Besides spatial and channel selection, we also add a coarse-grained module for extracting coarse-grained scene-level features without feature selection. The complete coarse-grained features and selective fine-grained features complement each other, being integrated into a multi-grained representation to better represent fine-grained images.

The main contributions of this paper are summarized as follows:

- We propose a fine-grained feature extraction method

*Corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

via highlighting discriminative convolutional channels, which selects the semantic channels and discards the noisy ones.

- We propose to integrate coarse-grained scene-level features into our extracted fine-grained features for a better image representation.
- We evaluate our proposed fine-grained image representation on six popular benchmarks for fine-grained image retrieval, which achieves significant improvement over existing state-of-the-art methods.

II. Related Work

A. Fine-Grained Representation for Image Classification

Due to the large intra-class variance and small inter-class variance, fine-grained image classification is more challenging in feature representation than general image classification. Therefore, strong supervisions like bounding boxes of objects or part locations, are used in many works [6], [21]. For example, Zhang et al. addressed fine-grained image classification by extracting interested part features through detector selection for classifier training [6]. Wei et al. utilized Fully Convolutional Networks [22] to locate the informative objects and extract deep descriptors [21].

As supervised information, object bounding boxes or object parts are laborious and expensive to be obtained. To alleviate this problem, He et al. selected the semantic parts by spatial constraints of object and part proposals [9]. Similarly, Qi et al. exploited spatial distances between object parts in a weakly supervised way [11]. Yang et al. proposed a self-supervision method to localize semantic parts and designed a novel loss function to enable localizing the most informative regions with image-level labels [23]. Zheng et al. learned subtle features from part proposals and distilled fine-grained features into scene-level image features by a teacher-student network [10]. To enable part localization and feature learning mutually reinforcing each other for a discriminative feature representation, Lin et al. presented bilinear CNN representation by outer product of two CNN features [8]. Despite the above successes, few works study fine-grained image representation for image classification in an unsupervised way.

B. Fine-Grained Representation for Image Retrieval

In addition to classification, fine-grained representation is also of vital importance to fine-grained image retrieval (FGIR). Xu et al. proposed a multi-view cross-modal matching algorithm based on view selection for fine-grained sketch-based image retrieval [24]. Similarly, Song et al. introduced a spatial-semantic attention method [20]. Wei et al. localized the main object by selecting areas with higher activation values in feature maps for fine-grained descriptions [12] which significantly improved the retrieval performance unsupervisedly. Recently, image labels have been leveraged in FGIR. For example, Zheng et al. designed a novel loss function and proposed a weakly supervised object localization and representation [25]. In general, most of previous works only consider spatial relation of convolutional parts, which ignored the channel information [12], [20]. In this paper, we focus on both spatial selection and channel selection for fine-grained image retrieval without any annotations.

C. Transfer Learning for Image Representation

Transfer learning and domain adaptive methods in computer vision have a long history [26]. Deep networks have shown impressive transferable ability by reusing pre-trained models. The common strategy is fine-tuning [27]–[30]. For example, Girshick et al. proposed a transfer learning strategy that is to replace the input layer of the network and continue training with task-related data [29]. Xiao et al. applied three types of attention to train domain-specific deep nets [30]. Furthermore, Ganin et al. proposed a new domain adaptive representation learning method by Domain-Adversarial Neural Network for extracting domain-invariant features [27]. Ge et al. selected some samples similar to the target domain from the source training set for joint fine-tuning [28]. Our proposed method, in contrast, has no fine-tuning, but adopts an unsupervised manner to manipulate extracted features by a pre-trained deep model. Specially, we study feature selection and multi-grained fusion to produce a better visual representation for FGIR.

III. Methodology

In order to learn good image representations for FGIR, we extract and fuse two-grained visual features for each image. Fig. 2 shows two corresponding feature extraction modules, where the fine-grained module (green flow) learns discriminative fine-grained features and the coarse-grained module (yellow flow) captures the global context information of fine-grained objects.

A. Notations

The following notations are used in the rest of this paper. The term "feature maps" ($\boldsymbol{F}$) denotes the results by convolution; the term "activation maps" ($\boldsymbol{A}$) denotes the feature maps of one channel; the term "$pool_5$" denotes the feature maps from operating max-pooling before fully connected layer.

Given an input image, the feature maps through $pool_5$ can be represented as a 3D matrix $\boldsymbol{F} \in \mathbb{R}^{H \times W \times C}$. $\boldsymbol{F}$ can be regarded as having $H \times W$ spatial units. Every spatial unit contains one C-dimensional descriptor. Similarly, $\boldsymbol{F}$ can be regarded as consisting of C activation maps $\boldsymbol{A} \in \mathbb{R}^{H \times W}$.

B. Discriminative Fine-Grained Feature Selection

Rich details in images play an important role in fine-grained image retrieval. We utilize VGG16 pre-trained on ImageNet [18] to extract convolutional feature maps $\boldsymbol{F}$ for each input image. Since parts in an image are in general spatially connected and activated on most channels, we perform spatial selection and channel selection to discard

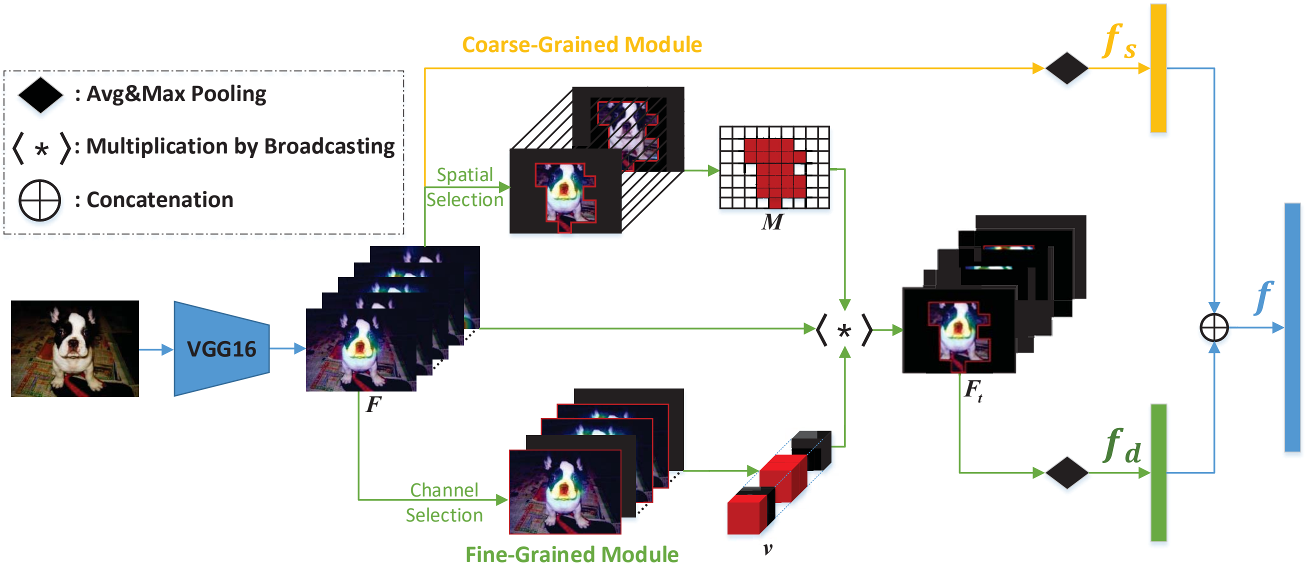


Fig. 2. Overview of our proposed MGSF. It consists of coarse-grained and fine-grained modules. We extract the scene-level feature $\boldsymbol{f_s}$ by Avg&Max Pooling of $pool_5$ feature maps $\boldsymbol{F}$ in coarse-grained module. For fine-grained module, the transformed fine-grained feature maps $\boldsymbol{F_t}$ can be obtained by channel selection and spatial selection, which has the characteristics of discarding the background of images and highlighting semantic information of channels. Avg&Max Pooling of $\boldsymbol{F_t}$ will lead to the discriminative fine-grained feature $\boldsymbol{f_d}$. The final image representation $\boldsymbol{f}$ is generated by weighted concatenation of $\boldsymbol{f_d}$ and $\boldsymbol{f_s}$. Note that $\boldsymbol{M}$ and $\boldsymbol{v}$ are binary masks obtained from spatial selection and channel selection, where the entry is 1 if colored red, 0 otherwise. Best viewed in color.

the background and highlight the informative channels on $\boldsymbol{F}$, so as to obtain more discriminative image representations.

Spatial Selection. To localize fine-grained parts, we follow [12] to conduct a spatial selection on $\boldsymbol{F}$. Specially, we opearte sum pooling in channel direction on $\boldsymbol{F}$, then obtain a 2D map $\boldsymbol{S} \in \mathbb{R}^{H \times W}$ by $\boldsymbol{S} = \sum_{i=1}^{C} \boldsymbol{A}^i$, where $\boldsymbol{A}^i$ is the i^{th} activation map in $\boldsymbol{F}$. As shown in [12], the areas with higher values in $\boldsymbol{S}$ are more likely to be parts of an object. We calculate the mean value T_s of $\boldsymbol{S}$ as the threshold to decide which descriptors are selected. If the entry of $\boldsymbol{S}$ in the position (h,w) is lower than T_s, the corresponding spatial unit of $\boldsymbol{F}$ is set to 0, 1 otherwise. As there may still remain small noisy parts, the maximum connected graph is subsequently adopted. Such a spatial selection is essentially equivalent to applying a binary mask $\boldsymbol{M} \in \mathbb{R}^{H \times W}$ to $\boldsymbol{F}$.

Channel Selection. As has been shown in Fig. 1, not all channels are equally important. There are some channels not activated at all and even some activated in noisy areas. For the channel activated in most of the spatial locations, there is more likely to contain semantic information of an object. If few positions are activated, this channel may only contains the background. For example in Fig. 1, in the 1^{th} activation map, the activated region highlights the nose of a French bulldog. In the 199^{th} and 200^{th} activation maps, the areas are activated on the head and ears. On the contrary, background is involved in the 512^{th} activation map and there is even no activated region in the 2^{th} activation map. Therefore, it is critical to select semantic channels.

Based on this observation that the more positions a channel activates, the more semantic information it has, we present a novel approach for channel selection. Concretely, we use $T_c = T_s/C$ to filter out the activated positions which results in a 3D matrix $\boldsymbol{B} \in \mathbb{R}^{H \times W \times C}$ in (1), where $A^i_{h,w}$ denotes the $(h,w)^{th}$ entry of the i^{th} activation map A^i, and $B^i_{h,w}$ denotes the $(h,w)^{th}$ entry of the i^{th} slice of $\boldsymbol{B}$.

$$B^i_{h,w} = \begin{cases} 1 & A^i_{h,w} \geq T_c, \\ 0 & otherwise. \end{cases} \tag{1}$$

We next sum the elements of $\boldsymbol{B}$ along the first and the second dimension, leading to a statistical vector $\boldsymbol{s} \in \mathbb{R}^C$. Then we sort the elements of $\boldsymbol{s}$ to obtain an ascending list $\boldsymbol{l}$. The N^{th} entry of $\boldsymbol{l}$ defines a threshold T to discard N channels with less activation areas:

$$T = \boldsymbol{l}(N). \tag{2}$$

If the j^{th} entry of $\boldsymbol{s}$ is lower than T, the corresponding channel is set to 0, 1 otherwise. Such a channel selection is essentially equivalent to applying a binary mask $\boldsymbol{v} \in \mathbb{R}^C$ to $\boldsymbol{F}$.

Transformed Fine-Grained Feature Maps. After spatial and channel selection on the original feature maps $\boldsymbol{F}$, we further integrate the selected results into the transformed fine-grained feature maps $\boldsymbol{F_t} \in \mathbb{R}^{H \times W \times C}$ which can be formulated as:

$$\boldsymbol{F_t} = \langle \langle \boldsymbol{F} * \boldsymbol{M} \rangle * \boldsymbol{v} \rangle \,, \tag{3}$$

where $\langle * \rangle$ indicates multiplication of two matrices by broadcasting. Therefore, $\boldsymbol{F_t}$ has the effect of both spatial selection and channel selection to highlight fine-grained parts. Average and max pooling of $\boldsymbol{F_t}$ will lead to two discriminative fine-grained feature vectors of 512-dimensionatliy, which are concatenated into $\boldsymbol{f_d} \in \mathbb{R}^{1024}$. For convenience, we define this operation from $\boldsymbol{F}$ to $\boldsymbol{f_d}$ as Avg&Max Pooling.

C. Fusing with Coarse-Grained Features

In addition to extracting the selective fine-grained feature vector $\boldsymbol{f_d}$ from the transformed fine-grained feature maps $\boldsymbol{F_t}$, we also extract the coarse-grained scene-level feature vector $\boldsymbol{f_s}$ from the original feature maps $\boldsymbol{F}$ by Avg&Max Pooling.

As $\boldsymbol{f_d}$ and $\boldsymbol{f_s}$ complement each other in two granularities, we normalize them using l_2 norm for a weighted concatenation, resulting in the multi-grained feature vector:

$$\boldsymbol{f} = \left[\frac{\boldsymbol{f_d}}{\left\| \boldsymbol{f_d} \right\|_2}, \alpha \times \frac{\boldsymbol{f_s}}{\left\| \boldsymbol{f_s} \right\|_2} \right], \tag{4}$$

where α is the weighting coefficient for $\boldsymbol{f_s}$. We empirically set α to 0.5 for all datasets except the Oxford Flowers for which we set α to 1.0.

IV. Experiments

A. Experimental Settings

The datasets for evaluation include:

- CUB200 [34] with 11,788 images in 200 classes.
- Standford Dogs [35] with 20,580 images in 120 classes.
- Oxford Flowers [36] with 8,189 images in 102 classes.
- Oxford Pets [37] with 7,349 images in 37 classes.
- FGVC-Aircrafts [38] with 10,000 images in 100 classes.
- Standford Cars [39] with 16,185 images in 196 classes.

We compare the following methods:

- pool$_5$+pooling: This baseline is directly from the *pool*$_5$ layer by Avg&Max Pooling, which generates a 1024-dimensional feature vector.
- selectFV: We use FV [40] to encode the *pool*$_5$ features after spatial selection, which generates a 2048-dimensional feature vector.
- selectVLAD: Similar to selectFV, but replacing FV [40] by VLAD [41], which generates a 1024-dimensional feature vector.
- SPoC [31]: This method uses sum-pooling on convolutional feature maps to generate a 256-dimensional feature vector.
- CroW [32]: This method exploits space and channel weights before sum-pooling based on SPoC, which generates a 256-dimensional feature vector.
- R-MAC [33]: This method encodes multiple regions on convolutional feature maps for a 512-dimensional image representation.
- SCDA [12]: This method performs spatial selection in *pool*$_5$ feature maps to generate a 1024-dimensional feature vector.
- SCDA_flip+ [12]: This method combines *pool*$_5$, *Relu*$_{5\_2}$ and image flipping based on SCDA to generate a 4096-dimensional feature vector.

B. Implementation Details

We extract feature maps $\boldsymbol{F}$ from the convolutional layers *pool*$_5$ of the publicly available pre-trained VGG16 model. Each image is fed into the model without cropping, before which each pixel value is subtracted by the mean of the image. Training and test sets are divided according to the default setting of each dataset. The training set serves as a database to be queried and the test set provides query images. The cosine similarity is used for nearest neighbor search. For evaluation, we use top-1 and top-5 mean average precision (mAP).

For selectFV and selectVLAD, the number of clusters in VLAD and the number of Gaussian components in FV are both set to 2 following [12]. It is worth noting that we implement the results of SCDA and SCDA_flip+ using the code provided by [12]. The results of pool$_5$+pooling, selectFV, selectVLAD, SPoC [31], CroW [32], and R-MAC [33] are from [12].

C. Comparisons with other Methods

The retrieval mAP results are reported in Tabel I. Compared with the proposed MGSF, the retrieval performance of pool$_5$+pooling is fairly low. The performance of selectFV and selectVLAD are also not satisfactory, and some lower than pool$_5$+pooling in CUB200 and Standford Dogs.

SPoC, CroW and R-MAC are originally used for general image retrieval which are not competitive. It is because that general image retrieval is completely different from FGIR and general deep learning image retrieval methods can not be directly applied to FGIR.

SCDA and SCDA_flip+ are existing state-of-the-art methods for FGIR. SCDA produces a more compact representation, but performs slightly worse than SCDA_flip+. SCDA_flip+ is the best amongst the previous methods on all datasets except Standford Dogs.

As shown in Table I, our proposed MGSF obtains remarkably best performance on all datasets except Standford Cars. For CUB200, Standford Dogs and Oxford Flowers, our method significantly outperforms SCDA_flip+ by 2.23% (62.34% vs. 60.11%) , 1.59% (75.82% vs. 74.23%) and 4.16% (81.38% vs. 77.22%) in top-1 mAP respectively, and by a margin of 1.50%, 1.35% and 3.24% in top-5 mAP respectively. These increases are quite significant using with only a half length of feature vector compared with SCDA_flip+. It shows the ability of our fine-grained and coarse-grained modules to learn powerful representation for fine-grained image retrieval. For Oxford Pets, although SCDA_flip+ results in quite high performance,

TABLE I
Comparison results of different methods in fine-grained image retrieval. The best results are in bold.

		CUB200		Standford Dogs		Oxford Flowers		Oxford Pets		FGVC-Aircrafts		Standford Cars	
		mAP (%)		mAP (%)		mAP (%)		mAP (%)		mAP (%)		mAP (%)	
Method	Dimension	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
pool_5+pooling	1,024	57.54	63.66	69.98	75.55	70.73	74.05	85.09	87.74	47.37	53.61	34.88	41.86
selectFV	2,048	52.04	59.19	68.37	73.74	70.47	73.60	85.04	87.09	48.69	54.68	35.32	41.60
selectVLAD	1,024	55.92	62.51	69.28	74.43	73.62	76.86	85.50	87.94	50.35	56.37	37.16	43.84
SPoC (w/o cen.) [31]	256	34.79	42.54	48.80	55.95	71.36	74.55	60.86	67.78	37.47	43.73	29.86	36.23
SPoC (with cen.) [31]	256	39.61	47.30	48.39	55.69	65.86	70.05	64.05	71.22	42.81	48.95	27.61	33.88
CroW [32]	256	53.45	59.69	62.18	68.33	73.67	76.16	76.34	80.10	53.17	58.62	44.92	51.18
R-MAC [33]	512	52.24	59.02	59.65	66.28	76.08	78.19	76.97	81.16	48.15	54.94	**46.54**	**52.98**
SCDA [12]	1,024	59.34	65.47	74.99	79.45	75.15	77.96	87.63	89.26	53.26	58.61	38.39	45.20
SCDA_flip+ [12]	4,096	60.11	66.29	74.25	78.83	77.22	79.65	87.90	89.84	53.98	59.63	39.90	46.52
$\boldsymbol{f_d}$	1,024	59.64	65.66	75.59	79.89	76.47	78.91	88.36	90.16	**54.01**	**59.87**	38.61	45.37
MGSF	2,048	**62.34**	**67.79**	**75.82**	**80.18**	**81.38**	**82.89**	**88.80**	**90.60**	52.69	58.31	39.14	46.16

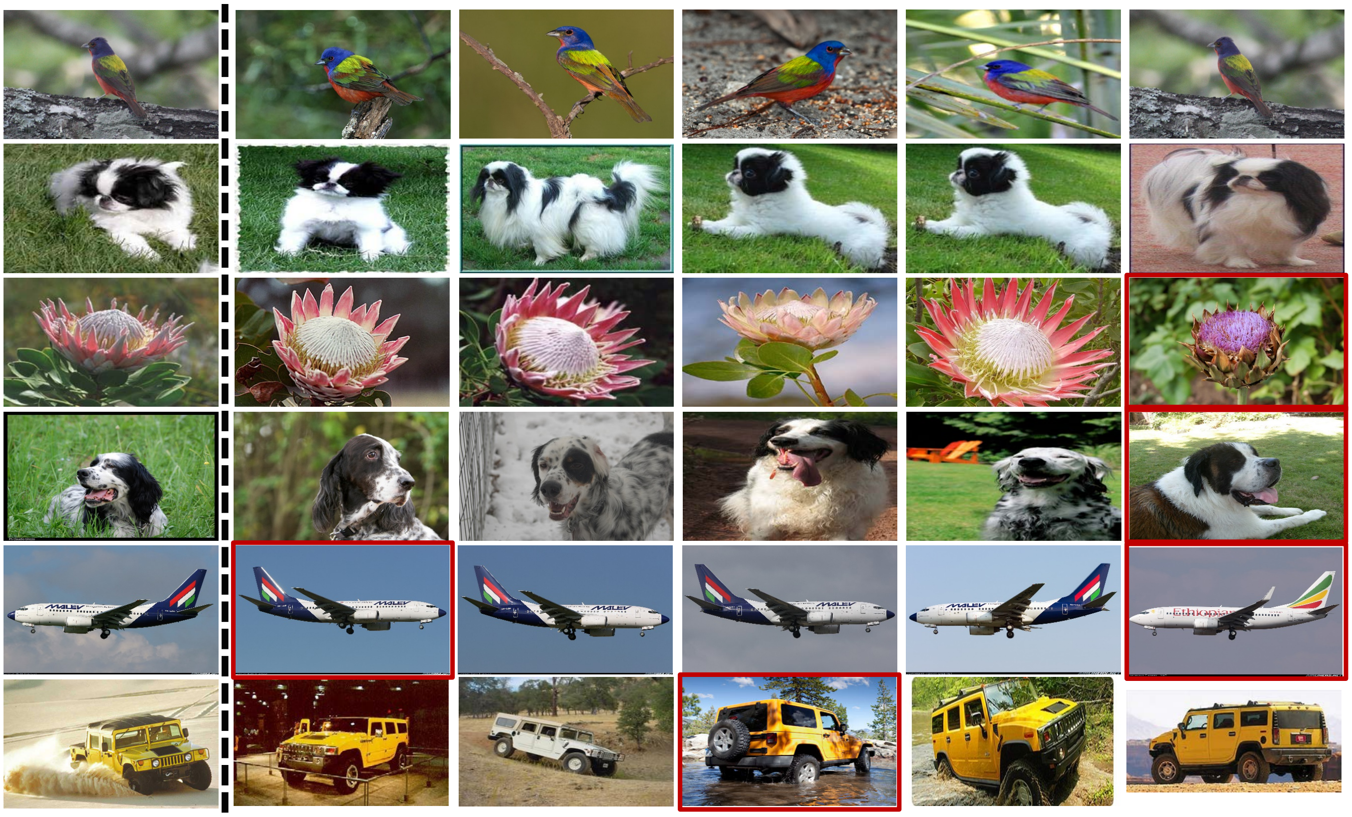
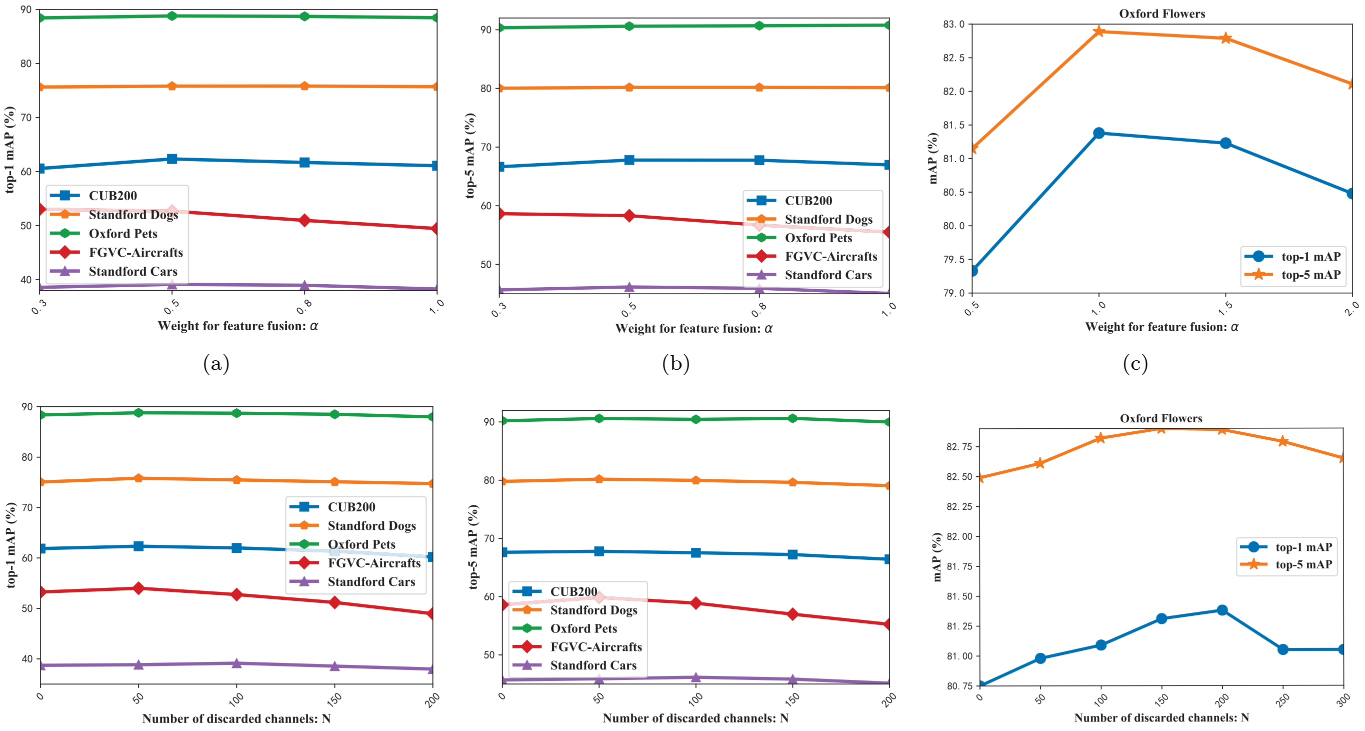


Fig. 3. Some retrieval results of six fine-grained datasets. Each row shows a retrieval example from a specific dataset. Queries are in the first column. Retrieval results are listed in descending order of similarities with the query. Images in red rectangle indicate incorrectly retrieved results. Best viewed in color.

our method is still 0.9% and 0.76% better in terms of top-1 and top-5 mAP. For FGVC-Aircrafts, the feature representation needs to be more discriminative compared to other datasets. This is also shown by the top ranked incorrect retrieval result in the penultimate row of Fig. 3. The differences between the incorrect image and the query are so slight that experts may also not be able to distinguish the differences between them. In Standford Cars, although not the best, $\boldsymbol{f_d}$ improves 0.22% and 0.17% compared to SCDA in top-1 and top-5 mAP respectively, and further boosts the performance to 39.14% and 46.16% after combining with $\boldsymbol{f_s}$, which still demonstrate the effectiveness of channel selection and feature fusion in our proposed MGSF. According to suggestions from [12], we can also take feature compression methods such as Singular Value Decomposition and Principal Component Analysis to further improve the performance.

By comparing SCDA and $\boldsymbol{f_d}$, we can find that our channel selection suppresses the noise of background and highlights the discriminative channels. MGSF shows better performance for fine-grained image retrieval in most datasets. As shown in Fig. 3, we show one example of the retrieval results using the proposed MGSF for each dataset. The results of the first two rows show that our



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4. Retrieval performance of our proposed method on six datasets with parameters α(a-c) and N(d-f).

method achieves competitive performance. The failure cases in the third and sixth rows show similar appearance to the queries. For the failure case of Oxford Pets, the dog in the returned image has the same pose as the query image. For FGVC-Aircrafts, the first image returned is almost identical to the query except that the left and right sides are interchanged. But their categories are Boeing 737-600 and Boeing 737-700 respectively. These errors are unavoidable, especially without any annotation information.

D. Parameter Setting

The main parameters include α, the weighting coefficient for $\boldsymbol{f_s}$ in multi-grained fusion, and N, the number of discarded channels.

Weight for feature fusion : α. The coefficient α is set to 0.5 for all datasets apart from Oxford Flowers. For the training and test splits in the Oxford Flowers dataset, the training set has 1020 images while the test set has 6149 images. Different from other datasets, the number of training sets almost one sixth of the test set. Fine-grained module learns discriminative fine-grained features, but $\boldsymbol{f_d}$ no longer has great influence in this case, especially in no supervision. Therefore, we doubled the coefficient α to 1.0 for better balancing $\boldsymbol{f_d}$ and $\boldsymbol{f_s}$. Fig. 4(a-c) shows the feasibility of our setting for α.

Number of Discarded Channels : N. We report retrieval performance under different N in Fig. 4(d-f). For Oxford Flowers, N is set to 200, while the best N is 50 for the other datasets. This also indicates that the Oxford Flowers dataset is different from the other datasets, and most of fine-grained features for flowers are encoded in a smaller number of channels. Although setting a larger N will miss some scene information, it can be compensated by multi-grained fusion with another scene-level feature $\boldsymbol{f_s}$.

TABLE II
Ablation study on CUB200.

		mAP (%)	
Method	Dimension	top-1	top-5
$\boldsymbol{f_s}$	1,024	57.54	63.66
Channel selection	1,024	59.03	65.31
SCDA [12]	1,024	59.34	65.47
$\boldsymbol{f_d}$(SCDA [12]+Channel selection)	1,024	59.64	65.66
Channel selection+$\boldsymbol{f_s}$	2,048	61.53	66.86
SCDA [12]+$\boldsymbol{f_s}$	2,048	61.87	67.62
MGSF	2,048	**62.34**	**67.79**

E. Ablation Study

To investigate the contribution of channel selection and multi-grained feature fusion in our proposed method, we report the top-1 and top-5 mAP for CUB200 in Table II. The results in the middle of the table show that the proposed channel selection improves fine-grained image retrieval performance of SCDA [12]. It can bring an improvement of 0.3% (59.64% vs. 59.34%) and 0.19%

(65.66% vs. 65.47%) in top-1 and top-5 mAP respectively. Comparing the results of the last two rows, it also demonstrates the effectiveness of channel selection, where our proposed MGSF with channel selection obtains a performance gain of 0.47% in top-1 mAP against SCDA+$\boldsymbol{f_s}$ without channel selection.

By comparing results of the last three rows and the first row, we can see that multi-grained fusion achieves great success. MGSF obtains the best results, 62.34% and 67.79% in top-1 and top-5 mAP respectively, which siginificantly outperforms $\boldsymbol{f_s}$. It further verifies FGIR will benefit greatly from fusing $\boldsymbol{f_s}$ and $\boldsymbol{f_d}$ in our MGSF.

V. Conclusion

In this paper, we propose a novel Multi-Grained Selection and Fusion (MGSF) method for fine-grained image representation in an unsupervised way. Based on a pre-trained CNN model, we higlight informative object parts as fine-grained discriminative features by spatial and channel selection, and combine them with coarse-grained scene-level features for obtaining a better image representation. Fine-grained image retrieval results show the advantageous of our MGSF in comparsion with previous state-of-the-art methods on six fine-grained image datasets.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants 61976029 and 61602069, the Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2016jcyjA0468, and the Fundamental Research Funds for the Central Universities under Grant 2018CDXYRJ0030.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[4] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in European Conference on Computer Vision, 2014.

[5] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," arXiv preprint arXiv:1406.2952, 2014.

[6] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in European Conference on Computer Vision, 2014.

[7] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[8] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in IEEE International Conference on Computer Vision, 2015.

[9] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in AAAI Conference on Artificial Intelligence, 2017.

[10] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[11] L. Qi, X. Lu, and X. Li, "Exploiting spatial relation for fine-grained image classification," Pattern Recognition, vol. 91, pp. 47–55, 2019.

[12] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," IEEE Transactions on Image Processing, vol. 26, no. 6, pp. 2868–2881, 2017.

[13] A. Jimenez, J. M. Alvarez, and X. Giro-i Nieto, "Class weighted convolutional features for visual instance search," in British Machine Vision Conference, 2017.

[14] J. Xu, C. Shi, C. Qi, C. Wang, and B. Xiao, "Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval," in AAAI Conference on Artificial Intelligence, 2018.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[19] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in International Conference on Artificial Intelligence and Statistics, 2010.

[20] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in IEEE International Conference on Computer Vision, 2017.

[21] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," Pattern Recognition, vol. 76, pp. 704–714, 2018.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[23] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in European Conference on Computer Vision, 2018.

[24] P. Xu, Q. Yin, Y. Qi, Y.-Z. Song, Z. Ma, L. Wang, and J. Guo, "Instance-level coupled subspace learning for fine-grained sketch-based image retrieval," in European Conference on Computer Vision, 2016.

[25] X. Zheng, R. Ji, X. Sun, Y. Wu, F. Huang, and Y. Yang, "Centralized ranking loss with weakly supervised localization for fine-grained object retrieval." in International Joint Conference on Artificial Intelligence, 2018.

[26] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[27] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The Journal of Machine Learning Research, 2016.

[28] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[30] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[31] A. Babenko and V. Lempitsky, “Aggregating deep convolutional features for image retrieval,” arXiv preprint arXiv:1510.07493, 2015.

[32] Y. Kalantidis, C. Mellina, and S. Osindero, “Cross-dimensional weighting for aggregated deep convolutional features,” in European Conference on Computer Vision, 2016.

[33] G. Tolias, R. Sicre, and H. Jégou, “Particular object retrieval with integral max-pooling of CNN activations,” arXiv preprint arXiv:1511.05879, 2015.

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.

[35] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel dataset for fine-grained image categorization: Stanford dogs,” in CVPR Workshop on Fine-Grained Visual Categorization, 2011.

[36] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in Indian Conference on Computer Vision, Graphics & Image Processing, 2008.

[37] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, “Cats and dogs,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[38] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, “Fine-grained visual classification of aircraft,” arXiv preprint arXiv:1306.5151, 2013.

[39] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in IEEE International Conference on Computer Vision Workshops, 2013.

[40] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image classification with the fisher vector: Theory and practice,” International Journal of Computer Vision, vol. 105, no. 3, pp. 222–245, 2013.

[41] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1704–1716, 2011.