Weakly Supervised Fine-Grained Categorization with Part-Based Image Representation

Yu Zhang, Xiu-Shen Wei, Jianxin Wu, *Member, IEEE*, Jianfei Cai, *Senior Member, IEEE*, Jiangbo Lu, *Senior Member, IEEE*, Viet-Anh Nguyen, and Minh N. Do, *Fellow, IEEE*

***Abstract*—In this paper, we propose a fine-grained image categorization system with easy deployment. We do not use any object/part annotation (weakly-supervised) in the training or in the testing stage, but only class labels for training images. Fine-grained image categorization aims to classify objects with only subtle distinctions (e.g., two breeds of dogs that look alike). Most existing works heavily rely on object/part detectors to build the correspondence between object parts, which require accurate object or object part annotations at least for training images. The need for expensive object annotations prevents the wide usage of these methods. Instead, we propose to generate multi-scale part proposals from object proposals, select useful part proposals, and use them to compute a global image representation for categorization. This is specially designed for the weakly-supervised fine-grained categorization task, because useful parts have been shown to play a critical role in existing annotation-dependent works but accurate part detectors are hard to acquire. With the proposed image representation, we can further detect and visualize the key (most discriminative) parts in objects of different classes. In the experiments, the proposed weakly-supervised method achieves comparable or better accuracy than state-of-the-art weakly-supervised methods and most existing annotation-dependent methods on three challenging datasets. Its success suggests that it is not always necessary to learn expensive object/part detectors in fine-grained image categorization.**

***Index Terms*—Fine-grained categorization, weakly-supervised, part selection.**

I. Introduction

Fine-grained image categorization has been popular for the past few years. Different from traditional general image recognition such as scene or object recognition, fine-grained categorization deals with images with subtle distinctions, which usually involves the classification of subclasses of objects belonging to the same class like birds [1], [2], [3], [4], dogs [5], planes [6], plants [7], [8], [9], etc. As shown in Fig. 1, fine-grained categorization needs to discriminate between objects that are visually similar to each other. In the red box of Fig. 1, Siberian Husky and Malamute are two different breeds of dogs that might be difficult to distinguish for humans that are not experts. However, general image categorization is comparatively easier, e.g., most people can easily recognize that the red box in Fig. 1 contains dogs while the blue box contains a kangaroo. Image representations that used to be useful for general image categorization may fail in fine-grained image categorization, especially when the objects are not well aligned, e.g., the two dogs are in different pose and the backgrounds are cluttered. Therefore, fine-grained categorization requires methods that are more discriminative than those for general image classification.

Fig. 1. Fine-grained categorization vs. general image categorization. Fine-grained categorization (red box) processes visually similar objects, e.g., to recognize Siberian Husky and Malamute. General image categorization usually distinguishes an object such as dogs (red box) from other objects that are visually very different (e.g., a kangaroo).

Fine-grained categorization has wide applications in both industry and research societies. Different datasets have been constructed in different domains, e.g., birds [1], butterflies [10], cars [11], etc. These datasets can have significant social impacts, e.g., butterflies [10] are used to evaluate the forest ecosystem and climate change.

One important common feature of many existing fine-grained methods is that they *explicitly use annotations of an object or even object parts* to depict the object as precisely as possible. Bounding boxes of objects and / or object parts are the most commonly used annotations. Most of them heavily rely on object / part detectors to find the part correspondence among objects.

For example, in [12], [13], the poselet [14] is used to detect object parts. Then, each object is represented with a bag of poselets, and suitable matches among poselets (parts) could be found between two objects. Instead of using poselets, [15] used the deformable part models (DPM) [16] for object part

This work was mainly done when Yu Zhang was working in ADSC and NTU. J. Wu is supported in part by the National Natural Science Foundation of China under Grant No. 61422203. J. Cai is supported in part by Singapore MoE AcRF Tier-1 Grant RG138/14. Y. Zhang, J. Lu, V.-A. Nguyen and M. N. Do are supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). M. N. Do is supported in part by the US National Science Foundation (NSF) grants CCF-1218682 and IIS 11-16012.

Y. Zhang is with the Bioinformatics Institute, A*SATR, Singapore. E-mail: zhangyu@bii.a-star.edu.sg.

X.-S. Wei and J. Wu are with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: weixs@lamda.nju.edu.cn, wujx2001@nju.edu.cn.

J. Cai is with the School of Computer Engineering, Nanyang Technological University, Singapore. E-mail: asjfcai@ntu.edu.sg.

J. Lu, and V.-A. Nguyen are with the Advanced Digital Sciences Center, Singapore. E-mail: {Jiangbo.Lu, vanguyen}@adsc.com.sg.

M. N. Do is with the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. E-mail: minhdo@illinois.edu.

detection. In [15] DPM is learned from the annotated object parts in training objects, which is then applied on testing objects to detect parts. Some works, like [17], [18], transfer the part annotations from objects in training images to those sharing similar shapes in testing images. Instead of seeking precise part localization, [17] proposed an unsupervised object alignment technique, which roughly aligns objects and divides them into corresponding parts along certain directions. It achieves better results than the label transfer method. Recently, [19] proposed to use object and part detectors with powerful CNN feature representations [20], which achieves state-of-the-art results on the Caltech-UCSD Birds (CUB) 200-2011 [1] dataset. The geometric relationship between an object and its parts are considered in [19]. [21] also shows that part-based models with CNN features are able to capture subtle distinctions among objects. [22] used object bounding boxes to cosegment objects and align the parts. Some other works, e.g., [23], [24], recognize fine-grained images with human in the loop.

In this paper, a part refers to a subregion in an object. For example, the parts in a bird include head, body, legs, etc. To achieve accurate part detection, most existing fine-grained works require annotated bounding boxes for objects, in both training and testing stages. As pointed out in [19], such a requirement is not so realistic for practical usage. Thus, a few works, such as [19], [20], have looked into a more realistic setup, i.e., only utilizing the bounding box in the training stage but not in the testing stage. However, even with such a setup, it is still hard for the wide deployment of these methods since accurate object annotations needed in the training stage are usually expensive to acquire, especially for large-scale image classification problems. It is an interesting research problem that frees us from the dependency on detailed manual annotations in fine-grained image categorization tasks. [25] has shown promising results without using the detailed manual annotations. They try to detect accurate objects and parts with complex deep learning models for fine-grained recognition.

In this paper, it is also our aim to *categorize fine-grained images with only category labels and without any bounding box annotation in both training and testing stages, while not degrading the categorization accuracy*. Our setup is the same as that of [25]. Notice that in the existing annotation-dependent works, representative parts like head and body in birds [19] have been shown to play the key role in capturing the subtle differences of fine-grained images. Different from general image recognition which usually uses a holistic image representation, we also try to make use of part information. However, unlike state-of-the-art fine-grained categorization methods, we do not try to find accurate part detections. Since the existing accurate part detectors (e.g., [19]) rely on the bounding box annotation while we consider a weakly-supervised setup in this research. Our key idea is to generate part proposals from object proposals, then select useful part proposals, and encode the selected part proposals into a global image representation for fine-grained categorization.

Fig. 2 gives a system overview, where there are three major steps: *part proposal generation*, *useful part selection*, and *multi-scale image representation*.

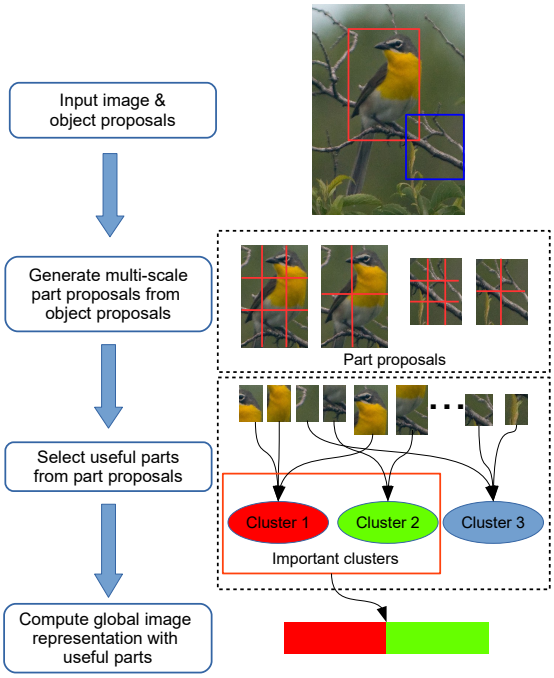


Fig. 2. System overview. This figure is best viewed in color. Note that we do not use any bounding box or part annotation.

- In the first step, we extract object proposals which are image patches that may contain an object. Part proposals are the sub-regions of the object proposals in each image, as illustrated in Fig. 2. We propose an efficient multi-max pooling (MMP) strategy to generate features for multi-scale part proposals by leveraging the internal structure of CNN.
- Considering the fact that most part proposals generated in the first step are from background clutters (which are harmful to categorization), in the second step, we propose to select useful part proposals from each image by exploring useful information in part clusters (all part proposals are clustered). For each part cluster, we compute an importance score, indicating how important the cluster is for the fine-grained categorization task. Then, those part proposals assigned to the useful clusters (i.e., those with the largest importance scores) are selected as useful parts.
- Finally, the selected part proposals in each image are encoded into a global image representation. To highlight the subtle distinction among fine-grained objects, we encode the selected parts at different scales separately, which we name as SCale Pyramid Matching (ScPM). ScPM provides a better discrimination than encoding all parts in one image altogether, i.e., without using the proposed scale pyramid matching.

Note that we propose to select *many* useful parts from multi-scale part proposals of objects in each image and compute a global image representation for it, which is then used to learn a linear classifier for image categorization. We believe that selecting many useful part proposals is better than selecting only the best part proposal in the final global

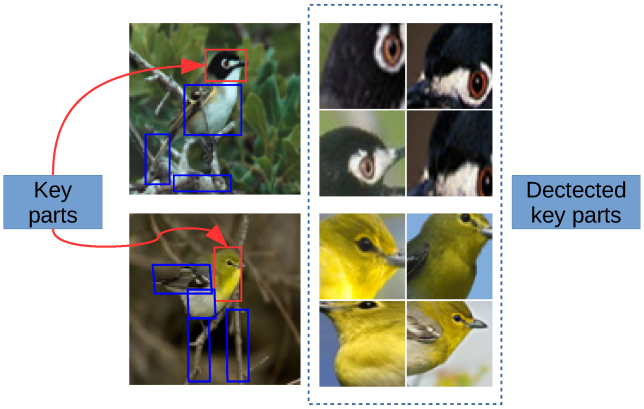


Fig. 3. Black-capped Vireo and Yellow-throated Vireo. They have the most distinctive parts in multiple part proposals: black cap and yellow throat, respectively, which are specified in red boxes. On the right, we show the key parts detected using the proposed representation from the two species. More examples of detected discriminative parts can be found in Fig. 8. This figure is best viewed in color.

representation. This is because it is very difficult to determine the exact location of an object/part in the image in our weakly-supervised scenario. Multiple useful part proposals can compensate each other to provide more useful information in characterizing the object. Experimental results show that the proposed method achieves comparable or better accuracy than state-of-the-art weakly-supervised work [25] and even most of the existing annotation-dependent methods on three challenging benchmark datasets. Its success suggests that it is not always necessary to learn expensive object / part detectors in fine-grained image categorization.

In addition, utilizing the proposed weakly-supervised fine-grained image representation, we can detect the key (most discriminative) object parts for different classes, which coincide well with the rules used by human experts (e.g., the yellow-throated vireo and the black-capped vireo differ because the yellow-throated vireo has a yellow throat while the black-capped vireo has a black head, cf. Fig. 3).

Overall, our main contribution lies in the explicit part proposal generation and selection, which, to the best of our knowledge, is for the first time proposed for fine-grained image categorization in a weakly-supervised setup. Another major contribution is the proposed framework which coherently integrates the three modules, part proposal generation, useful part selection and multi-scale image representation, and achieves state-of-the-art results.

II. Related Works

In this section, we review several works from two aspects of fine-grained categorization: part based image representation and weakly supervised methods.

A. Part Based Methods

Part representation has been investigated in general image recognition. In [26], over-segmented regions in images are used as parts and LDA (linear discriminant analysis) is used to learn the most discriminative ones for scene recognition. In [27], discriminative parts/modes are selected through the mean shift method on local patches in images for each class. In [28], a set of representative parts are learned using an SVM (support vector machine) classifier with the group sparse constraint for each class in image recognition and segmentation. All these methods tried to evaluate each part, which may be very computationally expensive when the part number is very large.

Part based methods have also been used in fine-grained image categorization for a long time. Detailed part annotations are provided with some datasets like CUB 200-2011 [1], where each bird in the image has 15 part annotations. Some methods, for instance [17], [18], directly extract feature vectors from these annotated parts for recognition. [17] also considers generating parts from aligned objects by dividing each object into several segments and assuming that each segment is a part in the object.

Some works consider a more practical setup when part annotations are missing in the testing phase. They learn part detectors from annotated parts in the training images and apply them on testing images to detect parts. These part detectors include DPM or object classifiers learned for each object class. [19] used selective search to generate object/part proposals from each image, and applied the learned part detectors on them to detect the head and body in the bird. The proposal which yields the highest response to a certain part detector is used as the detected part in the object.

Convolutional neural networks (CNN) have been widely used in image recognition. The outputs from the inner convolutional (CONV) layers can be seen as the feature representations of sub-regions in the image. When CNN is used on an object proposal, the outputs from the inner convolutional layers can be seen as the part representations, e.g., [25] used CNN on detected objects, and used the outputs from CONV4 (in Alexnet) as the parts. [29] used the outputs from all layers in CNN and selected some important ones as parts.

Recently, CNN aided by region proposal methods, has become popular in object recognition/detection, e.g., RCNN [30], fast-RCNN [31], faster-RCNN [32], and RCNN-minus-R [33]. All these four methods focus on the supervised object detection, where object bounding boxes in training images are necessary to learn the object detectors. They cannot be directly used in our weakly-supervised fine-grained image categorization. These methods generate object level representations, while ours used fine-grained part level representations. In RCNN, CNN is applied on each object proposal (bounding box acquired by selective search on the input image) and the output from the fully connected layer is used as the feature vector, where CNN is applied multiple times on an image. In Fast-RCNN, CNN is only applied once on the whole image. The bounding boxes of object proposals are mapped to the final convolutional (CONV) layer to get the object feature. Similarly, RCNN-minus-R used sliding windows to map to the last CONV layer in CNN in order to get the object representation. In Faster-RCNN, instead of mapping object proposal from input images, sliding windows are directly used on the last CONV layer to get the object feature.

Some existing works are related to the proposed method.

The proposed MMP is an efficient way to generate multi-scale part proposals to characterize fine-grained objects. It can be easily applied on millions or even billions of object proposals in a dataset. Unlike [25], where the outputs of CONV4 in CNN are used as parts, MMP provides dense coverage on different scales from part level to object level for each object proposal. The large number of part proposals provide us more opportunity to mine subtle useful information of objects.

Part selection can automatically explore those parts which are important for categorization by only using image-level labels. It is more efficient and practical than trying to learn explicit part detectors without groundtruth object/part annotations. [25] also worked on fine-grained categorization without object/part annotations, which requires much more computation than ours. [25] used two CNN models to detect interesting objects and further learned accurate part detectors from them. In contrast, we only need to select important parts from all part proposals, which are generated by applying one CNN model. More importantly, our method shows that without explicitly detecting the fine-grained objects/parts, the proposed image representation can acquire a better discriminance than [25] (cf. Table III).

ScPM is different from the Multi-scale Pyramid Pooling (MPP) method in [34], where MPP encodes local features from images resized on different scales into separate Fisher vector (FV) [35], and aggregates all the FVs into one to represent an image. Such aggregation may not highlight the subtle differences of object parts on different scales, which is especially important in fine-grained objects with complex backgrounds. In contrast, in ScPM, we automatically select different numbers of important part clusters on different scales using the proposed part selection method described in Sec. III-B. We will also use FV to encode the parts on each scale. The final FV representations from different scales are likely to have different lengths, which cannot be simply aggregated as MPP. We denote the strategy used in MPP as sum pooling, and compare it with the proposed ScPM in the experiment. Spatial pyramid matching (SPM) [36] is also not suitable for fine-grained image categorization. This is because spatial correspondence does not necessarily exist among manually split regions in fine-grained images, which may cause possible spatial mismatching problems [37].

B. Weakly Supervised Fine-grained Categorization

Most existing fine-grained works heavily rely on the object / part annotations in categorization when the objects are in complex backgrounds. [25] is the first work which categorizes fine-grained images without using human annotations in any image (both training and testing), but with only image labels. In [25], a CNN that is pre-trained from ImageNet is first used as an object detector to detect the object from each image. Then, part features (outputs from CONV4 in CNN) are extracted from objects and clustered into several important ones by spectral clustering. For each part cluster, a part detector is learned to differentiate it from other clusters. Finally, these part detectors are used to detect useful parts in testing images. In [25], each part is evaluated extensively by the learned part detectors and the detected ones are concatenated into the final image representation. In contrast, our method first encodes the large number of parts into a global image representation and then performs part selection on it, which can save much more computational effort than [25].

[29] also categorized fine-grained images in the same setup. They first generated a pool of parts by using the outputs from all layers in CNN. Then, they selected useful ones for categorization. They consider two ways of selection: one is to randomly select some parts; the other is to select a compact set by considering the relationship among them. These parts are concatenated to represent the image.

[38] learned to detect and align objects in an end-to-end system. This system includes two parts: one is an object detector, which is followed by a spatial transformer. The spatial transformer is learned to align the detected objects automatically to make the parts match accurately.

This paper is different from [25], [29], [38], in that, we do not explicitly detect/align the object/part in the image, but propose an efficient part selection method to extract the most discriminative information for categorization.

III. Fine-grained image representation without using object / part annotations

The proposed part-based image representation includes three parts: part proposal generation, part selection, and multi-scale image representation, which are detailed in Sections III-A to III-C, respectively.

A. Part Proposal Generation

Regional information has been shown to improve image classification with hand-crafted methods like spatial pyramid matching [36] and receptive fields [39]. When a CNN model is applied on an image, features of local regions can be acquired automatically from its internal structure. Assume the output from a layer in CNN is $N \times N \times d$ dimension, which is the output of d filters for $N \times N$ spatial cells. Each spatial cell is computed from a receptive field in the input image. The receptive fields of all the spatial cells in the input image can highly overlap with each other. The size of one receptive field can be computed layer by layer in CNN. In a convolution (pooling) layer, if the filter (pooling) size is $a \times a$ and the stride is s, then $T \times T$ cells in the output of this layer corresponds to $[s(T-1)+a] \times [s(T-1)+a]$ cells in the input of this layer. For example, one cell in the CONV5 (the 5th convolutional) layer of CNN model (imagenet-vgg-m) [40] corresponds to a 139×139 receptive field in the 224×224 input image (cf. Fig. 4).

We generate features of multi-scale receptive fields for an image by leveraging the internal outputs of CNN with little additional computational cost (cf. Fig. 5). Considering the outputs of one layer in CNN, we can pool the activation vectors of adjacent cells of different sizes, which correspond to receptive fields with different sizes in the input image. Max-pooling is used here.

In particular, given the $N \times N \times d$ output X in one layer in CNN, we use max-pooling to combine information from all

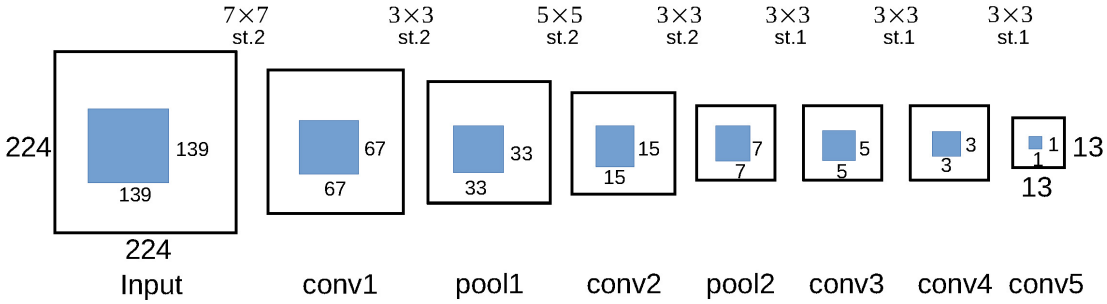


Fig. 4. Receptive fields computed using the CNN model (imagenet-vgg-m) [40]. One cell in the CONV5 layer corresponds to a 139×139 receptive field in the input image. We only show the spatial sizes of the image and filters, where $a \times a$ is the filter (pooling) size, and 'st' means the stride.

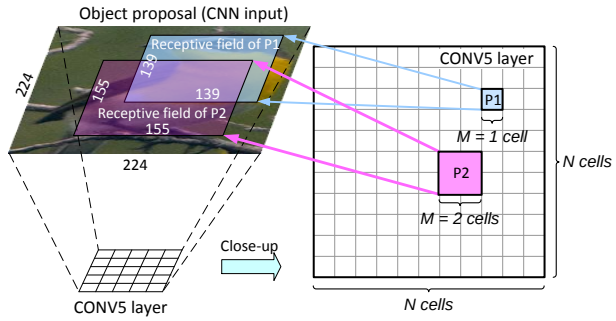


Fig. 5. Generating multi-scale part proposals. For an input object proposal, by applying CNN on it, spatial cells of different sizes on the CONV5 layer in CNN correspond to parts of different scales. This figure is best viewed in color.

$M \times M$ adjacent cells, that is:

$$z^M_{i,j,k} = \max_{\substack{i \le p < i+M, \\ j \le q < j+M}} X_{p,q,k}, \quad \text{s.t.} \;\; 1 \le M \le N, 1 \le k \le d, \tag{1}$$

where M ranges from 1 (single cell) to N (all the cells). In Eq. 1, an $M \times M$ spatial neighborhood is represented by a d-dimensional feature mapping $\boldsymbol{z}^M$. When M is assigned to different values, the corresponding cells can cover receptive fields of different sizes (scales) in the input image, thus providing a more comprehensive information. We name this proposed part proposal generation strategy as multi-max pooling (MMP) and apply it to the CONV5 layer (last CONV layer in CNN). This is because the CONV5 layer can capture more meaningful object/part information than those shallow layers in CNN [41]. When a CNN model is applied on an object bounding box in an image, the acquired receptive fields from MMP can be seen as the part candidates for the object. Thus, we can acquire a multi-scale representation of parts in objects with MMP.

To compute the part proposals, we first generate object proposals from each image. Object proposals are those regions inside an image that have high *objectness*, i.e., having a higher chance to contain an object. Since no object/part annotations are utilized, we could only use unsupervised object detection methods. Selective search [42] is used in our framework given its high computation efficiency, which has also been used in [30], [19] to generate initial object/part candidates for object detectors. After generating multiple object proposals, we apply the CNN model on each bounding box/object proposal, and use the proposed MMP to get a large number of part proposals from each object proposal.

B. Part Selection

We then propose to select useful (i.e., discriminative) part clusters, and form a global representation from these useful parts in each image.

Among the object/part proposals, most of them are from background clutters, which are harmful for image recognition. For example, in the CUB200-2011 [1] dataset, when we use the intersection over union criteria, only 10.4% object proposals cover the foreground object. The part proposals from those unsuccessful object proposals will contribute little to the classification, or even be noisy and harmful. Thus, we need to find those useful part proposals (discriminative parts of the foreground object) for our final image representation.

Our basic idea is to select useful parts through mining the useful information in part clusters. We first cluster all part proposals in the training set into several groups. Then, we compute the importance of each cluster for image classification. Those part proposals assigned to the useful clusters (clusters with the highest importance values) are selected as the useful parts.

We compute the cluster importance with the aid of Fisher vector (FV) [35].[1] We first encode all the part proposals in each image into a FV with a GMM (Gaussian Mixture Model). The GMM is learned using part proposals extracted from training images. Each Gaussian corresponds to a part cluster. Then, for each dimension in FVs of all training images $\boldsymbol{x}_{:i}$, we compute its importance using its mutual information (MI) with the class labels $\boldsymbol{y}$ [45]. [45] shows that different dimensions in FV have weak correlations, which advocates processing each dimension separately. The MI value of each dimension $\boldsymbol{x}_{:i}$ in FV is computed as:

$$I(\boldsymbol{x}_{:i}, \boldsymbol{y}) = H(\boldsymbol{y}) + H(\boldsymbol{x}_{:i}) - H(\boldsymbol{x}_{:i}, \boldsymbol{y}), \tag{2}$$

where H is the entropy of a random variable. Since $\boldsymbol{y}$ remains unchanged for different i, we simply need to compute $H(\boldsymbol{x}_{:i}) - H(\boldsymbol{x}_{:i}, \boldsymbol{y})$. In order to compute the value distribution

[1] VLAD can be used in our framework, which is used in [43] to encode CNN of multiple spatial regions for general image classification. We choose FV because it has a better discriminance than VLAD [44].

of $\boldsymbol{x}_{:i}$ in Eq. 2, an efficient 1-BIT quantization method [45] is used. For a scaler x in $\boldsymbol{x}_{:i}$, it is quantized according to

$$x \leftarrow \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}. \tag{3}$$

Finally, the cluster (Gaussian) importance is the summation of the MI values of all FV dimensions computed from this Gaussian. For a Gaussian G, its importance is computed as:

$$m(G) = \sum_{i \in G} I(\boldsymbol{x}_{:i}, \boldsymbol{y}). \tag{4}$$

We only keep those dimensions in FV from the most important Gaussians with the largest importance values. As will be shown in Sec. IV, this novel strategy greatly improves categorization accuracy, even when object or part annotations are not used at all.

C. Multi-scale Image Representation

Considering our part proposals are generated at different scales (with different M in Eq. 1), aggregating all of them into a single image representation cannot highlight the subtle distinction in fine-grained images. Thus, we propose to encode part proposals in an image on different scales separately and we name it SCale Pyramid Matching (ScPM). The steps are as follows:

- **Generate parts on different scales**. Given an image I, which contains a set of object proposals $I = \{o_1, ..., o_{|I|}\}$, each object proposal o_i contains a set of multi-scale part proposals $o_i = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_{|o_i|}\}$. For part proposals in I on different scales $M \in \{1, ..., N\}$, we compute separate FVs. In practice, the scale number can be very large ($N = 13$ in the CNN setting), which may lead to a severe memory problem. Since the part proposals on neighboring scales are similar in size, we can divide all the scales into m ($m \leq N$) non-overlapping groups $\{g(j), j = 1, \ldots, m, g(j) \subseteq \{1, \ldots, N\}\}$.
- **Compute FV using selected parts on each scale**. For an image I, its part proposals belonging to the scale group $g(j)$ are used to compute one FV $\phi_j(I)$ as:

$$\phi_j(I) = [f_{\boldsymbol{\mu}_1^j}(I), f_{\boldsymbol{\sigma}_1^j}(I), \ldots, f_{\boldsymbol{\mu}_i^j}(I), f_{\boldsymbol{\sigma}_i^j}(I), \ldots], \tag{5}$$

$$f_{\boldsymbol{\mu}_i^j}(I) = \frac{1}{\sqrt{w_i^j}} \sum_{c(t) \in g(j)} \gamma_t^j(i) \left(\frac{\boldsymbol{z}_t^{c(t)} - \boldsymbol{\mu}_i^j}{\boldsymbol{\sigma}_i^j} \right), \tag{6}$$

$$f_{\boldsymbol{\sigma}_i^j}(I) = \frac{1}{\sqrt{2w_i^j}} \sum_{c(t) \in g(j)} \gamma_t^j(i) \left[\frac{(\boldsymbol{z}_t^{c(t)} - \boldsymbol{\mu}_i^j)^2}{(\boldsymbol{\sigma}_i^j)^2} - 1 \right], \tag{7}$$

  where $\{w_i^j, \boldsymbol{\mu}_i^j, \boldsymbol{\sigma}_i^j\}$ are the mixture weights, mean vectors, and standard deviation vectors of the i-th selected diagonal Gaussian in the j-th scale group $g(j), j = 1, \ldots, m$, respectively. $\{\boldsymbol{z}_t\}$ are the selected part proposals in an image, $c(t)$ is the scale index of the t-th part and $\gamma_t^j(i)$ is the weight of the t-th instance to the i-th Gaussian in the j-th scale group.
- **Image representation**. Following [35], two parts corresponding to the mean and the standard deviation in each Gaussian of FV are used. Each of the m FVs is power and ℓ_2 normalized independently, and then concatenated to represent the whole image as $\phi(I)$:

$$\phi(I) = [\phi_1(I), \ldots, \phi_m(I)]. \tag{8}$$

- **Feature normalization**. Because of the ℓ_2 normalization, each $\phi_i(I)$ satisfies that $\|\phi_i(I)\|_2 = 1$. After part selection, however, this property ceases to hold. Because only a few parts are selected, we expect $\|\phi_i(I)\|_2 < 1$ for all $1 \leq i \leq m$. Data normalization has been shown to effectively improve the discriminative power of a representation [46]. For the image representation after part selection, we apply power normalization and ℓ_2 normalization again.

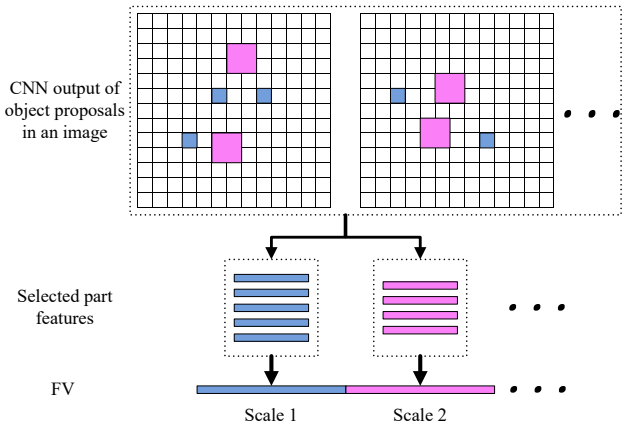


Fig. 6. The process of generating image representation using ScPM.

The whole process is illustrated in Fig. 6.

IV. Experiments

In this section, we evaluate the proposed weakly-supervised method for fine-grained categorization. The selective search method [42] with default parameters is used to generate object proposals for each image. The pre-learned CNN models [40] from ImageNet are used to extract features from each object proposal as [30], which has been shown to achieve state-of-the-art results. It is fine-tuned with training images and their labels. We would like to point out that we do not fine tune CNN using object proposals because many of them are from background clutters, which may deteriorate the CNN performance. We use the 'imagenet-vgg-m' model [40], given that its efficiency and accuracy are both satisfactory. It has a similar structure (with 5 convolutional layers) to that of AlexNet [47].

The part proposals in each scale group are assigned into 128 clusters. Each part feature is reduced into 128 dimensions by PCA. All 13 part scales ($N = 13$ in the CNN model) are divided into 8 scale groups: the first 4 scales form the first 4 groups, the subsequent 6 scales form 3 groups with 2 scales in one group, and the last 3 scales form the last scale group. This arrangement makes the number of parts in each group roughly balanced. The dimension of the global image representation using FV becomes: $128 \times 2 \times 128 \times 8 = 262144$, from which different fractions of useful part clusters will be selected and evaluated.

TABLE I
EVALUATION OF DIFFERENT MODULES IN THE PROPOSED IMAGE REPRESENTATION ON CUB 200-2011 DATASET.

	Accuracy (%)
CONV5+MMP+ScPM	75.47
CONV5+MMP	73.14
CONV5	63.75
whole image	54.68

We evaluate the proposed method on three benchmark fine-grained datasets:

- **CUB200-2011 [1]**: The Caltech-UCSD Birds 200-2011 dataset contains 200 different bird classes. It includes 5994 training images and 5794 testing images.
- **StanfordDogs [5]**: This dataset contains 120 different types of dogs and includes 20580 images in total.
- **VMMR-40 [11]**: It contains 928 classes. Each class has at least 40 images. The dataset contains 78651 images in total. We use 20 images in each class for training and the rest for testing.

For all datasets, we only use the class labels of images in the training stage.

We choose LIBLINEAR [48] to learn linear SVM classifiers for classification. All the experiments are run on a computer with Intel i7-3930K CPU, 64G main memory, and an Nvidia Titan GPU.

A. Influences of Different Modules

We evaluate different modules in the proposed part based image representation (without part selection) on the CUB 200-2011 dataset in Table I:

- The effect of MMP in the proposed image representation. We compare the part proposals generated using the outputs of CONV5 and CONV5+MMP. All part proposals in each image are encoded into one FV (without part selection and ScPM). It can be seen that multi-scale part proposals (CONV5+MMP) can greatly improve the recognition accuracy over single-scale part proposals (CONV5) by about 10%. This is because MMP can provide very dense coverage of object parts at different scales. The part based image representation is also shown to be significantly better than the object based image representation.
- The influence of ScPM in the proposed image representation. Using the multi-scale part proposals generated by MMP, ScPM achieves a better accuracy (2.3% higher) than that of the method encoding all part proposals altogether. This shows that it is beneficial to encode parts at different scales separately.
- Evaluation of the global image representation using CNN, indicated as 'whole image' in Table I. The CNN model is applied on the whole image, which is represented using the output of FC7. It leads to a significantly worse accuracy rate than our part based method.
- We evaluate the proposed multi-scale image representation with different numbers of GMMs in Table II. The classification accuracy increases when the number of GMMs increases. After the GMM number exceeds 128, the accuracy improvement becomes slower. As a tradeoff between the accuracy and computational efficiency (including both memory footprint and computation time), we use 128 GMMs in the following experiments as the default value.
- We compare ScPM with the sum pooling method used on FV [34] in Table II. ScPM shows better classification results than the sum pooling [34] when different GMMs are used in FV. This is because ScPM can highlight the difference of fine-grained objects on various scales.

TABLE II
CLASSIFICATION ACCURACY (%) OF THE PART BASED IMAGE REPRESENTATION WITH DIFFERENT NUMBERS OF GMM.

#GMM	ScPM	Sum pooling
16	72.09	68.24
32	72.93	69.11
64	74.12	70.25
128	75.47	71.58
256	75.50	71.56

TABLE III
CLASSIFICATION ACCURACY COMPARISONS ON CUB 200-2011 DATASET USING VGG-CNN-M MODEL.

Without annotations in either training or testing		
Methods	Selection fraction	Acc. (%)
Proposed	100% (All)	75.47
	75.0% (3/4)	76.02
	50.0% (1/2)	77.71
	25.0% (1/4)	**78.92**
	12.5% (1/8)	77.89
Proposed (AlexNet, selection fraction 1/4)		75.29
Feature selection (selection fraction 1/4) [45]		77.54
Two-level attention (AlexNet) [25]		69.70
Two-level attention (VGG verydeep) [25]		77.90
Activation Constellation (AlexNet) [29]		68.50
Activation Constellation (VGG verydeep) [29]		81.01
Spatial Transformer [38]		**84.10**
Use annotations in training, not in testing		
DPD+DeCAF [20]		44.94
Part based R-CNN (without parts) [19]		52.38
Part based R-CNN-ft (without parts) [19]		62.75
CL-45C (without parts) [49]		73.50
Part based R-CNN-ft (with parts) [19]		73.89
Pose Normalized CNN [50]		75.70
Co-segmentation [22]		82.80

We summarized the observations from the above evaluations. First, MMP+ScPM can compute an efficient multi-scale part representation. Second, ScPM is better than the sum pooling when pooling multiple FVs into a global representation. Finally, we fix the the number of Gaussian components in GMM as 128 when computing FV in the following experiments. In the following section, we will show that the proposed part selection can further improve the accuracy.

B. Part Selection

We show the classification accuracy using part selection on the proposed image representation (MMP+ScPM) for CUB 200-2011 in Table III.

It can be seen that part selection can greatly improve accuracy. We show the results corresponding to selecting different fractions of part clusters in the image representation.

TABLE IV
CLASSIFICATION ACCURACY ON CUB 200-2011 WITH OBJECT BASED IMAGE REPRESENTATION.

Methods	Selection fraction	Acc. (%)
Object based FV	100% (All)	60.13
	75.0% (3/4)	61.60
	50.0% (1/2)	**62.89**
	25.0% (1/4)	60.63
	12.5% (1/8)	52.78
Fast-RCNN [31]		**63.41**

TABLE V
CLASSIFICATION ACCURACY (%) ON CUB 200-2011 DATASET USING THE METHOD IN [26] ON THE SAME PARTS IN TABLE III.

#part classifiers	80	120	160	200
Acc. (%)	67.42	72.76	75.08	75.13

When selecting the most important quarter of the part clusters (fraction 25%), a peak is reached, and it is better than the one without part selection (fraction 100%) by 3.5%. Even when fewer part proposals are selected (fraction 12.5%), its accuracy is still better than the one without part selection by 2.4%. This shows that part selection can efficiently reduce the noise introduced by those part proposals from background clutters. We also compare part selection with feature selection [45] on the same feature representation with the same selection fraction (25%). Feature selection (77.54%) is worse than part selection (78.92%). This is because part selection can keep more semantic information of parts.

As a comparison to our proposed part based image representation, we evaluated an object based image representation for fine-grained image categorization. We applied CNN on each object proposal and extracted the output from the FC7 layer as the object feature (reduced to 128 by PCA). The objects in each image were encoded into a FV with 128 GMMs. We applied feature selection [45] on the FVs and computed their classification accuracies. The results are shown in Table IV. When the background noise is discarded with different selection fractions, the classification can be improved to the highest 62.89% on the object based image representation. We also evaluated the object generation method using fast-RCNN [31], i.e., mapping object proposals to the last CONV layer in CNN to get the object features. The object features are encoded into FV and applied feature selection (25% fraction), which has 63.41% accuracy. Although their computation can be faster, they have much lower accuracy than our part based image representation.

Our best accuracy (78.92%) significantly outperforms the state-of-the-art weakly-supervised methods [25], [29] by over 9% and 10% respectively when similar CNN models (vgg-cnn-m and AlexNet) are used. With a deeper and more powerful CNN model (vgg-verydeep), [25] reduces the gap to ours to 1% while [29] achieves higher accuracy. Note that, in addition to the high complexity of using the very deep CNN model, [29] is expensive because it needs to evaluate each part to select the best ones. In contrast, ours only selects best part clusters, which has a much smaller number than that of parts. [38] achieves much higher results than other works because they used a more powerful baseline CNN structure. We also compared with the 'blocks that shout' method [26] on our parts used in Table V. Useful parts are selected through learned part classifiers and then encoded into a FV for each image.The accuracy does not improve when more part classifiers are used, which is also lower than ours in Table III.

TABLE VI
CLASSIFICATION ACCURACY (%) ON CUB 200-2011 DATASET USING VLAD. FV RESULTS ARE ALSO CITED FOR COMPARISON.

Without annotations in either training or testing			
Methods	Selection fraction	FV	VLAD
Proposed	100% (All)	75.47	73.82
	75.0% (3/4)	76.02	74.13
	50.0% (1/2)	77.71	76.21
	25.0% (1/4)	**78.92**	**77.09**
	12.5% (1/8)	77.89	76.42

We also show the accuracy of annotation-dependent methods using object / part annotations in the training stage but not in the testing stage, which use the least annotations and are closest to our weakly-supervised setup. Most of these methods try to learn expensive part detectors to get accurate matching for recognition. However, the superior performance of our method shows that they are not always necessary, especially in weakly-supervised fine-grained categorization.

We would like to highlight that part selection is more important in fine-grained categorization than feature selection in general image categorization. With part selection, the accuracy is 3.5% (78.92% vs. 75.47%) higher than the original image representation. In [45], feature selection is used to compress FV for general image recognition like object recognition. Much smaller (around 1%) improvement after selection (worse in most time) is achieved over the original FV, which is significantly different from the improvement observed in Table III. This fact clearly shows the distinction between the two applications. In the weakly-supervised fine-grained tasks, selecting proper object parts is critical, while in general image recognition, the global image representation without selection is usually already good.

We also compare the proposed image representation (using FV) with using VLAD [43]. The classification accuracy using VLAD is shown in Table VI. VLAD leads to inferior results than FV using different selection fractions. On each selection fraction, the accuracy of VLAD is about 2% worse than that of FV. In the following experiments, we will only use FV in the proposed image representations.

We further evaluate the proposed method with the very deep CNN model (VGG-verydeep-16) [51]. The classification results are shown in Table VII. The very deep CNN model has 13 convolutional layers. It has a much deeper structure than our previously used CNN model (the vgg-m model), which has only 5 convolutional layers. Thus, the very deep CNN model can provide more discrimination in image recognition tasks. We also use the outputs from the layer before the last convolutional layer in our method. We find that the very deep CNN model has better results than the shallow model (77.28% vs. 75.47%), when part selection is not used. However, after part selection is used, the difference shrinks, where the best classification accuracies of the two models are 79.34% vs. 78.92%. This shows that a weak (shallow) CNN

TABLE VII
CLASSIFICATION ACCURACY (%) ON CUB 200-2011 DATASET USING VGG-VERYDEEP-16 CNN MODEL.

Without annotations in neither training nor testing			
Methods	Selection fraction	Verydeep	Shallow
Proposed	100% (All)	77.28	75.47
	75.0% (3/4)	77.99	76.02
	50.0% (1/2)	**79.34**	77.71
	25.0% (1/4)	78.32	78.92
	12.5% (1/8)	77.65	77.89

model can benefit from part selection in the proposed image representation. Besides, the very deep CNN model introduces much more computation than the shallow model. Thus, in the following experiments, we will only use the shallow CNN model (imagenet-vgg-m) in the proposed method.

We evaluate the time cost in each module of the proposed method on CUB 200-2011. The image representation generation time is 3.4 seconds per image, where CNN costs 0.9 second, part generation 2.3 seconds, FV 0.2 second. The cost of part selection is almost negligible. Learning 8 GMMs (for 8 scales) costs about 1 hour (using 1/5 training images). Learning part selection parameters costs 1500 seconds. SVM classifiers take 40 minutes during training and 5 minutes during testing for features without part selection. With part selection, the time is proportionally reduced with respect to the selection fraction.

Overall, these results show that: 1) part selection is important in weakly-supervised fine-grained categorization; 2) it is not always necessary to learn expensive object/part detectors in fine-grained categorization; 3) a very deep CNN model is not necessary in extracting parts when part selection is used; and 4) FV is better than VLAD in generating the image representation.

C. Understand Subtle Visual Differences: with the help of Key Part Detection

We want to detect and show the key (most discriminative) parts in fine-grained images of different classes to give a more insightful understanding of the critical property in objects, which may help us in feature design for fine-grained images.

We learn a binary SVM (support vector machine) classifier in each selected part cluster to compute the part score. This classifier is used to propagate the image labels to parts. In the training phase, for each selected part cluster, we aggregate the part features in one image assigned to this cluster altogether (similar to VLAD). The aggregated features of training images are ℓ_2 normalized and are then used to train a classifier with image labels. In the testing phase, given a part, its score is computed as the dot-product between the classifier for the part cluster it falls in (only considering those parts in the selected part clusters) and its feature (the CNN activation vector). Note that in both training and testing processes, the part features are centered (i.e., minus the cluster center in each part cluster).

Fig. 7 shows parts that belong to two clusters. The parts are sorted according to their importance scores in descending order. We can see that parts in the same cluster are relatively coherent, corresponding mainly to the head region of the two species of birds.

Fig. 7. Part variations. Parts are from two different part clusters. They are shown according to their importance scores in the descending order within each part cluster.

Then, we show more examples of key part detection in Fig. 8. In each pair, we show one sample image and 20 detected key parts with the highest (smallest) scores from all testing images of the positive (negative) class. The bird names are given in the captions, which clearly indicates how humans characterize different birds.

It can be seen that the detected parts capture well the key parts in these species, which are consistent with human-defined rules. We also find that the proposed method can capture some tiny distinction that might not be easily discriminated by human eyes. For example, in the first pair, the key parts in the red-bellied woodpecker and red-headed woodpecker are both red, and the locations are very close. From the detected parts, we can find that the red color of the red-headed woodpecker is darker and the feather of red-bellied woodpecker is finer.

From the detected parts, we can also understand the necessity to select many useful parts in the proposed image representation. Only using the best part may cause possible loss of useful information in characterizing an object. Multiple good parts can compensate each other from different aspects like location, view, scale, etc. This also explains why the proposed representation works better than [25], which only uses the detected best part for categorization.

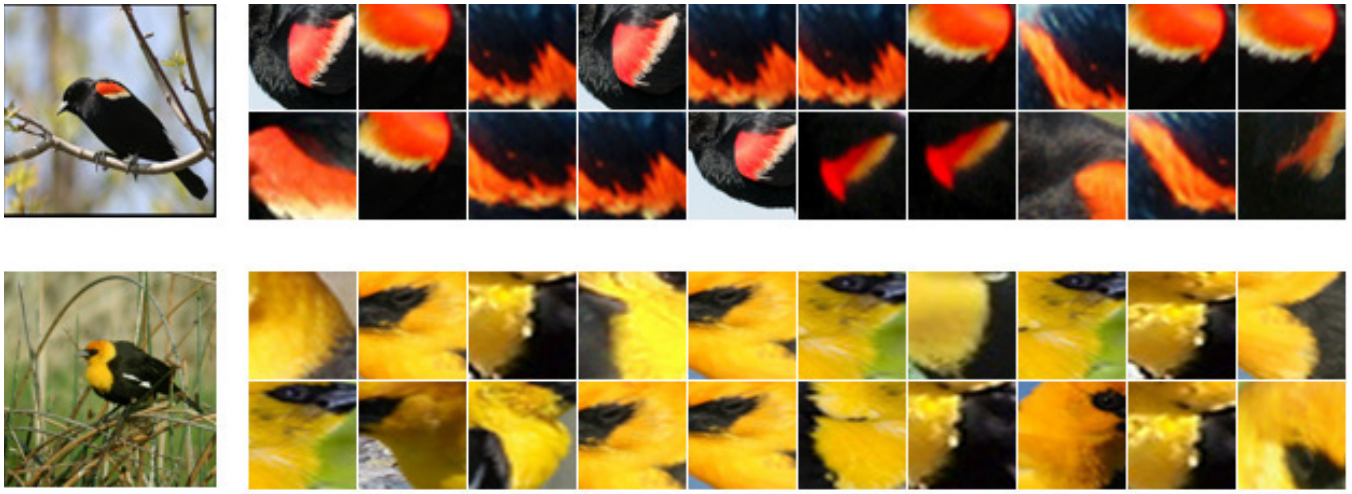
D. Classification Results on Stanford Dogs

We show the categorization accuracy for Stanford Dogs in Table VIII. The proposed method (either with or without part selection) shows much better accuracy than the existing annotation-dependent works. Part selection also plays an important role in the proposed image representation, which leads to a 2.69% improvement over the original representation. Stanford Dogs is a subset in ImageNet. It is also evaluated in state-of-the-art weakly-supervised works [25], [29], whose results are significantly lower than ours.

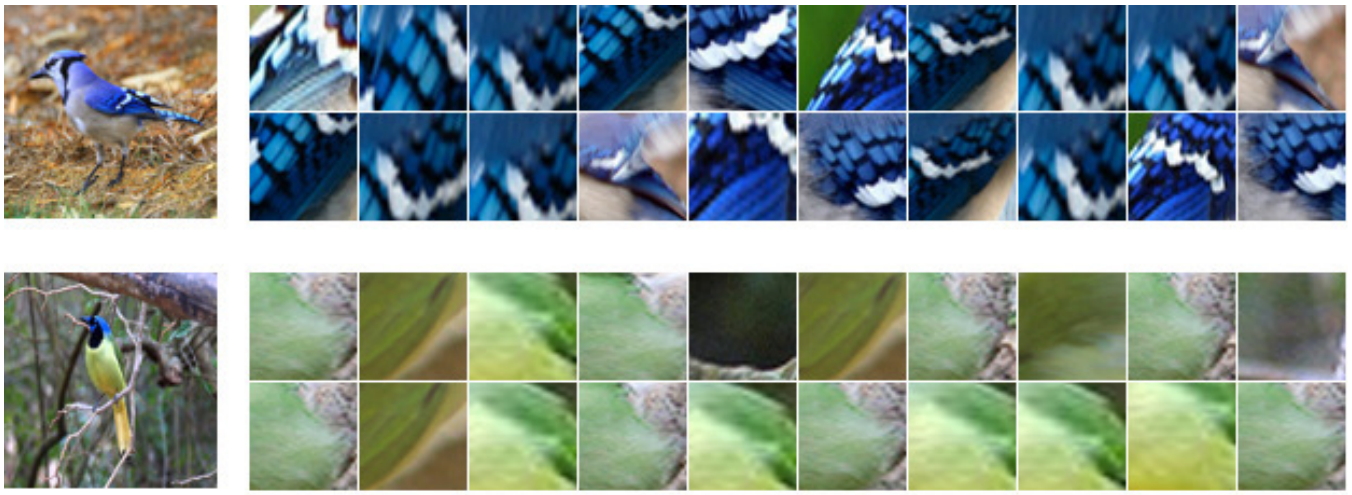
E. Classification Results on VMMR-40

VMMR-40 is a recently released large-scale dataset for car recognition. The images are captured from different angles by different users and devices. The cars are not well aligned. Some images contain irrelevant backgrounds. We show the

(a) Red-bellied Woodpecker vs. Red-headed Woodpecker



(b) Red-winged Blackbird vs. Yellow-headed Blackbird



(c) Blue Jay vs. Green Jay

Fig. 8. Key (most discriminative) parts visualization for pairwise classes. Key parts are detected from testing images using the classifier learned from training images. Top 20 key parts are shown for each class. The important parts found by the proposed method coincide well with the rules human experts use to distinguish these birds. This figure is best viewed in color.

TABLE VIII
CLASSIFICATION ACCURACY ON STANFORDDOGS.

Without annotations in either training or testing		
Methods	Selection fraction	Acc. (%)
Proposed	100% (All)	77.74
	75.0% (3/4)	78.52
	50.0% (1/2)	79.66
	25.0% (1/4)	**80.43**
	12.5% (1/8)	78.38
Two-level attention [25]		71.90
Activation Constellation (AlexNet) [29]		68.61
Use annotations in both training and testing		
Edge templates [52]		38.00
Unsupervised alignments [17]		50.10
MTL [53]		39.30

Fig. 9. Key part detection of two models in VMMR-40 dataset [11]: acura-integra-1991 and acura-integra-1994.

classification accuracy in Table IX. We first test the classification accuracy using the CNN FC7 feature extracted from the whole image. Then, we test our proposed part based image representation with different part selection fractions. We can see that the performance of the part based image representation greatly outperforms that of the whole image representation. Part selection does not improve as much accuracy as those observed in the previous two datasets. This is because the backgrounds in VMMR-40 images are less complex than those in CUB 200-2011 and Standford Dogs. The classification results validate the capability of the proposed method in characterizing the unaligned fine-grained objects in complex backgrounds. We also show the detected key parts in Fig. 9. We can see that the main difference of the two models lie in the rear lights, which are accurately detected.

F. Discussions

The major argument of this paper is that part selection is a more natural and efficient choice than using part detectors in weakly-supervised fine-grained image categorization. Particularly, we find that:

- It is hard to learn accurate part detectors to align objects without object / part annotations in fine-grained image categorization (cf. Table III).
- Multi-scale part representation is important to characterize fine-grained objects on different scales (cf. Table I).
- Selecting multiple good parts is better than detecting one best part in fine-grained object recognition (cf. Table III and Fig. 8).
- Selected parts are discriminative for categorization by discarding the background noise in images (cf. Fig. 8).

We have provided the following methods for efficient representation of fine-grained objects in the weakly-supervised setup:

- Multi-max pooling (MMP) is an efficient way to generate multi-scale part proposals from the CNN outputs on object proposals.
- Part selection is necessary to reduce the background noise in images, which is more efficient than those methods trying to learn accurate object / part detectors.
- Encoding useful part proposals on different scales separately (ScPM) can highlight the subtle distinctions among fine-grained objects.

In our experience, there is one issue with the proposed framework: the part proposal generation process may introduce heavy computations, when the numbers of images and object proposals are very large in the dataset. Our part proposals are generated from CNN which is applied on each object proposal. It is important to research on how to reduce the number of effective object proposals (so that we can reduce the times of CNN applied on object proposals) or how to generate part proposals directly from CNN computed on images.

TABLE IX
CLASSIFICATION ACCURACY ON VMMR-40.

Without annotations in either training or testing		
Methods	Selection fraction	Acc. (%)
Proposed	100% (All)	40.12
	75.0% (3/4)	**40.58**
	50.0% (1/2)	39.97
	25.0% (1/4)	39.10
Whole image		25.93

V. CONCLUSIONS

In this paper, we have proposed to categorize fine-grained images without using any object/part annotation either in the training or in the testing stage. Our basic idea is to select multiple useful parts from multi-scale part proposals and use them to compute a global image representation for categorization. This is specially designed for fine-grained categorization in the weakly-supervised scenario, because parts have been shown to play an important role in the existing annotation-dependent works. Also, accurate part detectors are usually hard to acquire. Particularly, we propose an efficient multi-max pooling strategy to generate multi-scale part proposals by using the internal outputs of CNN on object proposals in each image. Then, we select useful parts from those part clusters which are important for categorization. Finally, we encode the selected parts at different scales separately in a global image representation. With the proposed image / part representation technique, we use it to detect the key parts of objects in different classes, whose visualization results are intuitive and coincide well with rules used by human experts.

In the experiments, on three challenging datasets, our proposed weakly-supervised method achieves comparable or better results than those of state-of-the-art weakly-supervised

works [25], [29] and most existing annotation-dependent methods. Future works would include utilizing the part information mined from the global image representation to help localize objects and further improve classification.

REFERENCES

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 2019 – 2026.

[3] A. Iscen, G. Tolias, P.-H. Gosselin, and H. Jegou, "A comparison of dense region detectors for image search and fine-grained classification," *IEEE Trans. on Image Processing*, vol. 24, no. 8, pp. 2369–2381, 2015.

[4] L. Xie, Q. Tian, M. Wang, and B. Zhang, "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. on Image Processing*, vol. 23, no. 5, pp. 1994–2008, 2014.

[5] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.

[6] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, B. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed, "Understanding objects in detail with fine-grained attributes," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 3622–3629.

[7] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conf. on Computer Vision, Graphics and Image Processing*, 2008, pp. 722–729.

[8] A. R. Sfar, N. Boujemaa, and D. Geman, "Vantage feature frames for fine-grained categorization," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 835–842.

[9] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," *IEEE Trans. on Image Processing*, vol. 23, pp. 623–634, 2014.

[10] E. Rodner, M. Simon, G. Brehm, S. Pietsch, J. W. Wagele, and J. Denzler, "Fine-grained recognition datasets for biodiversity analysis," in *Third Workshop on Fine-Grained Visual Categorization (FGVC3), CVPRW*, 2015.

[11] A. Ben Khalifa and H. Frigui, "A dataset for vehicle make and model recognition," in *Third Workshop on Fine-Grained Visual Categorization (FGVC3), CVPRW*, 2015.

[12] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2011, pp. 161–168.

[13] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for subcategory recognition," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 3665–3672.

[14] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proc. European Conf. Computer Vision*, vol. LNCS 6316, 2010, pp. 168–181.

[15] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 729–736.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[17] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2013, pp. 1713–1720.

[18] C. Goring, E. Rodner, A. Freytag, and J. Denzler, "Nonparametric part transfer for fine-grained recognition," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 2489–2496.

[19] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. European Conf. Computer Vision*, vol. LNCS 8689, 2014, pp. 834–849.

[20] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int'l Conf. on Machine Learning*, 2014, pp. 647–655.

[21] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

[22] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2015.

[23] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 580 – 587.

[24] C. Wah, G. V. Horn, S. Branson, S. Maji, P. Perona, and S. Belongie, "Similarity comparisons for interactive fine-grained categorization," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 859–866.

[25] 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 *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2015.

[26] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2013.

[27] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. Advances in Neural Information Processing Systems*, 2013.

[28] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2013.

[29] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2015, pp. 1143–1151.

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[31] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2015.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems*, 2015.

[33] K. Lenc and A. Vedaldi, "R-cnn minus r," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[34] D. Yoo, S. Park, J.-Y. Lee, and I. S. Kweon, "Fisher kernel for deep neural activations," arXiv:1412.1628v2, 2014.

[35] 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.

[36] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2006.

[37] Y. Zhang, J. Wu, J. Cai, and W. Lin, "Flexible image similarity computation using hyper-spatial matching," *IEEE Trans. on Image Processing*, vol. 23, pp. 4112–4125, 2014.

[38] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Advances in Neural Information Processing Systems*, 2015.

[39] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 3370–3377.

[40] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. British Machine Vision Conference*, 2014.

[41] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. European Conf. Computer Vision*, vol. LNCS 8689, 2014, pp. 818–833.

[42] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, pp. 154–171, 2013.

[43] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activations features," in *Proc. European Conf. Computer Vision*, vol. LNCS 8695, 2014, pp. 392–407.

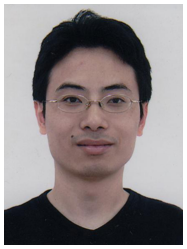
[44] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[45] Y. Zhang, J. Wu, and J. Cai, "Compact representation for image classification: To choose or to compress?" in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 907–914.

[46] J. Wu, Y. Zhang, and W. Lin, "Towards good practices for action video encoding," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 2577–2584.
[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012.
[48] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, Aug 2008.
[49] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2015.
[50] S. Branson, G. V. Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *Proc. British Machine Vision Conference*, 2014.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
[52] S. Yang, L. Bo, J. Wang, and L. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 3131–3139.
[53] J. Pu, Y.-G. Jiang, J. Wang, and X. Xue, "Which looks like which: Exploring inter-class relationships in fine-grained visual categorization," in *Proc. European Conf. Computer Vision*, vol. LNCS 8691, 2014, pp. 425–440.

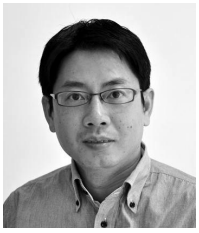
Yu Zhang received his BS and MS degrees in telecommunications engineering from Xidian University, China, and his PhD degree in computer engineering from Nanyang Technological University, Singapore. He is currently a postdoctoral fellow in the Bioinformatics Institute, A*STAR, Singapore. His research interest is computer vision.

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. 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 ICCV 2015 and senior PC member for AAAI 2016. His research interests are computer vision and machine learning. He is a member of the IEEE.

Jianfei Cai (S'98-M'02-SM'07) received his PhD degree from the University of Missouri-Columbia. He is currently an Associate Professor and has served as the Head of Visual & Interactive Computing Division and the Head of Computer Communication Division at the School of Computer Engineering, Nanyang Technological University, Singapore. His major research interests include computer vision, visual computing and multimedia networking. He has published more than 170 technical papers in international journals and conferences. He has been actively participating in program committees of various conferences. He has served as the leading Technical Program Chair for IEEE International Conference on Multimedia & Expo (ICME) 2012 and the leading General Chair for Pacific-rim Conference on Multimedia (PCM) 2012. Since 2013, he has been serving as an Associate Editor for IEEE Trans on Image Processing (T-IP). He has also served as an Associate Editor for IEEE Trans on Circuits and Systems for Video Technology (T-CSVT) from 2006 to 2013.



Jiangbo Lu (M'09-SM'15) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering, Katholieke Universiteit Leuven, Leuven, Belgium, in 2009. From April 2003 to August 2004, he was with VIA-S3 Graphics, Shanghai, China, as a Graphics Processing Unit (GPU) Architecture Design Engineer. In 2002 and 2005, he conducted visiting research at Microsoft Research Asia, Beijing, China. Since October 2004, he has been with the Multimedia Group, Interuniversity Microelectronics Center, Leuven, Belgium, as a Ph.D. Researcher. Since September 2009, he has been working with the Advanced Digital Sciences Center, Singapore, which is a joint research center between the University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), Singapore, where he is leading a few research projects as a Senior Research Scientist. His research interests include computer vision, visual computing, image processing, video communication, interactive multimedia applications and systems, and efficient algorithms for various architectures. Dr. Lu was an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). He received the 2012 TCSVT Best Associate Editor Award.

Viet-Anh Nguyen (M'10) received the B.S and Ph.D. degrees in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2004 and 2010, respectively. He is currently working with the Advanced Digital Sciences Center (ADSC) in Singapore, which was jointly founded by University of Illinois at Urbana-Champaign (UIUC) and the Agency for Science, Technology and Research (A*STAR), a Singapore government agency. Before joining ADSC, he worked in the School of Electrical & Electronic Engineering, NTU as a Research Fellow from 2008 to 2010. His research interests include image and video processing, media compression and delivery, computer vision, and real-time multimedia system.

Minh N. Do (M'01-SM'07-F'14) was born in Vietnam in 1974. He received the B.Eng. degree in computer engineering from the University of Canberra, Australia, in 1997, and the Dr.Sci. degree in communication systems from the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland, in 2001. Since 2002, he has been on the faculty at the University of Illinois at Urbana-Champaign (UIUC), where he is currently a Professor in the Department of Electrical and Computer Engineering, and hold joint appointments with the Coordinated Science Laboratory, the Beckman Institute for Advanced Science and Technology, and the Department of Bioengineering. His research interests include signal processing, computational imaging, geometric vision, and data analytics. He received a Silver Medal from the 32nd International Mathematical Olympiad in 1991, a University Medal from the University of Canberra in 1997, a Doctorate Award from the EPFL in 2001, a CAREER Award from the National Science Foundation in 2003, and a Young Author Best Paper Award from IEEE in 2008. He was named a Beckman Fellow at the Center for Advanced Study, UIUC, in 2006, and received of a Xerox Award for Faculty Research from the College of Engineering, UIUC, in 2007. He was a member of the IEEE Signal Processing Theory and Methods Technical Committee, Image, Video, and Multidimensional Signal Processing Technical Committee, and an Associate Editor of the IEEE Transactions on Image Processing. He is a Fellow of the IEEE for contributions to image representation and computational imaging. He was a co-founder and CTO of Personify Inc., a spin-off from UIUC to commercialize depth-based visual communication.