LEARNING DEEP AND SPARSE FEATURE REPRESENTATION FOR FINE-GRAINED OBJECT RECOGNITION

M. Srinivas[1] *Yen-Yu Lin*[2] *Hong-Yuan Mark Liao*[1]

[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

ABSTRACT

In this paper, we address fine-grained classification which is quite challenging due to high intra-class variations and subtle inter-class variations. Most modern approaches to fine-grained recognition are established based on convolutional neural networks (CNN). Despite the effectiveness, these approaches still suffer from two major problems. First, they highly rely on large sets of training data, but manually annotating numerous training data is expensive. Second, the learned feature presentations by these approaches are often of high dimensions, leading to less efficiency. To tackle the two problems, we present an approach where on-line dictionary learning is integrated into CNN. The dictionaries can be incrementally learned by leveraging a vast amount of weakly labeled data on the Internet. With these dictionaries, all the training and testing data can be sparsely represented. Our approach is evaluated and compared with the state-of-the-art approaches on the benchmark dataset, CUB-200-2011. The promising results demonstrate its superiority in both efficiency and accuracy.

***Index Terms*—** Part-based RCNN, sparse representation, dictionary learning, fine-grained categorization

1. INTRODUCTION

Fine-grained classification (FGC) aims to distinguish fine-level categories of images, such as bird species, airplane or car types [1, 2], and animal breeds [3]. In addition to the difficulties inheriting from generic object recognition such as large intra-class variations, fine-grained classification is much more challenging due to subtle inter-class variations. In this work, we propose to learn a deep and sparse feature representation to address the difficulties of FGC, and illustrate it with the application to fine-grained bird species recognition. This application is considered challenging, since some of the species are difficult to recognize even for humans.

The major difficulties hinder the advances in accurate fine-grained classification come from diverse factors. First, there exist small inter-class variations and large intra-class variations. Second, the training data of a category in the benchmarks of FGC are often too few to reliably represent the data distribution of this category in the feature space. Meanwhile, the number of categories to be recognized is large. Third, the learned feature presentations by conventional approaches are often of high dimensions, leading to a high computational cost.

To address the first difficulty, two research trends of FGC arise. The first trend is the use of part-based representations [4]. Part-based models recognize objects by referring to not only the appearances of object parts but also their spatial relationships. Thus, these models are robust to intraclass variations caused by different poses. Meanwhile, the distinct characteristics for fine-grained recognition are often carried by object parts, instead of the whole objects. The second trend is to extract more discriminative feature representations [5] by using *convolutional neural networks* (CNN) [6]. Recent work based on CNN has shown notable improvement over the work that adopts handcrafted features [7, 8].

However, the two trends of FGC have worsened the other two difficulties, the demand for large training data and the high dimensions of the resultant feature presentations. Learning part-based models typically needs training data with part-level annotation, which leads to the expensive cost of manual labeling in collecting training data. The powerful CNN can extract discriminative features and learn non-linear classifiers simultaneously, but the resultant feature representations, i.e. the input to the last decision layer, are of high dimensions. Take *part-based RCNNs* (PRCNN) [10] as an example. It achieved promising results for FGC by using part annotations and CNN-based feature representations. However, further improvement based on PRCNN is very difficult, because learning part-based CNN requires a large set of strongly annotated training data, which is currently not available in the benchmark datasets for FGC.

The main contribution of this work lies in the development of a fine-grained classification approach that addresses these drawbacks of the part based models and reduces the computational cost caused by the high-dimension representation. We address the lack of training data by using additional web images, which can be obtained by querying the categories to be recognized in search engines. This additional dataset contains a large volume of images with image-level labels. Weakly supervised learning is conducted on the small

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

strongly-labeled data and the abundant weakly-labeled data. Specifically, we leverage the strongly labeled dataset to learn the part-based representation, and transfer the learned representation to the object parts in the weakly labeled dataset. In this manner, both the strongly and weakly labeled data can be used to derive a more reliable feature representation.

Using deep convolutional layers can extract features of high quality, but the dimension of the extracted features is quite large. Dimensionality reduction methods [9] can be used to reduce the dimension of feature representations. Specifically, *kernel principal component analysis* (KPCA) is employed. In this work, we propose a new method for fine-grained categorization that learns robust CNN-based feature representations, and carries out classification based on dictionary learning and sparse representation. With the aid of the extra data borrowed from the Internet, the problem caused by the lack of training data problem is alleviated. The use of dictionary learning further reduces the computational cost, and improves the accuracy of FGC.

Our approach increases the classification performance while reduces the computation cost. The main contribution of this work is three-fold. First, additional training data are used to prevent the problem of overfitting when training CNN with a small dataset available for FGC. Second, we learn category-specific dictionary using both strongly and weakly labeled training data. Thus, our approach is more efficient compared with the method in [11], since training data can be represented by just a few dictionary atoms. Another advantage of using on-line dictionary learning (ODL) is the reusability of existing dictionaries [12]. Namely the dictionary can be incrementally learned by using the extra weakly labeled data obtained from the Internet. The learned dictionary can be used to reduce the computational cost for processing web images. Third, we employ l_1-lasso sparsity to enhance the performance of predicting test data. Our method can efficiently search the sparsest representation of a test sample in the trained dictionary, which is composed of training samples of all classes. Thus, there is no need to derive the decision boundaries. These sparsely learned dictionaries give better classification performance. An overview of the our proposed method is shown in Fig. 1.

The experimental results show the effectiveness of the proposed approach. With the additional weakly labeled dataset acquired from the web, we achieve the recognition rate of 84.3% on the CUB-200-2011 dataset, which is comparable to those by the state-of-the-art approaches. Besides, our approach reduces the computational cost by 30% when comparing with the related work [11].

2. RELATED WORK

Significant progress has been made on fine-grained classification (FGC) in the field of computer vision. Recent methods for FGC using CNNs have achieved performance gains over

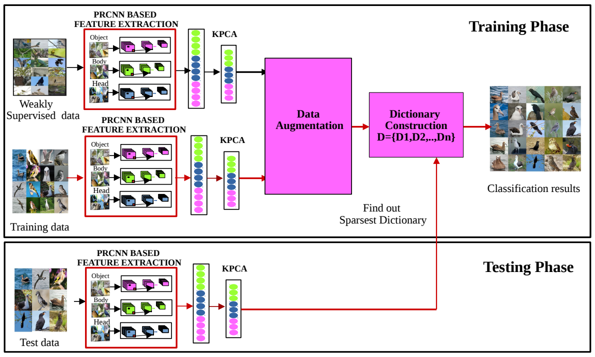

Fig. 1. Overview of our approach.

the methods that adopt handcrafted features [7, 8, 13]. The major disadvantage of part-based methods is the high cost of manually annotating object parts in collecting training data. Branson et al. [14] described an architecture for bird species fine-grained classification. In this work, features are computed by applying deep CNN to image patches. These image patches are located and normalized by poses. For learning a compact space of normalized poses, higher order functions for geometric warping and graph-based clustering algorithms are included in their work. However, this approach does not give satisfactory classification accuracy in predicting bird species.

For fine-grained recognition, many current methods employ a two-stage pipeline, part detection followed by classification, to accomplish this challenging task. Zhang et al. [10] learned a more robust model for fine-grained classification without using the bounding boxes of objects at the phase of testing. They method adopts the RCNN detector [15] to learn the whole object, locate the parts and make the prediction. It employs the constraints to enforce the learned geometric relationship between the detected parts and the whole objects. The learned pose-normalized representation is used to carry out fine-grained categorization. Huang et al. [16] proposed the part-stacked CNN for fine-grained classification. They leveraged a two-stream structure to capture both object-level and part-level information. Krause et al. [17] presented an approach that carries out fine-grained recognition without using part-level annotation. In their approach, part-level information is derived by using co-segmentation and image alignment. Compared to [10, 16, 17], our proposed method has the advantages in both classification accuracy and computational efficiency.

In [18], Lin et al. proposed a bilinear model for representing bird species, and applied it to fine-grained classification. They employed two CNN models, and generated the feature representations by computing the outer product of feature maps. Namely, every two feature maps are multiplied at each location and pooled across locations to obtain an image descriptor. In [19], Shih et al. presented a network layer, called the co-occurrence layer, which explores the co-occurrence of visual patterns to enhance the performance of fine-grained recognition. The work in [11] compiles more discriminative CNN features. It uses a training set augmented by additional part patches obtained from weakly labeled web images, and generates more robust feature representations. A multi-instance learning algorithm is applied to both the strongly and weakly labeled datasets to learn a more robust classifier. The disadvantage of the method in [11] is that the dimension of the feature representation is large. It results in the high computational cost of processing the two CNN models. Our proposed method instead balances classification accuracy and computation cost in fine-grained recognition.

3. THE PROPOSED METHOD

In this section, we present our method for bird species fine-grained categorization. Our method learns a CNN based feature representation and employs detailed object part annotation in a unified framework. With the aid of weakly labeled web data, it alleviates the problem caused by the lack of training data. Besides, it integrates dictionary learning into CNN. Classification based on sparsity is carried out for bird species prediction.

3.1. Feature Representation

Recently feature representations based on deep CNN are widely used for fine-grained categorization. Specifically, PRCNN has demonstrated the effectiveness of such feature representations in this scenario. The main limitation of PRCNN is that learning such feature representations is almost infeasible with a small set of training data. In particular, CNN requires a large set of training data to learn robust feature representations. By using weakly labeled web images from Flicker, our approach overcomes the lack of training data.

Suppose that we are given a training set with part-level annotation. The set contains images of R fine-grained categories, and each image is annotated with a bounding box x_0 of the whole object and the bounding boxes $\{x_1, x_2, ..., x_m\}$ of m different parts. We use the training set, and learn a detector for the whole object and a detector for each part. Then, the part detectors are used to collect additional part patches from the weakly labeled images by localizing object parts during testing. Specifically, we follow the augmented PRCNN to learn the detectors where the part-based CNNs with k-way fc8 classification layer are fine tuned.

Let $\{y_0, y_1, ..., y_m\}$ denote the weights of R-CNN detectors for the whole object x_0 and the m different parts $\{x_i\}_{i=1}^{m}$. The corresponding detector scores $\{e_0, e_1, ..., e_m\}$ are computed for every region proposal b via

$$e_i(b) = \sigma(y_i^T \phi^{(i)}(b)), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $\phi^{(i)}(b)$ represents the features extracted by the ith part detector at location b.

Although weakly labeled images contain only image-level annotation, we use the learned part detectors to discover the part patches in these weakly labeled images. After applying the part detectors to all image locations, the detection scores constrained by geometric relations are used to infer the locations of the object and its parts. The detected locations $H^* = \{h_0, .., h_m\}$ are given by maximizing the following function:

$$H^* = \arg\min_{H} \prod_{i=1}^{m} g_{h_0}(h_i) \prod_{i=0}^{m} e_i(h_i), \tag{2}$$

where

$$g_{h_0}(h_i) = \begin{cases} 1, & \text{if } h_i \text{ falls outside } h_0 \text{ by at most 10 pixels,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In this way, the part-based information of the weakly labeled images is revealed. We then use the newly obtained information to fine-tune the CNN-based part detectors. On the one hand, the part detectors can be more reliably trained. On the other hand, the part locations can be more precisely determined owing to the better part detectors.

However, the dimension of the feature vector ϕ is often too high to build an effective dictionary. In order to reduce the dimension, kernel principle component analysis is applied. In the following, we will discuss the dictionary learning and sparse representation for classification.

3.2. Dictionary Learning and Sparse Representation

Sparsity-dominated dictionaries give us an effective representation for fine-grained classification. To process training data, several dictionary learning methods have been developed such as K-SVD [20], on-line dictionary learning (ODL) [12] and incremental dictionary learning (IDL) [21]. Earlier work on dictionary learning and sparsity representation based classification is typically used for high-level image classification problems. However, these methods are not particularly useful for fine-grained classification. In the following, we explain how to construct category-specific dictionaries from training data to resolve this issue.

In fine-grained datasets, the differences between categories are very subtle. In these datasets, similar parts across categories may degrade the performance of the learned dictionary. When we are applying sparse coding techniques, the learned shared dictionary is possible to be influenced by the matching parts of images. Most of the dictionary atoms encode common features, while only a few of atoms encode the differences. Thus, the shared dictionary degrades the performance of fine-grained classification. By using category specific dictionaries, most of the dictionary atoms are helpful in encoding the differences between data of different classes.

In the proposed method, ODL is used to train a dictionary for each category. Namely, for the training data of R

categories, we construct R dictionaries to represent the categories. Each image is associated with the dictionary that gives the sparsest representation. For a given test sample, it can be represented as a linear combination of training data from category specific dictionary atoms. Specifically, given a database of N training images of R classes, training samples $\{\mathbf{y}_i\}_{i=1}^{N}$ collected from both the strongly and weakly labeled data are denoted by $C = [C_1, \ldots, C_d, \ldots, C_R]$, where C_d is the data matrix of class d. Let L_d denote the number of training data of class d. Let a be an image belonging to the dth class. Then it can be represented as a linear combination of these training samples:

$$a = D_d \Phi_d, \tag{4}$$

where D_d, a matrix of size $m \times L_d$, is a dictionary whose columns are (atoms) the training samples in the dth class and Φ_d represents the dth class related sparse coefficients. The proposed method is a two-step process: the first step is dictionary learning and the second one is sparse representation based classification. These steps are detailed below:

3.2.1. Dictionary Construction:

The ODL algorithm is used to construct the dictionary for each class of training samples. The sparse stage in ODL is a Cholesky-based implementation of LARS-lasso algorithm. Thus, the dictionaries $D = [D_1, \ldots, D_R]$ are computed by using the following equation:

$$(D_i, \Phi_i) = \arg \min_{D_i, \Phi_i} \frac{1}{2} \| C_i - D_i \Phi_i \|_2^2 + \lambda \| \Phi_i \|_1, \quad \text{for } i = 1, 2, \ldots, R. \tag{5}$$

3.2.2. Sparsity based Classification:

In the classification phase, we seek the sparse vector Φ for a given test image g. Using the dictionaries of training samples $D = [D_1 \ldots D_N]$, the sparse representation Φ is obtained by solving the following optimization problem:

$$\Phi = \arg \min_{\Phi} \frac{1}{2} \| g - D\Phi \|_2^2, \quad \text{subject to} \| \Phi \|_1 \leq T_1, \tag{6}$$

where T_1 represents the sparsity threshold. Then, the prediction is made based on

$$\hat{i} = \arg \max_{i} \| \delta_i(\Phi) \|_1, \text{ for } i = 1, 2, \cdots, R, \tag{7}$$

where δ_i is a characteristic function that selects the coefficients for class i. Define $\delta_i(\Phi) = (\Phi_{i,1}, .., \Phi_{i,R_i})$ which is the contribution of i^{th} class to the represent of g in the dictionary. This test clip g is assigned to class $\hat{i}$ if the absolute sum of sparsity coefficients associated with the $\hat{i}$th dictionary is the maximum among other classes. This method chooses l_1 norm as it is found to be better for the classification results. According to our implementation, sparsely learned dictionaries give better classification performance.

Table 1. Accuracy comparison between APRCNN [11] and our method on the CUB200-2011 dataset. A = APRCNN, P = our method, ft = fine-tuning, and dn = denoising.

Part Localization	Predict BBox			
Method	Train		Train+Weak	
	A	P	A	P
ft on Train	78.6	82.1	79.9	83.4
ft on Train/Weak	81.2	83.3	82.2	84.3
ft on Train/Weak dn	83.2	83.6	84.6	84.3

4. EXPERIMENTAL RESULTS

In this section, we describe the used dataset and explain the experimental results with different dictionary sizes. Experiments are carried out on CUB-200-2011 birds database [22], which contains 11,788 images of 200 bird classes. About 30 images of each class are used for training while about 30 of each class are used for testing. Each image is associated with an image-level label, object bounding boxes and part landmarks. To solve the problem of lacking such strongly labeled data in fine-grained classification, we collect an additional weakly labeled dataset from Flicker. We augment the strongly labeled dataset with the weakly labeled one. This additional dataset contains 100 images for each category to be recognized. These additional web images are with only image-level labels, which may not be correct due to the ambiguity of query words and label noise.

We implement this work on the deep learning package `Caffe` and use part-based RCNN for robust feature representations. The performance of our approach is evaluated with various settings, such as with or without part-level annotations in the training and/or testing phases and with or without using the weakly labeled data to fine-tune the networks. Specifically, we follow the work in [11] and use the same setting for the ease of comparison in the experiments. Table 1 shows the performance of the proposed method and the related paper [11] with these different settings. When fine-tuning the part CNN on the training set, the proposed method gives the recognition rate of 82.1%. By augmenting the part patches from weakly labeled dataset and fine-tuning the part CNNs, the accuracy is improved to 83.3%. The results show the advantages of using the additional dataset to better train the CNN. A further improvement in the classification accuracy can be obtained by denoising the weakly labeled dataset. The resultant accuracy is 83.6%. In these two cases the proposed method gives promising results compared with APRCNN [11]. Finally, our proposed method achieves the recognition rate of 84.3% with the aid of the weakly labeled data.

Table 2 reports the classification accuracy of the proposed method and the state-of-the-art methods on CUB-200-2011 dataset. On this challenging dataset, we achieve a promising recognition rate of 84.3%. In addition, our approach inte-

Table 2. Performance of different approaches on the CUB200-2011 dataset. Part = part-level annotation, BBox = bounding box, ACC = accuacy rate (%).

Method	Train BBox	Train Part	Test BBox	ACC
DPD+DeAF [23]	✓	✓	✓	65.0
POOF [24]	✓	✓	✓	56.8
Symbiotic [25]	✓		✓	61.0
Alignment [26]	✓		✓	62.7
CNNaug [27]	✓		✓	61.8
PRCNN [10]	✓	✓		73.9
PoseNorm CNN [14]	✓	✓		75.7
Co-segmentation [17]	✓			82.0
Bilinear Model [18]				84.1
Co-occur. Layer [19]				85.8
APRCNN [11]	✓	✓		84.6
Proposed Method	✓	✓		84.3

grates online dictionary learning for producing a sparser feature representation. Thus, another advantage of our approach is that it is more computationally efficient when comparing with most CNN-based approaches.

It is worth mentioning that our approach induces less computational cost. The dimensions of CNN-based feature representations are often very high. For instance, the dimension is 12288 in RCNN. Working on high-dimensional data takes more computational time. In this work, we use kernel PCA to reduce the dimension to 1000. We evaluate the performance of the proposed method with three different types of kernels, including the sigmoid kernel, the RBF kernel, and the polynomial kernel. Results are shown in Fig. 2(b).

We conduct the experiments with different dictionary sizes. The results are shown in Fig. 2(a). In the experiments, the CUB-200-2011 dataset consists of 200 bird classes, each of which has about 30 images for training. As can be observed in Fig. 2(a), 80 atoms, i.e. $D = 80$, suffices to achieve satisfactory results. It means that the coefficient vector of each image is of dimension 80, which is much lower than those in CNN and those after applying KPCA. Thus, the main advantage of our approach over the existing CNN-based methods is that data can be represented in a more compact way, but the representation still achieves the state-of-the-art classification accuracy. Besides, online dictionary learning is particularly suitable in our cases where weakly labeled images are obtained sequentially.

We evaluate the performance of the proposed method with three different feature dimensions when applying kernel PCA, i.e. $V_d = 600$, 1000 and 2000. The results are shown in Fig. 2(c). When the dimension is set to 1000, the proposed method gives better classification accuracy. Fig. 2(d) shows the results of computational cost when comparing the proposed method with the related work APRCNN [11]. Our approach is executed with a lower computational cost. It results from the use of KPCA and dictionary learning for sparse data representation. Therefore, our method reduces the computation cost compared with other techniques. In addition, sparsely learned dictionaries give better classification performance results in the experiments.

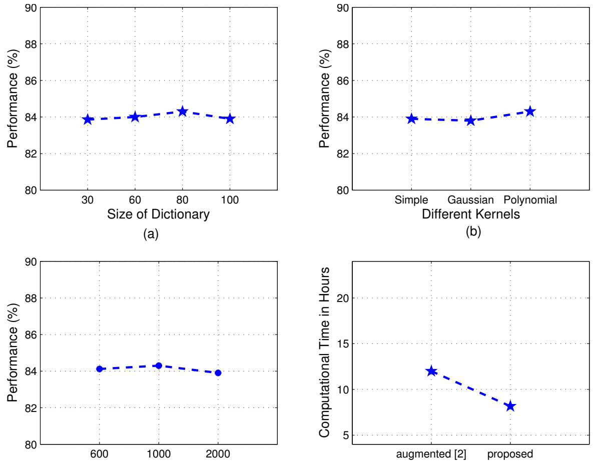

Fig. 2. Performance comparison be with (a) different dictionary sizes ($D = 10, 20, 30, 60$ and 80), (b) different Kernel functions, and (c) different feature dimensions ($V_d = 600$, 1000 and 2000). (d) Comparing our approach with APRCNN [11] in running time.

5. CONCLUSIONS

In this paper, we present an approach to compile a deep and sparse feature representation, and illustrate it with the application to fine-grained classification. The approach integrates dictionary learning and sparsity classification into the CNN framework. We also alleviate the problem caused by the lack of fully labeled training data via using additional weakly labeled dataset collected from Flicker. The experimental results show that augmenting the additional part patches from weakly labeled dataset to the strongly labeled training data helps fine-tune the CNN models as well as achieve better classification results. Using dictionary learning also helps alleviate the issue of the high computational cost required for fine-grained categorization. It follows that a given training or testing data is represented by a compact vector, whose dimension is equivalent to the number of dictionary atoms. In our case, the dimension is reduced from 12288, the number of the features in the last layer of CNN, to 80, the dictionary size. More importantly, such a low dimensional representation still achieves the recognition rate of 84.3% on the CUB-200-2011 database, which is comparable to the state-of-the-art accuracy.

Acknowledgement. The authors would like to thank the anonymous reviewers for their thorough review and constructive comments. This work was supported in part by grants MOST 105-2221-E-001-030-MY2, MOST 105-2218-E-001-006, and MOST 105-2221-E-001-018-MY3.

6. REFERENCES

[1] M. Stark, J. Krause, B. Pepik, D. Meger, J. J. Little, B. Schiele, and D. Koller, "Fine-grained categorization for 3d scene understanding," in British Machine Vision Conference, 2012.

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

[3] J. Liu, A. Kanazawa, D. W. Jacobs, and P. N. Belhumeur, "Dog breed classification using part localization," in Euro. Conf. on Computer Vision, 2012.

[4] E. Rosch, C. B. Mervisa, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," Cognitive Psychology, 1976.

[5] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in Neural Information Processing Systems, 2010.

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

[7] L. D. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in Int. Conf. on Computer Vision, 2011.

[8] 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 Int. Conf. on Computer Vision, 2011.

[9] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," IEEE Tran. on Pattern Analysis and Machine Intelligence, 2011.

[10] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell "Part-based R-CNNs for fine-grained category detection," in Euro. Conf. on Computer Vision, 2014.

[11] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Augmenting strong supervision using web data for fine-grained categorization," in Int. Conf. on Computer Vision, 2015.

[12] J. Mairal, F. R. Bach, J. Ponce, and G. Sapiro, "On-line dictionary learning for sparse coding," in Int. Conf. on Machine Learning, 2009.

[13] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for sub-category recognition," in Computer Vision and Pattern Recognition, 2012.

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

[15] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE Tran. on Pattern Analysis and Machine Intelligence, 2016.

[16] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," arXiv preprint arXiv:1512.08086, 2015.

[17] J. Krause, H. Jin, J. Yang, and F.-F. Li, "Fine-grained recognition without part annotations," in Computer Vision and Pattern Recognition, 2015.

[18] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in Int. Conf. on Computer Vision, 2015.

[19] Y.-F. Shih, Y.-M. Yeh, Y.-Y. Lin, M.-F. Weng, Y.-C. Lu, and Y.-Y. Chuang, "Deep co-occurrence feature learning for visual object recognition," in Computer Vision and Pattern Recognition, 2017.

[20] M. Aharon, M. Elad, and A. Bruckstein "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Tran. on Signal Processing, 2006.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in Computer Vision and Pattern Recognition, 2010.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Technical Report CNS-TR-2011-001, CalTech, 2011.

[23] 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 Int. Conf. on Machine Learning, 2014.

[24] T. Berg and P. N. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in Computer Vision and Pattern Recognition, 2013.

[25] Y. Chai, V. S. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in Int. Conf. on Computer Vision, 2013.

[26] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in Int. Conf. on Computer Vision, 2013.

[27] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in Computer Vision and Pattern Recognition Workshops, 2014.