DEEP DICTIONARY LEARNING FOR FINE-GRAINED IMAGE CLASSIFICATION

M. Srinivas   *Yen-Yu Lin*   *Hong-Yuan Mark Liao*

Academia Sinica, Taipei, Taiwan

ABSTRACT

Fine-grained image classification is quite challenging due to high inter-class similarity and large intra-class variations. Another issue is the small amount of training images with a large number of classes to be identified. To address the challenges, we propose a model for fine-grained image classification with its application to bird species recognition. Based on the features extracted by *bilinear convolutional neural network* (BCNN), we propose an *on-line dictionary learning* algorithm where the principle of sparsity is integrated into classification. The features extracted by BCNN encode pairwise neuron interaction in a translation-invariant manner. This property is valuable to fine-grained classification. The proposed algorithm for dictionary learning further carries out sparsity based classification, where training data can be represented with a less number of dictionary atoms. It alleviates the problems caused by insufficient training data, and makes classification much more efficient. Our approach is evaluated and compared with the state-of-the-art approaches on the CUB-200-2011 dataset. The promising experimental results demonstrate its efficacy and superiority.

Index Terms— Sparse representation, on-line dictionary learning, deep learning, fine-grained image classification

1. INTRODUCTION

Fine-grained image classification targets at distinguishing fine-level image categories in images, such as bird species, airplane types, and animal breeds. In addition to the difficulties inheriting from generic image classification such as large intra-class variations and insufficient training data, fine-grained classification is much more challenging due to subtle inter-class differences. For instance, the inter-class differences between a glaucous-winged gull and a Larus mainly lie in the patterns in their beaks, which are significantly more subtle than the intra-class variations on a popular fine-grained dataset for birds [1]. In this work, we investigate the task of fine-grained bird species recognition, which is considered quite challenging since some of the species are difficult to recognize, even for humans.

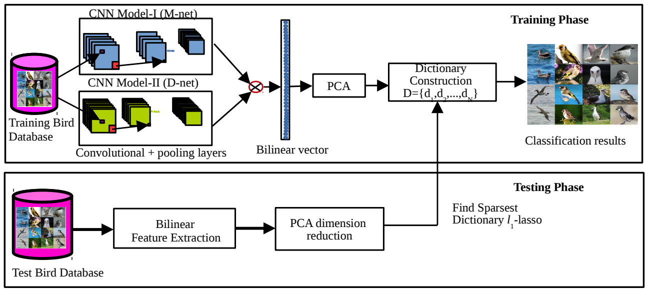

Fig. 1. The overview of our approach including both the training and the testing phases.

The major difficulties hindering the advances in accurate fine-grained image classification result from diverse factors. First, modern benchmarks datasets for fine-grained classification consist of a small amount of training data per category. Second, there exist large intra-class variations and small inter-class variations. Third, the number of fine-grained categories to be recognized is often large in increasingly complex applications. Conventional approaches to generic image classification suffer from these difficulties, and cannot be applied to fine-grained image classification without modifications.

Two research focuses are widely adopted in existing approaches to fine-grained classification. The first one is the use of the part-based representation as stated in Rosch *et al.* [2]. Part-based models recognize images by considering not only the appearances of individual parts but also their geometric layout. They are more robust to unfavorable intra-class variations. However, part-based algorithms rely on the extra efforts on collecting part-level annotation, such as adopting human-in-the-loop methods [3, 4], using data with part-level annotations [5, 6, 7], or carrying out unsupervised patch discovery [8]. The second focus is to explore more discriminative feature representations [9]. Recent studies have shown that this task can be accomplished by using *convolutional neural networks* (CNNs) [10, 11]. By employing deep CNN features extracted from pre-trained models on large datasets and domain-specific fine tuning approaches, significant improvements in a image classification and detection tasks can be achieved [12]. Modern approaches using CNNs has shown notable improvement over the conventional approaches that adopt handcrafted features [13, 14].

However, the two aforementioned research focuses lead to two unfavorable effects, i.e. the demand for part-level annotation and the high dimensions of the resultant feature presentations. Part-level annotation results in an expensive cost

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

of manual labeling in training data collection. CNNs extract discriminative features and learn non-linear classifiers simultaneously, but the resultant feature representations, i.e. the input to the final decision layer, are typical of high dimensions. The main contribution of this paper lies in the development of a fine-grained classification approach that addresses these drawbacks.

In this work, we use the features extracted by *bilinear CNN* (BCNN) [1], which computes the outer products between feature maps for image description. The pairwise correlations between the feature maps can be considered the interaction between object parts since the corresponding neurons of these feature maps are often sensitive particular patterns (or parts) in the image. In this way, part-level annotation is not required. But the output dimension of BCNN is quite large. To reduce the dimension of outputs, we will adopt dimensionality reduction methods with different data supervision conditions [15, 16] to make data more compact.

We learn a category-specific dictionary with *on-line dictionary learning* (ODL) for fine-grained categorization. The main advantage of training a dictionary [17] is that the training data are sparsely represented. Such a representation facilitate various follow-up applications [18]. Another advantage is that it gives more robust classification performance even with a small dictionary size and a few training data. For data in the sparse representation, we employ l_1-lasso sparsity-based classification method. It efficiently searches 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. Sparsely learned dictionaries give better classification performance results in our experiments. To sum up, the overview of the proposed method is shown in Fig. 1.

2. RELATED WORK

Fine-grained image classification has gained significant progress in the fields of image processing and computer vision within a short period of time. Recent CNNs-based models [19, 7, 20] have demonstrated significant performance improvement over the models using handcrafted features [13, 14, 21]. Models employing part-based CNNs achieve significant performance gains for fine-grained recognition. For instance, the method in [20] learns the part detectors and the whole-object detectors. It also applies the constraints to enforce the learned geometric relationship between the detected parts and objects. It follows that the resulting pose-normalized representation is used to carry out fine-grained categorization. Their work does not require object bounding boxes at the testing phase. However, the main disadvantage of models of this category is that the part-based annotation is the high cost of manually annotating object parts in collecting training data. Branson *et al.* [22] proposed a system for bird species classification by mingling human interaction with the system. In their work, a non-localized computer vision method is used to extract the bag-of-words features from the entire image. The work in [23] is composed of a human user and a machine. The machine interactively provides two heterogeneous forms of information, i.e. clicking on object parts and answering binary questions arisen by a user who is not able to carry out the recognition task. In general, these human-in-the-loop methods are not practical. In [24], Shih *et al.* proposed a deep co-occurrence feature learning method for visual object recognition.

Encoding the image content into a highly-discriminative visual signature by using both segmentation and part localization techniques is proposed in [8]. The model is symbiotic in the sense that part localization helps segmentation and conversely segmentation facilitates part localization. Specifically in [8], part localization is accomplished by using the part-based detector in [8], while the segmentation results are obtained by using *GrabCut*. Despite effectiveness, extra work for part localization and segmentation is required. In [19], an architecture for bird species fine-grained classification was presented. it computes features by applying deep CNNs to image patches. These image patches are located and normalized by poses. For learning a compact pose normalization space, higher order geometric warping functions and a graph-based clustering algorithm are included in the work. We instead use the BCNN model [1] for image description. The major difference is that the reduced dimension bilinear features are used in our work to learn category specific dictionaries. Our approach can derive a better representation model and a sparse-based classification technique, which are jointly applied to a test sample to get better performance.

The method in [17] employs category specific and shared dictionaries for fine-grained classification. Our work is developed with three major differences from [17]. First, HOG features are used in [17]. It has been pointed out in the literature that handcrafted features typically lead to sub-optimal performance. We instead integrate the deep model for feature extraction into our work. Second, the category-specific and the shard dictionaries for feature encoding are derived by the K-SVD algorithm. However, shared dictionaries often degrade the classification accuracy in fine-grained classification, like bird species identification. Another drawback of the K-SVD algorithm is that it is much slower than ODL, especially when the batch mode is used for dictionary update. Third, an SVM classifier is used for classification in [17]. In [25], we address the fine-grained classification problem by using on-line dictionary learning with its integration into CNN. In this work weakly supervised data is used to solve the less availability of data problem. Our work uses *sparsity representation based classification* (SRC), which searches for the sparsest representation of a test data in a dictionary composed of all training data of all classes, and achieves better performance.

3. OUR APPROACH

In this section, we first describe the feature extraction process by using bilinear CNN. Then, we present the key components of the proposed approach, including on-line dictionary construction and sparse representation based classification.

3.1. Bilinear CNN for feature extraction

In the following, we describe bilinear CNN (BCNN) for feature extraction, which is composed of two CNNs and their combination via computing the outer products between feature maps.

BCNN can serve as an image descriptor that compiles bilinear feature representation from two CNNs whose outputs are multiplied using the outer product pooled across locations. Specifically, a bilinear model B consists of two features functions $B = (f_{V1}, f_{V2})$, where f_{V1} and f_{V2} are derived from CNN model-I and model-II, respectively. A feature function is defined as $f_{Vi} : M \times I \rightarrow R^{1 \times L_i}$, where it takes an image I and a location $m \in M$ as inputs, and outputs features of size L_i, for $i \in \{1, 2\}$. A matrix outer product is then used at each location to combine the feature outputs. At any given location m, the bilinear feature combination of f_{V1} and f_{V2} is given by the bilinear function $f(m, I, f_{V1}, f_{V2})$=$f_{V1}(M, I)^T f_{V2}(M, I)$. If the extracted features by f_{V1} and f_{V2} are of size L_1 and L_2 respectively, the bilinear features, i.e. $\phi(I) = \sum_{m \in M} f(m, I, f_{V1}, f_{V2})$, are of size $L_1 \times L_2$. The bilinear feature matrix is then reshaped to obtain the bilinear vector of size $L_1 L_2 \times 1$.

Following [1], the sum pooling technique is used to aggregate the bilinear features across the image. If the bilinear vector obtained above is denoted by $x = \phi(I)$, we can obtain a signed square-root representation ($y \leftarrow \text{sign}(x)\sqrt{|x|}$) followed by l_2 normalization ($z \leftarrow y/||y||_2$) to obtain the final bilinear representation. However, the dimension of the bilinear feature vector z is too large to build an effective dictionary. In order to reduce the dimensionality, PCA is applied. In the next section, we discuss the dictionary learning and sparse representation based 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 [18], *on-line dictionary learning* (OLD) [26] and *incremental dictionary learning* (IDL) [27]. These approaches to dictionary learning and sparsity based classification are typically used for high-level, generic image classification. However, these methods are not particularly suitable for fine-grained classification problems where category-specific features are required to separate data of a class from the rest. Hence, we describe how to construct the category-specific dictionaries to enhance fine-grained classification.

The differences between classes are very subtle in fine-grained classification. The shared dictionary is probably dominated by atoms that capture patterns commonly shared across classes. Only a few of atoms encode the discriminative differences. Thus, the shared dictionary degrades the performance of fine-grained classification. By using category specific dictionaries, most of dictionary atoms will become helpful in encoding the differences between data of different classes.

In the proposed method, on-line dictionary learning is used to train the dictionaries from training data. Specifically for a given number of categories R, we learn R dictionaries, one for each category. Based on a learned dictionary, an image can be compactly described by the coefficients that give the most sparsest representation. Consider a database of N training images and of R classes. Training samples $\{z_i\}_{i=1}^{N}$ in form of bilinear features are denoted by $C = [C_1 ... C_r ... C_R]$, where C_r is the data matrix of class r. Let I be an image belonging to class r. Then it can be approximated by a linear combination of the dictionary atoms of that class, i.e.

$$I \simeq D_r \psi_r, \tag{1}$$

where D_r is the learned dictionary of class r, whose columns are the atoms. ψ_r is the resultant coefficient vector. The proposed method has a two-step process. The first step is to construct category-specific dictionaries. The second one is to yield the sparse representation for classification. The two steps are detailed below:

1) Dictionary Construction: The ODL algorithm is used to construct a dictionary for training samples of each class. The dictionaries $D = [D_1, \dots, D_R]$ are constructed by solving the following optimization problem:

$$(D_r, \Psi_r) = \arg\min_{D_r, \Psi_r} \frac{1}{2} \| C_r - D_r \Psi_r \|_2^2 + \lambda \| \Psi_r \|_1, \quad \text{for } r = 1, 2, \dots, R, \tag{2}$$

where λ is a positive constant and Ψ_r is the matrix whose columns are the coefficient vectors of data of class r.

2) Sparsity based Classification: Given a test sample z, its coefficient vector ψ based on all the category-specific dictionaries, $D = [D_1 \dots D_R]$, is firstly computed via solving the following optimization problem:

$$\psi = \arg\min_{\psi} \frac{1}{2} \| z - D\psi \|_2^2, \quad \text{subject to} \| \psi \|_1 \leq T, \tag{3}$$

where T is the sparsity threshold. Then, z is predicted as the class with the least reconstruction error, i.e.

$$i = \arg\min_{i} \| z - D\delta_i(\psi) \|_2^2, \tag{4}$$

where δ_i is a characteristic function that selects only the coefficients of class i. Namely for a given test sample, we find its sparsest representation based on all the category-specific dictionaries by using l_1-lasso algorithm. Then, the test sample is assigned to the class with the least reconstruction error. In the experiments, we will show that the sparsity based classification improves the recognition performance.

Table 1. Performance of different approaches on CUB-200-2011 dataset. [BB=bounding box]

method	BB	ACC (%)
DeCAF6 [11]	✓	58.8
Symbiotic [8]	✓	61.0
CNNaug [12]	✓	61.8
Alignment [29]	✓	67.0
Part R-CNN [7]		73.9
PoseNorm CNN [19]		75.7
Xu *et al.* [20]		78.6
BCNN[D,M] [1]		84.1
Ours		**84.6**

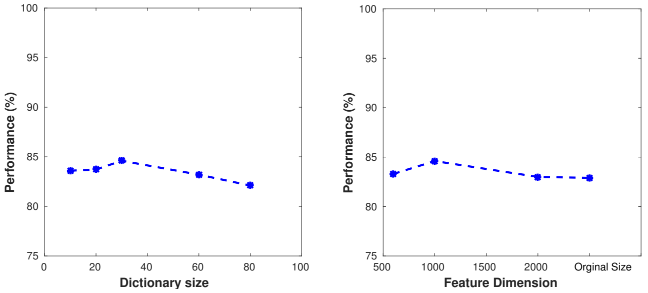

Fig. 2. Performance with different (a) dictionary sizes and (b) feature dimensions (V_d = 600, 1000, 2000, and the original dimensions).

4. EXPERIMENTAL RESULTS

In this section, we evaluate our approach on the CUB-200-2011 birds dataset [28], which contains 11,788 images of 200 bird classes. A central square patch is cropped from each image in this dataset. The patch is then resized to resolution 448×448. Different results have been reported with various experimental settings on this dataset, such as the availability of the bounding boxes and the part-level annotation at training and/or test phases. We follow the setup in [1] where neither part-level annotation nor bounding boxes are available.

In our experiments, BCNN [1] is used for feature extraction. It is composed of two different CNN models, i.e. the very deep network *D-net* [30] and *M-net* [31]. In both models, we consider the feature vector as the outputs of the last convolution layer with non-linearity activation, i.e. layer $\texttt{conv}_{5,4}\texttt{+Relu}$ for D-net and layer $\texttt{conv}_{5}\texttt{+Relu}$ for M-net. Both models produce 1×512-dimensional features at each location. The D-net produces slightly larger output features compared with M-net. Like [1], we downsample the output of the D-net by removing one row and one column. The output spatial sizes of D-net and M-net are then equal, i.e. 27×27. The pooled bilinear features, the pairwise inner product between feature maps, are of size 512×512, or 1×262144 equivalently. We employ PCA to reduce its dimension to 1×1000, which serves as the input to our approach to category-specific dictionary learning and sparsity-based classification.

Our approach is compared with the powerful approaches on CUB-200-2011 dataset. Their performance is reported in Table 1. In competing approaches [1, 7, 19, 20] and our approach, bounding boxes are not used. The approaches [7, 19] give the recognition rates of 73.9% and 75.5%, respectively. Both approaches employ part-based detectors to capture part-level information for improving performance, but lead to extra computational costs. In [1, 20], their approaches apply SVM to the extracted deep features for classification, and achieve the better performance of 84.1% and 78.6% respectively. Our approach instead learns category-specific dictionary to grab the discriminative information, and gives a promising accuracy rate of 84.6%. It is superior to the competing approaches, and achieves the state-of-the-art performance.

We evaluate the effects of the dictionary size and the reduced dimension on our approach. The performance of our approach with various dictionary sizes is shown in Fig. 2(a). The results indicate that dictionaries of small sizes suffice to achieve satisfactory results. We also evaluate the proposed method with four different feature dimensions, including $V_d = 600$, 1000, 2000 and 262144 (the original dimensions). The results in Fig. 2(b) show that the proposed method gives better classification accuracy when the dimension is set to 1000, even better than with the original dimensions. The results demonstrate that our approach works well with low-dimensional features and small dictionaries. Thus in addition to accuracy, our approach has the advantage in a set of follow-up applications such as retrieval or visualization where a compact representation of data is appreciated.

5. CONCLUSIONS

In this paper, we have presented an effective and efficient approach to fine-grained image classification. Our approach adopts a bilinear feature representation, and carries out category-specific dictionary learning and sparsity-based classification. The category-specific dictionaries can capture the discriminative features for fine-grained classification, and alleviate the problems of high computational costs since data are sparsely represented. The promising experimental results demonstrate that our approach can tackle the issues of insufficient training samples and the large number of categories, and achieve the state-of-the-art performance. For further work, we plan to leverage the flexibility of dictionary learning, and explore weakly labeled or even unlabeled training data to further enhance fine-grained image classification.

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] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, “Bilinear CNN models for fine-grained visual recognition,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015.

[2] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem, “Basic objects in natural categories,” *Cognitive psychology*, 1976.

[3] Jia Deng, Jonathan Krause, and Li Fei-Fei, “Fine-grained crowdsourcing for fine-grained recognition,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[4] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman, “Discovering localized attributes for fine-grained recognition,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[5] Thomas Berg and Peter Belhumeur, “Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[6] Zhenyang Li, Efstratios Gavves, Thomas Mensink, and Cees GM Snoek, “Attributes make sense on segmented objects,” in *Proc. Euro. Conf. Computer Vision*, 2014.

[7] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, “Part-based R-CNNs for fine-grained category detection,” in *Proc. Euro. Conf. Computer Vision*, 2014.

[8] Yuning Chai, Victor Lempitsky, and Andrew Zisserman, “Symbiotic segmentation and part localization for fine-grained categorization,” in *Proc. Int’l Conf. Computer Vision*, 2013.

[9] Liefeng Bo, Xiaofeng Ren, and Dieter Fox, “Kernel descriptors for visual recognition,” in *Proc. Neural Information Processing Systems*, 2010.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Neural Information Processing Systems*, 2012.

[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, “DeCAF: A deep convolutional activation feature for generic visual recognition,” in *Proc. Int’l Conf. Machine Learning*, 2014.

[12] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in *Computer Vision and Pattern Recognition Workshops*, 2014.

[13] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, “Describing people: A poselet-based approach to attribute classification,” in *Proc. Int’l Conf. Computer Vision*, 2011.

[14] Ryan Farrell, Om Oza, Ning Zhang, Vlad I Morariu, Trevor Darrell, and Larry S Davis, “Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance,” in *Proc. Int’l Conf. Computer Vision*, 2011.

[15] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh, “Multiple kernel learning for dimensionality reduction,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2011.

[16] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh, “Dimensionality reduction for data in multiple feature representations,” in *Proc. Neural Information Processing Systems*, 2008.

[17] Shenghua Gao, Ivor Wai-Hung Tsang, and Yi Ma, “Learning category-specific dictionary and shared dictionary for fine-grained image categorization,” *IEEE Trans. on Image Processing*, 2014.

[18] Michal Aharon, Michael Elad, and Alfred Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. on Signal Processing*, 2006.

[19] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona, “Bird species categorization using pose normalized deep convolutional nets,” *arXiv preprint:1406.2952*, 2014.

[20] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao, “Augmenting strong supervision using web data for fine-grained categorization,” in *Proc. Int’l Conf. Computer Vision*, 2015.

[21] Ning Zhang, Ryan Farrell, and Trever Darrell, “Pose pooling kernels for sub-category recognition,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[22] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie, “Visual recognition with humans in the loop,” in *Proc. Euro. Conf. Computer Vision*, 2010.

[23] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie, “Multiclass recognition and part localization with humans in the loop,” in *Proc. Int’l Conf. Computer Vision*, 2011.

[24] Ya-Fang Shih, Yang-Ming Yeh, Yen-Yu Lin, Ming-Feng Weng, Yi-Chang Lu, and Yung-Yu Chuang, “Deep co-occurrence feature learning for visual object recognition,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[25] M Srinivas, Yen-Yu Lin, and Hong-Yuan Mark Liao, “Learning deep and sparse feature representation for fine-grained recognition,” in *Proc. Int’l Conf. Multimedia and Expo*, 2017.

[26] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, “Online dictionary learning for sparse coding,” in *Proc. Int’l Conf. Machine Learning*, 2009.

[27] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, “Locality-constrained linear coding for image classification,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010.

[28] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, “The Caltech-UCSD birds-200-2011 dataset,” 2011.

[29] Gavves Efstratios, Basura Fernando, Cees GM Snoek, Arnold WM Smeulders, and Tinne Tuytelaars, “Fine-grained categorization by alignments,” in *Proc. Int’l Conf. Computer Vision*, 2013.

[30] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint:1409.1556*, 2014.

[31] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” *arXiv preprint:1405.3531*, 2014.