Multiple Granularity Descriptors for Fine-grained Categorization

Dequan Wang[1], Zhiqiang Shen[1], Jie Shao[1], Wei Zhang[1], Xiangyang Xue[1] and Zheng Zhang[2]

[1]Shanghai Key Laboratory of Intelligent Information Processing,
School of Computer Science, Fudan University
[2]Department of Computer Science, New York University Shanghai

[1]{dqwang12; zhiqiangshen13; shaojie; weizh; xyxue}@fudan.edu.cn [2]zz@nyu.edu

Abstract

Fine-grained categorization, which aims to distinguish subordinate-level categories such as bird species or dog breeds, is an extremely challenging task. This is due to two main issues: how to localize discriminative regions for recognition and how to learn sophisticated features for representation. Neither of them is easy to handle if there is insufficient labeled data.

We leverage the fact that a subordinate-level object already has other labels in its ontology tree. These "free" labels can be used to train a series of CNN-based classifiers, each specialized at one grain level. The internal representations of these networks have different region of interests, allowing the construction of multi-grained descriptors that encode informative and discriminative features covering all the grain levels.

Our multiple granularity framework can be learned with the weakest supervision, requiring only image-level label and avoiding the use of labor-intensive bounding box or part annotations. Experimental results on three challenging fine-grained image datasets demonstrate that our approach outperforms state-of-the-art algorithms, including those requiring strong labels.

1. Introduction

Psychologists have showed that humans do well in basic-level recognition before developing their ability of subordinate-level recognition[20]. Perhaps not coincidentally, researches in the field of computer vision have followed a twin trajectory, moving from coarse-grained to fine-grained. Fine-grained categorization, referring to subordinate-level recognition, has emerged as a popular research area in the computer vision community. In contrast to basic-level categorization, subordinate-level classification needs to explicitly discriminate against subtle differences among similar subcategories. Progress in fine-grained categorization not only expands the scope of generic object recognition, but also directly benefits domain experts.

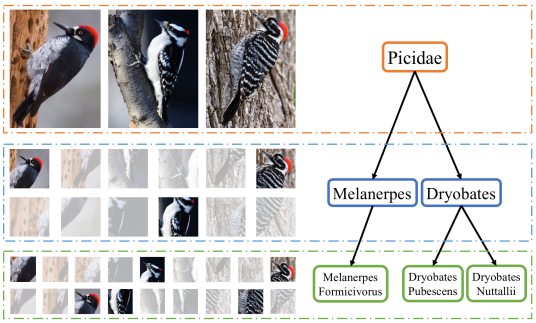

Figure 1. Any subordinate-level label has its parenting labels in the ontology tree that are informative and discriminative for classification, with respect to their grain levels. *Acorn Woodpecker*, *Downy Woodpecker* and *Nuttall's woodpecker* are three distinguish species, but all of them belong to the same family.

In general, fine-grained categorization is extremely challenging. This is due to two main issues: how to 1) localize discriminative regions and 2) learn their corresponding representations. Identifying such regions is critical, as they contain informative details for subordinate-level categorization. As an example, domain experts always characterize bird species with some special parts such as spotted chest and white back. Basic-level classification only requires the "where" of the object, while fine-grained one *additionally* asks for the "where" of the most discriminative parts.

Classification at subordinate-level is difficult because of large inner-class variations and inter-class confusions, which could be resolved if sufficient labels were available. Unfortunately, unlike basic-level classification where little domain expertise is required to label the data and tools such as crowd-sourcing can be leveraged, subordinate-level recognition demands well-annotated data which is comparatively scarce and expensive to acquire.

Our idea is based on the core observation that a subordinate-level label carries with an *implied* hierarchy of labels, each corresponding to a level in the domain ontology. Following the assumption that domain experts distinguish finer classes with visually distinctive features, hierarchies thus have embedded and latent knowledge. Our goal is therefore to explore the rich semantic relationships among these extra labels. For instance, Melanerpes Formicivorus, also known as Acorn Woodpecker, can also be called Melanerpes at genus level, or Picidae at family level (Figure 1). These labels are *free* for extracting their corresponding discriminative patches and features.

Our framework contains a parallel set of deep convolutional neural networks, each optimized to classify at a given granularity. In other words, our framework is composed of a set of single-grained descriptors. Saliency in their hidden layers guides the selection of regions of interest (ROI) from a common pool of bottom-up proposed image patches. ROI selection is therefore by definition granularity-dependent, in the sense that selected patches are results of the associated classifier of a given granularity. Meanwhile, ROI selections are also cross-granularity dependent: the ROIs of a more detailed granularity is typically sampled from those at the coarser granularities. This is built upon the intuition we discussed earlier, by emulating the process of multi-level attention. Finally, per-granularity ROIs are fed into the second stage of the framework to extract per-granularity descriptors, which are then merged to give classification result.

Our experiments on three challenging fine-grained benchmarks, CUB-200-2011[34], Birdsnap[3] and Aircraft[27], outperform the existing state-of-art approaches, while only requiring the weakest, per-image labels, validating our approach. We find that finding the discriminative regions and extracting the corresponding features are complementary, delivering 3% and 2% performance improvements respectively. We note that while specific techniques can vary, the approach to explore richer and semantically hierarchical labels is generalizable.

In summary, we make the following contributions:

1. We overcome the scarcity of labeled data by enriching a subordinate label with its parenting labels in the taxonomy hierarchy.
2. We derive a multi-granularity learning framework that leverages the hierarchical labels to generate comprehensive descriptors.
3. We propose a two-step fine-tuning mechanism consisting of salient region localization followed by classification of patches.

The rest of the paper is organized as follows. Section 2 covers related work, Section 3 describes our framework, experiment results are reported in Section 4 and we conclude and discuss future work in Section 5.

2. Related Work

Fine-grained classification, whose domains vary from animal breeds[34][3][21] to man-made objects[27][23], becomes increasingly popular recently. In contrast to coarse-grained classification tasks, differences in appearances among objects in the same basic-level category are extremely subtle. Consequently, it requires more sophisticated features which in turn heavily rely on abundance of labeled training data. We attack the problem by using more implied labels from the ontology structure. From that perspective, previous work exploring similar idea is [9], which introduces Hierarchy and Exclusion (HEX) graphs to capture semantic relations between labels instead.

To our best knowledge, our approach is the first to integrate taxonomic hierarchy and feature learning in fine-grained categorization. Other related works have focused on boosting performance from the following two aspects: part localization and representation learning.

2.1. Part-based Model

To tackle the issue of the insufficient discriminative power for fine-grained categorization, some of the existing works focused on feature encoding, such as part-based or pose normalized frameworks[5][36][38][37] and segmentation-based methods[1][7][14]. [38] made use of deformable part models[13] for pose-normalized representations of fine-grained objects. Several approaches used poselets[4], a part detector based on key-points to capture a specific viewpoint and pose, as a part localization method for fine-grained classification[12][37].

Part-based representations are very prevailing for object recognition from basic-level to subordinate-level, but it is unclear what those parts should be at different granularity levels. Some existing approaches[36][5][26] regard CNNs as visual descriptors and provide part-level supervisions to guide learning progress. These pipelines attempt to capture image patches containing subtle differences from training images. The strongest supervision setting requires part information in both training and testing phase, which is an unrealistic requirement in practice. Our approach localizes and learns granularity-specific features automatically, using only image-level labels and their ontology structures.

2.2. Representation Learning

Since domain training data is limited, [15] proposed a transfer learning strategy, where a convolutional neural network [25] is first pre-trained on ImageNet[24] and then fine-tuned on the smaller task-specific dataset. Meanwhile, several other researches reported similar settings on a wider range of visual tasks[29]. [35] designed domain-nets and filter-nets to simulate two-level attention mechanism like human beings. [30] proposed recurrent neural net-

work simulating attention mechanism while utilizing pre-trained CNNs as region feature extractors. [22] utilized co-segmentation and alignment to generate parts without extra annotations. After fine-tuning, CNNs have shown great promise in capturing a great amount of feature representations with its tremendous learning power. Our work shares the same spirit, but extends it in the multi-granularity framework with a carefully designed two-step refining process.

3. The Proposed Model

3.1. Overview

An object at subordinate-level carries with it all its parenting labels along the path of its taxonomy tree. As an example, downy woodpecker is Picidae, Dryobates and Dryobates Pubescens at the family-, genus- and species-grain, respectively. The question is how to leverage these labels. Our idea is to derive per-grain descriptor, and combine them into what we call the multiple granularity descriptors, or MGD in short.

Ideally, MGD should correspond well with how human detect such objects, and its constituent descriptors should complement each other. For instance, the family-grained descriptor focuses on body and shape, whereas genus- and species-grained descriptor describes increasingly localized details. Our goal is to derive these descriptors without knowing their spacial distribution on raw image in advance.

Consider the problem of getting an informative single-grain descriptor. Suppose a bottom-up process has proposed a pool of image patches (or regions). The next step is to select the patches that are most discriminative for classification at this grain-level. Once such patches are selected, refining a network to derive the grain-specific descriptor becomes possible (the right part of Figure 2 and Section 3.3).

We choose to generate grain-specific regions of interests (ROIs) first. This is described in the left half of Figure 2. Importantly, ROIs are generated by detecting high saliency points in the heatmap of a network fine-tuned for a given grain. Figure 3 gives a preview of multi-grained heatmap in comparison to crowd-sourcing interests[11]. The score of an image patch, a measure of how relevant it is to a given grain, is then evaluated against the ROIs. Section 3.2 describes this part of the pipeline in more detail.

The complete pipeline of a grain has a clear division of labor. The network that generates ROIs to pick grain-specific region candidates is called the *detection network*, and the network that generates feature representation is called the *description network*. Detection network is fine-tuned from a generic CNN using its associated grain labels, with original raw input image; description network is further refined from detection network with the same label set, but is instead fed with the image regions selected by ROIs from the detection network. To map well with the intuition given earlier, ROIs are loosely regularized across-grain: ROI candidates of a detailed grain are encouraged to be part of the ROIs of the next coarse level, and thus operate at a more localized level. This two-step refinement process and the composition of MGD are the subject of Section 3.3 and Section 3.4, respectively.

3.2. Region Discovery

Generating Saliency Heatmap We first create multiple granularity detection networks. These networks are refined from the same VGGNet pre-trained on ImageNet, feeding each with an entire image with one grained label (*e.g*. fine-grained detection CNN is fed with labels at species level). After training, we obtain 512 channels of filter response map from the last pooling layer of VGGNet. Detection CNNs produces heatmap of spatial distribution of ROIs deep inside network, since filter parameters are learned from domain-specific training data. Our goal is to uncover saliency in their hidden layers to guide selection of ROIs.

Stacking together the feature maps across all channels generates a heatmap that is polluted with cluttered activations driven by unrelated background noises. Therefore we perform data preprocessing. Inspired by [18], we applies a map normalization operator $\mathcal{N}(\cdot)$, its effect is to globally suppresses numerous comparable peak responses and simultaneously enhances strong peaks. The steps are:

1. Normalizing the map to a fixed range $[0 \dots M]$, here M is shared across maps
2. Computing the average $\bar{m}$ of all other elements
3. Cultiplying the map by $(M - \bar{m})^2$

The normalization operator $\mathcal{N}(\cdot)$ compares the maximum response in a map to its average overall activations by figuring out the difference between the maximum and the average. When this difference is large, the peak stands out naturally. On the contrary, when this difference is small, the map is suppressed as whole. $\mathcal{N}(\cdot)$ derives from biological mechanisms, cortical lateral inhibition, where the neighboring similar features inhibit each other via specific, anatomically defined connections[6].

The heatmap that describes the spatial distribution of interest points is then produced:

$$\varphi(\boldsymbol{I}) = \sum_{i=1}^{512} \mathcal{N}(\varphi_i(\boldsymbol{I})) \tag{1}$$

where $\varphi_i(\cdot)$ denotes the i-th filter response map of image $\boldsymbol{I}$ while $\varphi(\cdot)$ indicates the heatmap.

Identifying Region of Interest Our next step is to filter irrelevant patches among those proposed by bottom-up mechanism adaptively, guided by energy distribution of heatmap. Although heatmap highlights some discriminative area, between the poles of foreground and background there

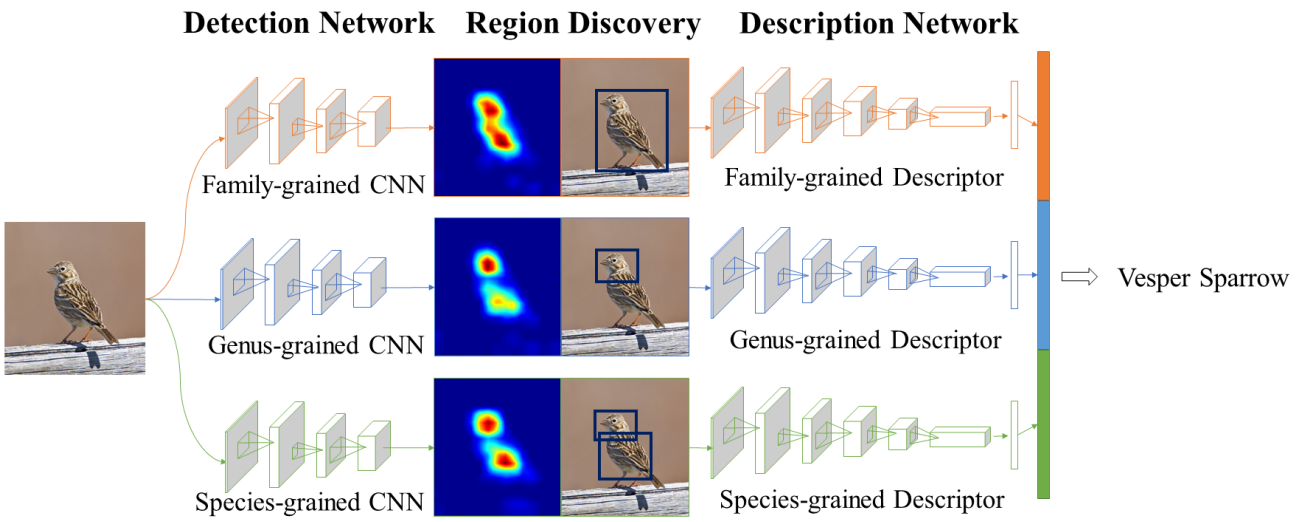

Figure 2. Overview of our Multiple Granularity Descriptors (MGD) framework.

still lie intermediary areas. Another issue we need to deal with is that CNNs are sensitive to certain texture pattern, and as a result background clutter may be more influential (*e.g.* due to bird's protective coloration).

We employ [17], an interactive image segmentation tool, which introduced geodesic star convexity for interactive image segmentation. Its principle is to fully utilize star-convexity as shape cues[33]. Global foreground estimation helps filter background patches while extending object-level regions from the highest energy points of interest. Here we define *energy* as the sum of all enclosing elements of heatmap. Instead of operating on input image, we apply it to heatmap and simulate a robot user. Pixels with higher energy are regarded as foreground seeds, whereas lower energy areas are background seeds, and the rest are labeled as unknown. Thus we establish a mapping of pixels of heatmap to three seedmap labels $(foreground, unknown, background)$, colored as $(white, gray, black)$ respectively. At the end, we generate foreground segmentation rather than bounding box, with the hope to produce more accurate location of the object.

Our next step is to choose the regions. We measure confidence of foreground regions by pixel-level overlap of region candidate and object segmentation. Specifically, we calculate ratio of Intersection-over-Union(IoU) as follows:

$$\phi(region) = \frac{Area(region \cap seg)}{Area(region \cup seg)} \tag{2}$$

where $\cap$ and $\cup$ denotes intersection and union while $Area(\cdot)$ is the number of pixels. seg is the foreground pixels discovered in the previous step. At the same time, we calculate the density of box's energy, which implies the confidence of discriminative regions:

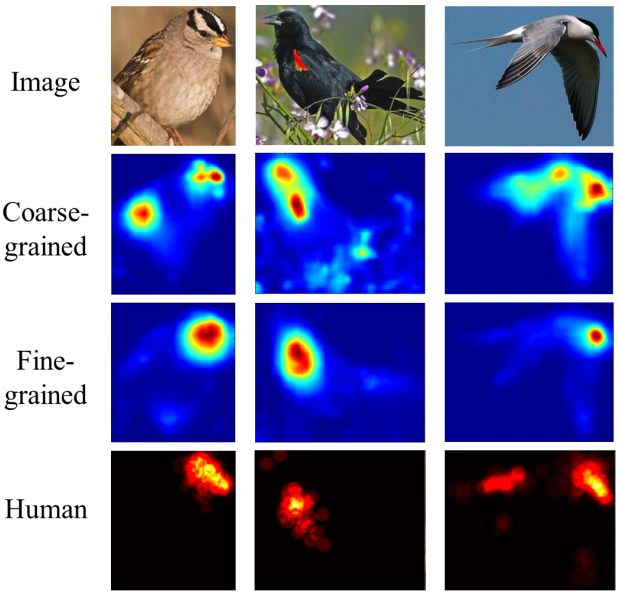

Figure 3. Heatmap of human and multi-grained CNNs interests.

$$\rho(region) = \frac{Energy(region)}{Area(region)} \tag{3}$$

where $Area(\cdot)$ denotes the area of box. Finally, we compute the score of a region as:

$$\psi(region) = \phi(region) \cdot \rho(region) \tag{4}$$

The idea is to unite two factors: low-level visual clues given by segmentation of foreground object, measured by pixel-level overlap, and high-level semantic hints obtained from heatmap via density of energy.

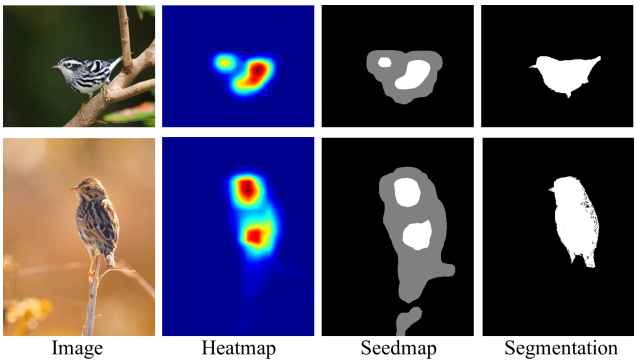

Figure 4. Process of interactive segmentation based on heatmap.

3.3. Two-Step Finetuning

The bottom-up proposed regions are divided into three groups according to their size: large, medium and small. They are now distributed to the detection networks according to this rule: coarse-grain takes the large regions only, mid-grain picks both large and medium, and the finer-grain takes all three. The rational is that finer-grain is the hardest to classify due to subtle differences, and thus needs as many candidate regions as possible.

Given the set of regions assigned to it, a detection work ranks them according to its ROIs as is described in Section 3.2, and picks positive and negative samples according to two thresholds (in our experiment, they are 0.35 and 0.15 determined by grid search, for high and low respectively). They are now fed to description network, which is initialized from its corresponding detection network.

Figure 5 illustrates the hierarchical supervision for multiple granularity CNNs. For instance, the top level of taxonomic tree, such as order-grained or family-grained labels, tells woodpecker from sparrow and tanager. The finer grained label, such as genus-grained or species-grained at the bottom level, provides more details about red-cockaded comparing to downy and pileated. As for input image patches, larger regions like object-level patches are fed into all three levels, while smaller regions like part-level patches are utilized to distinguish subtle differences.

3.4. Multiple Granularity Descriptor

Based on region proposals from detection networks, multi-grained descriptors can be extracted by description networks of different grains. The image I can be represented as a set of multiple granularity descriptors:

$$MGD = \{F_1(R_1), F_2(R_2), \ldots, F_L(R_L)\} \qquad (5)$$

where F, R denote the specific grained feature vector and the corresponding regions of interest respectively, and L is the total number of granularity. Note that at test time, each grain only selects the highest scored region.

Our final feature space concatenates the outputs of the first fully connected layer of multiple descriptor networks across grains. In general, the progression through the 19-layer VGGNet can be seen as a movement from low to mid to high-level features. The pooling layers aggregate complex structural information, with max-pooling operation grab hold of deformable parts, and the later fully connected layers summarize complex co-occurrence statistics and remove the influence of spatial displacement. To handle fully connected layers with different scales of magnitude, each representation is normalized independently.

Finally, we employ a linear SVM to learn weights for the final classification. The application of a linear SVM instead of softmax layers by CNNs is mostly for technical convenience to combine multi-grained features.

4. Experiments

In this section, we report our experiment results. First, we compare the framework's overall performance against state-of-the-art algorithms on three standard subcategory datasets. We then perform detailed analysis on the gains contributed by individual components.

We start with the 19-layers VGGNets[31] pre-trained on of ImageNet[10], with the publicly available implementation Caffe[19]. In our framework, each grain refines a VGGNet, and then 4096-dimensional representation after the first fully-connected layer is extracted. Therefore, a three-grained pipeline has an internal representation of 12,288 dimensional global features. We utilize Selective Search[32] to obtain the initial hypothesis of regions of interest for each image. Multi-grained labels references for a three-level taxonomic hierarchy (family, genus and species) come from the followings: American Ornithologists' Union Check-list of North American Birds[8] and WordNet[28]. For a fair comparison, we try to reproduce experiment of state-of-the-art method[36] on the same baseline network.

4.1. Comparison with State-of-the-art Methods

We report our results on three challenging datasets: CUB-200-2011[34], Birdsnap[3] and Aircraft[27] where classification accuracy indicates average over test samples.

CUB-200-2011 This dataset contains 200 bird species and 11,788 images, and is a wildly-used fine-grained classification benchmark. Each image in CUB-200-2011 has rich supervision, including image-level label, bounding box and fifteen part landmarks. For fair comparison, we use standard dataset split, with about 30 samples of each breed for both training and testing phases.

Birdsnap This dataset contains 49,829 images of 500 of the most common species of North American birds with most species have 100 images. Each image in BirdSnap is labeled with a bounding box and seventeen parts landmarks along with image-level label. Compared to CUB-200-2011,

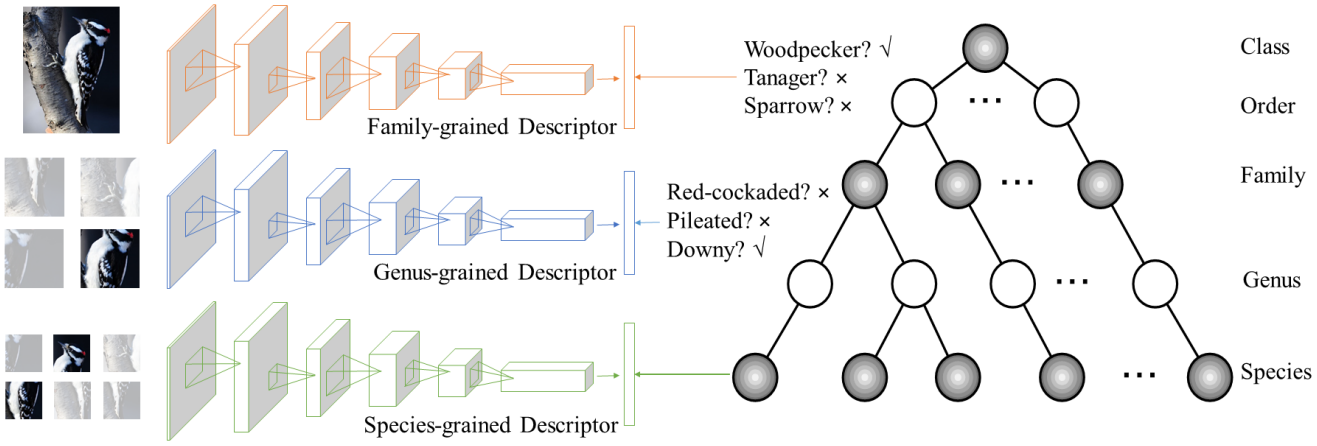

Figure 5. Illustration of the hierarchical supervision for multiple granularity descriptors.

it contains much more species and images, well reflecting the visual variation among almost all the commonly birds in the United States. We follow [3] experimental setting and hold out a test set of 2,443 images.

Aircraft This dataset contains 100 aircraft variants and 10,000 images. Each image in Aircraft has both image-level label and bounding box annotation. The training and test set consists of around 67 and 33 images per class, respectively. Compared to birds, airplanes usually occupy a large area of whole image with relatively clean background. However, the difficulty can be substantial inter-class confusion: the difference of Boeing 737-300 and Boeing 737-400 may only appear in the number of windows in the model. For fair comparison, we use standard evaluation metric mA, average classification accuracy by category.

Quantitative and Qualitative Results Comparisons of our performances (termed "multi-grained") against other algorithms are given in Table 1, 2 and 3, along with requirements of supervision. To understand the performance difference caused by imperfect location detection, we also report results when we replace segmentation in Eq.(2) with the ground truth bounding box.

Interestingly, our framework delivers the best result with and without bounding box. In CUB-200-2011 dataset, it is even better than the methods that require part-level annotation. The only exception is when the oracle bounding box is also provided at the test time.

Some example segmentation results of our multiple granularity descriptors are shown in Figure 6, including both successful and failure cases. The failures are mainly caused by intra-class variability which remains challenging, where cluttered background shares high visual similarities with the undetected objects.

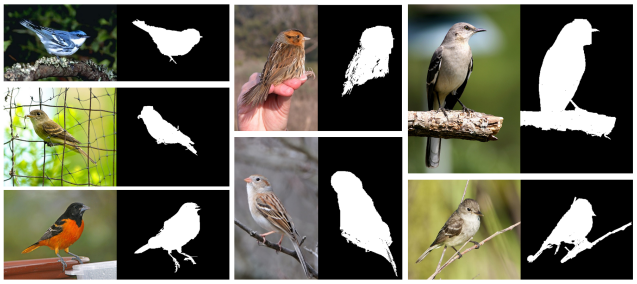

Figure 6. Some example segmentation results by our approach on CUB-200-2011[34] and Birdsnap[3] datasets. Successful segmentations (left 2 columns) and failure cases (right column).

4.2. Individual Component Effectiveness

The performance of our framework is affected by different factors as follows:

1. Choice of descriptor is the most significant
2. More accurate region discovery has an impact, more so when accuracy is overall quite low
3. Combining multiple regions of interest helps both fine-tuning and classification

We now elaborate on these conclusions, using the well studied CUB-200-2011. The results are listed in Table 4.

Analysis of Multiple Granularity In upper half of Table 4, we show how single-, double- and triple-grained differ. Independent of whether ground truth bounding box is used, performance steadily improves.

Analysis of Region Discovery The difference between results using the same number of grained pipeline but with or without bounding box is caused by region detection accuracy. The gap is not significant, but meaningful. Earlier, we have shown that this gap is more significant for the other two benchmarks, where the general accuracy is lower.

Methods	Feature	BBox	Part	Oracle BBox	Oracle Part	Accuracy (%)
Zhang *et al*. [38]	KDES	√		√		51.0
Chai *et al*. [7]	Fisher	√		√		61.0
Gavves *et al*. [14]	Fisher	√		√		62.7
Zhang *et al*. [38]	KDES	√	√	√	√	64.5
Berg *et al*. [2]	POOF	√	√	√	√	73.3
Zhang *et al*. [36]	AlexNet	√	√			73.5
Branson *et al*. [5]	AlexNet	√	√			75.5
Zhang *et al*. [36]	AlexNet	√	√	√		76.7
Lin *et al*. [26]	AlexNet	√	√			80.3
Zhang *et al*. [36]	VGGNet	√	√			81.6
Krause *et al*. [22]	VGGNet	√				82.0
Zhang *et al*. [36]	VGGNet	√	√	√		85.0
Multi-grained	VGGNet	√				**83.0**
VGG-19[31]	VGGNet					67.0
Xiao *et al*. [35]	AlexNet					69.7
Xiao *et al*. [35]	VGGNet					77.9
Multi-grained	VGGNet					**81.7**

Table 1. Quantitative results on the CUB-200-2011 dataset [34] in comparison with state-of-the-art methods.

Methods	Annotation	Accuracy (%)
VGG-19[31]	BBox	62.3
Berg *et al*. [3]	BBox	66.6
Multi-grained	BBox	**74.8**
VGG-19[31]	None	51.7
Multi-grained	None	**65.9**

Table 2. Quantitative results on the Birdsnap dataset [3] in comparison with state-of-the-art methods.

Methods	Annotation	mA (%)
VGG-19[31]	BBox	63.2
Chai *et al*. [7]	BBox	75.8
Gosselin *et al*. [16]	BBox	81.5
Multi-grained	BBox	**86.6**
VGG-19[31]	None	56.6
Multi-grained	None	**82.5**

Table 3. Quantitative results on the Aircraft dataset [27] in comparison with state-of-the-art methods.

Analysis of Two-Step Finetuning Once a region is discovered, how its associated features are extracted makes a difference. We could feed it into detection CNN (which is refined from ImageNet with the target grain labels) to extract features (second half of Table 4). Or, alternatively let the detection CNN act as region of interest generator which picks up potential region hypothesis according to feature map scores. The selected domain relative patches are used to train the description CNN. Results show that, for CUB-200-2011 dataset, this brings the most significant gain.

Methods	Annotation	Accuracy (%)
Single-grained	BBox	81.2
Double-grained	BBox	82.4
Multi-grained	BBox	83.0
Single-grained	None	79.5
Double-grained	None	81.0
Multi-grained	None	81.7
Detection CNN	BBox	77.3
Description CNN	BBox	81.2
Detection CNN	None	76.2
Description CNN	None	79.5

Table 4. Evaluation of individual components contributing to the overall performance on CUB-200-2011 dataset[34].

5. Conclusion

In this paper, we propose a fine-grained categorization framework that is trained from multiple granularity labels. Our framework can simultaneously handle both representation learning and region discovery, and discover features across grain levels fully automatically.

Experimental results on three challenging datasets, CUB-200-2011, Birdsnap and Aircraft demonstrate that our method outperforms most of the existing approaches, and makes it under the weakest supervision signal.

Our approach is generalizable, and we will actively look into ways to adapt the idea to other algorithms. We also plan to make more aggressive use of taxonomic hierarchy and on other difficult fine-grained domains.

Acknowledgements

We would like to thank anonymous reviewers for helpful feedback. We would also like to thank Tianjun Xiao and Hao Ye for useful discussions. This work was supported in part by Natural Science Foundation of China (No.61473091), STCSM (No.15JC1400103), EU FP7 QUICK Project under Grant Agreement (No.PIRSESGA-2013-612652) and Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment(CURE).

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR*, 2013.

[2] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[3] 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 *CVPR*, 2014.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[5] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *BMVC*, 2014.

[6] M. W. Cannon and S. C. Fullenkamp. A model for inhibitory lateral interaction effects in perceived contrast. *Vision research*, 1996.

[7] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.

[8] R. T. Chesser, R. C. Banks, C. Cicero, J. L. Dunn, A. W. Kratter, I. J. Lovette, A. G. Navarro-Sigüenza, P. C. Rasmussen, J. Remsen Jr, J. D. Rising, et al. Fifty-fifth supplement to the american ornithologists' union check-list of north american birds. *The Auk*, 131(4):CSi–CSxv, 2014.

[9] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*. 2014.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[11] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[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 *ICCV*, 2011.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

[14] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[16] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin. Revisiting the fisher vector for fine-grained classification. *Pattern Recognition Letters*, 2014.

[17] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *CVPR*, 2010.

[18] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 1998.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 2014.

[20] K. E. Johnson and A. T. Eilers. Effects of knowledge and development on subordinate level categorization. *Cognitive Development*, 1998.

[21] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPRW*, 2011.

[22] J. Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *CVPR*, 2015.

[23] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[25] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1995.

[26] D. Lin, X. Shen, C. Lu, and J. Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 2015.

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

[28] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[29] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *CVPRW*, 2014.

[30] P. Sermanet, A. Frome, and E. Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[32] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[33] O. Veksler. Star shape prior for graph-cut image segmentation. In *ECCV*. 2008.

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

[35] 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. *CVPR*, 2015.

[36] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

[37] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *CVPR*, 2012.

[38] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.