Cross-Category Cross-Semantic Regularization for Fine-Grained Image Recognition

Yelin Chen, Xianjie Mo, Zijun Liang, Tingting Wei, and Wei Luo(✉)

South China Agricultural University,
Guangzhou 510000, GD, People's Republic of China
{cvychen,liangzijun}@stu.scau.edu.cn, cedricmo.cs@gmail.com,
weitingting@scau.edu.cn, cswluo@gmail.com

Abstract. Fine-grained image recognition (FGIR) is challenging due to the local and subtle differences between subordinate categories. Existing methods adopt a two-step strategy by first detecting local parts from images, and then extracting features from them for classification. Although steady progress has been achieved, these methods localize object parts separately while neglecting the relationships between them. In this paper, we propose cross-category cross-semantic (C^3S), a regularization module that exploits the relationships between object parts from different images to regularize the fine-grained feature learning for FGIR. C^3S encourages the features of the same object part from different images to have strong correlations while decorrelating the features from different object parts as much as possible. C^3S can be incorporated into networks without introducing any extra parameters. Experiments on five benchmark datasets (CUB-200-2011, Stanford Dogs, Stanford Cars, FGVC-Aircraft and NABirds) validate the effectiveness of C^3S and demonstrate its comparable performance to existing methods.

Keywords: Fine-grained image recognition · Deep convolutional neural networks

1 Introduction

Fine-grained image recognition (FGIR) refers to distinguishing objects into subordinate categories, e.g., species of birds [29], models of cars [15], breeds of dogs [13], etc. Compared to base-class recognition, images from subordinate categories often exhibit more subtle and regional visual differences, which make recognition models difficult to learn discriminative and robust feature representations. The recognition performance of FGIR has undergone significant improvements in recent years, thanks to the progress in the design and training of deep neural networks. Generally, there are two broad types of attention mechanisms that

Y. Chen and X. Mo—Equal contributions. The first author is a student.

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11857, pp. 110–122, 2019.
https://doi.org/10.1007/978-3-030-31654-9_10

are widely employed for weakly-supervised FGIR, namely, part detection that explicitly searches discriminative regions in raw images, and soft attention that constructs multiple output branches on one or more top-level layers to build higher-level attention-induced abstractions. Our study in this paper falls into the second category.

To better learn discriminative and robust feature representations, previous work [1] typically adopts a two-step strategy by first detecting local parts from images, and then extracting features from them for fine-grained classification. However, the application of this strategy is limited by a trade-off between recognition and localization ability. Based on these observations, the end-to-end learning framework has emerged that provides accurate fine-grained recognition predictions as well as highly informative regions during inference. These methods eliminate the need for alternative and multistage strategies, but lack a mechanism to exploit the relationships between object parts from different images, which usually results in degraded accuracy. [27] explores relationships between object parts by adapting a soft attention model. The model first extracts attention aware features through an one-squeeze multi-excitation structure (OSME), a module that takes input as a feature map and outputs multi-branch feature maps, and then employs a metric learning framework to mine semantic features. While achieving decent results, the complicated sample selection procedure and non-trivial optimization involved in the objective of metric learning limit its application.

In this paper, we propose C^3S, a simple but effective method that exploits relationships between object parts from different images to regularize the fine-grained feature learning for FGIR. Similar to [27], our method first extracts attention aware features through multiple excitation modules, but it further employs C^3S to guide the attention features to be extracted from distinct discriminative regions. In theory, for the attention features from different images but extracted from the same attention regions, they should be more correlated than those from different attention regions. To this end, C^3S encourages the features of the same object part from different images to have strong correlations while decorrelating the features from different object parts as much as possible. With constraints introduced by C^3S, the learned excitation modules are likely to localize distinct informative regions distributed over the whole object. Therefore, our proposed model can provides accurate fine-grained classification predictions as well as distinct discriminative regions during inference. Different from the metric learning framework, C^3S can be easily integrated into deep convolutional neural networks (DCNNs) without any sample selection procedure, resulting in increased computational efficiency. Experiments on five benchmark datasets (CUB-200-2011, Stanford Dogs, Stanford Cars, FGVC-Aircraft and NABirds) validate the effectiveness of C^3S and demonstrate its comparable performance to existing methods. Moreover, C^3S can be trained with standard back-propagation, allowing for end-to-end training of our method. Our main contributions can be summarized as follows:

– We propose a soft-attention based framework for FGIR, which learns to refine attention features end-by-end by unsupervised exploiting the relationships between the attended features.

- We design a novel regularizer, C^3S, to exploit the relationship between different attention features. C^3S encourages semantics of every attention by forcing the features of the same object part from different images to have strong correlations between each other while decorrelating the features from different object parts as much as possible.
- We conduct extensive experiments on five fine-grained benchmark datasets and present a detailed comparison analysis. The state-of-the-art performance validates the effectiveness of our method.

The remainder of this paper is organized as follows: Related work is reviewed in Sect. 2. In Sect. 3 we elaborate the OSME for attention feature extraction and C^3S. Experimental results are presented and analyzed in Sect. 4 and we conclude our work in Sect. 5.

2 Related Work

2.1 Fine-Grained Image Recognition

In computer vision, the recognition performance of coarse-grained learning has improved greatly thanks to the marked progress of deep learning architectures. However, Fine-grained image recognition (FGIR) is still a challenging task due to the subtle visual differences between subordinate classes in the object. In the task of classifying birds and dogs, it is difficult to distinguish categories owe to the distinct object postures and scenarios such as the flying bird across the forest and the standing dog in the crowd. A straightforward way of locating vital discriminative regions is to exploit manual object part annotations for better recognition. Yet it is laborious and expensive to obtain datasets with detailed annotations such as bounding box, which seriously limits the effectiveness of these methods in practice.

For the sake of implementing general networks, increasing weakly-supervised mechanisms have emerged to improve the fine-grained features learning. Jaderberg et al. [12] raise the Spatial Transformer, which allows spatial manipulation of data within the network in the weakly-supervised way. Lin et al. [18] utilize a bilinear network to capture discriminative features among subordinate categories. Other advanced methods, such as the way of leveraging the label hierarchy on fine-grained classes [38] and the strategy of integrating the average with bilinear pooling to learn a pooling way during the training [26]. These methods also perform the great prediction accuracy.

In comparison to previous methods, our method also devotes to extracting features in the weakly-supervised way but boosts the performance by encouraging the relationships between identical object parts and decorrelating the distinct parts, instead of operating spacial transform of the data or utilizing the bilinear extractor.

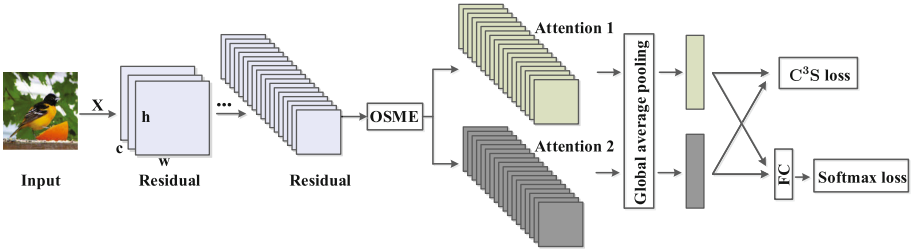

Fig. 1. Overview of our network architecture. Here we visualize the case of extracting two attention branches through OSME module in a residual network. The ultimate prediction according to softmax loss is regularized by C^3S loss which exploits the relationships between attention branches. Our model is trained end-to-end with gradients from C^3S loss and softmax loss.

2.2 Attention-Based Learning

Attention-based learning is a promising direction to address various issues of FGIR. There are two broad types of attention mechanisms that are effective for weakly-supervised FGIR. One is part detection, which explicitly searches discriminative regions in raw images. Faster R-CNN is a part detection method that contributes to perform without annotations by employing Region Proposal Network (RPN). Approaches like YOLO [25] and SSD [19] improve the detection speed in a single-shot architecture. Another attention mechanism is soft attention, which can lead the distribution of available processing resources to the most informative locations of images [10,11,16,21,24]. Wang et al. [31] propose the trunk-and-mask attention mechanism, which focuses on stacking Attention Modules for generating attention aware features that change adaptively as layers going deeper. The Squeeze-and-Excitation (SE) block [8] raised by Hu et al. exploits the channel-wise information to regulate channel weights. Above methods present the good performance in FGIR. What's more, our method is also part of soft attention, which utilizes the relationships between attention branches.

3 Method

In this section, we present our proposed method which can efficiently and accurately localize informative regions by exploiting the relationships between object parts from different images. As shown in Fig. 1, the framework of our method is composed of two parts: (1) Extracting attention features from multiple attention regions through the one-squeeze multi-excitation (OSME) module (Sect. 3.1); and (2) Guiding the attention features to represent distinct discriminative regions through C^3S (Sect. 3.2). Notably, our method can be easily integrated into existing backbone architecture; we use ResNet-50 with SE blocks as an instantiation.

3.1 Preliminaries

We briefly introduce the one-squeeze multi-excitation (OSME) [27] module for completeness before diving deep into the C^3S regularizer. OSME is a differential module that extracts attention features from multiple attention regions. Let $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_C] \in \mathbb{R}^{W \times H \times C}$ be the output feature maps of the last residual block τ. In order to generate multiple attention-specific feature maps, the OSME module extends the original SE block by performing one-squeeze and multiple-excitation operations.

Formally, in the first one-squeeze step, OSME adopts global average pooling to shrink $\mathbf{U}$ and generates a channel-wise statistics $\mathbf{z} = [z_1, \cdots, z_C] \in \mathbb{R}^C$. In the second multi-excitation step, a gating mechanism is independently employed on $\mathbf{z}$ for each attention module $p = 1, \cdots, P$:

$$\mathbf{m}^p = \sigma\Big(\mathbf{W}_2^p \delta\left(\mathbf{W}_1^p \mathbf{z}\right)\Big) = [m_1^p, \cdots, m_C^p] \in \mathbb{R}^C, \tag{1}$$

where δ and σ denote Sigmoid and ReLU [23] functions respectively. To obtain the output attention-specific features map $\tilde{\mathbf{U}}^p$ of each attention module p, we use corresponding $\mathbf{m}^p$ to re-weight the channels of the original feature maps $\mathbf{U}$:

$$\tilde{\mathbf{U}}^p = [m_1^p \mathbf{u}_1, \cdots, m_C^p \mathbf{u}_C] \in \mathbb{R}^{W \times H \times C}. \tag{2}$$

The OSME block can extract multiple attention-specific features, but still lacks a mechanism to guarantee that these features come from distinct discriminative regions. [27] addresses this by formulating a metric learning framework to pull same-attention same-class features closer and push different-attention or different-class features away. However, optimizing the non-trivial objective of metric learning is still challenging in practice.

3.2 Cross-Category Cross-Semantic Regularizer

Different from the metric learning framework introduced in [27], we propose to regularize the attention-specific features learning by exploiting the relationships between object parts from different images and different excitation modules. In theory, for the attention features from different images but extracted from the same attention regions, they should be more correlated than those from different attention regions. To this end, we design the cross-category cross-semantic regularizer (C^3S) that encourages the features from the same excitation module to have strong correlations between each other while decorrelating the features from different excitation modules as much as possible.

To obtain the feature vector $\mathbf{f}^p \in \mathbb{R}^C$, we first adopt global average pooling (GAP) on corresponding attention-specific features map $\tilde{\mathbf{U}}^p$, and then we scale them down to unit vectors through ℓ_2 normalization ($\mathbf{f}^p \leftarrow \mathbf{f}^p / \|\mathbf{f}^p\|$). Thus we encode the relationships between all pairs of excitation module p and p' into a symmetric matrix S:

$$S_{|p,p'|} = \frac{1}{N^2} \sum \mathbf{F}^{p\mathrm{T}} \mathbf{F}^{p'}, \tag{3}$$

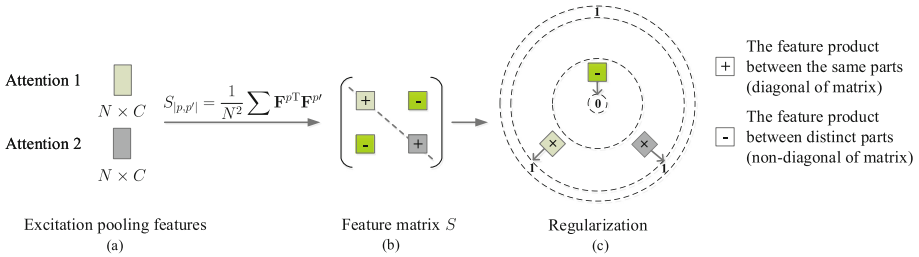

Fig. 2. An illustration of the cross-category cross-semantic regularizer. Assume the amount of excitation P in Eq. 4 is 2. (a) Each attention branch is composed of N image features extracted through OSME and GAP. (b) The interactive features between each branch are contained into a feature matrix. (c) Constraint leads feature products in symmetric matrix to the corresponding orientation to decrease the loss.

where $S_{|p,p'|}$ indicates an element of the matrix, N is the batch size, T is the transpose operator and $\mathbf{F}^p = [\mathbf{f}_1^p, \cdots, \mathbf{f}_N^p] \in \mathbb{R}^{C \times N}$ is a matrix storing the feature vectors from excitation module p for total samples in the batch.

The C^3S regularization loss is then constructed from two key components: (1) maximizing the diagonal of S to maximize the relationship between same -excitation features and (2) minimizing the non-diagonal of S to minimize the relationship between different-excitation features:

$$\mathcal{L}_{C^3S}(S) = \frac{1}{2}\Big(\sum S - \mathrm{tr}\,(S)\Big) + \Big(P - \mathrm{tr}\,(S)\Big) = \frac{1}{2}\Big(\sum S - 3\mathrm{tr}\,(S) + 2P\Big) \quad (4)$$

where $\mathrm{tr}(S)$ indicates the sum of the elements on the main diagonal of S. And in our method, we set the amount of excitation module P as 2. Different from the metric learning framework, C^3S regularization loss can be naturally inserted into the OSME block and is easily optimized without any sample selection procedure.

Therefore, we minimize the total loss:

$$L = L_s + \lambda L_{C^3S}, \quad (5)$$

where L_s is the softmax loss and λ is a balance weight. Our framework can be trained end-to-end using stochastic gradient descent (SGD).

4 Experiments

4.1 Datasets

Our experiments are carried out on five benchmark datasets, CUB-200-2011, Stanford Dogs, Stanford Cars, FGVC-Aircraft and NABirds, which are challenging to the network due to their complex textures and very subtle divergences among classes. All of these datasets are accompanied with additional bounding box despite of our weakly-supervised learning method needing no annotations but class labels. The statistics of these five datasets are shown in Table 1.

Table 1. Statistics of benchmark datasets.

Datasets	#total	#class	#train	#test
CUB-200-2011	11,788	200	5,994	5,794
Standford dogs	20,580	120	12,000	8,580
Standford cars	16,185	196	8,144	8,041
FGVG-Aircraft	10,000	100	6,667	3,333
NABirds	48,562	555	23,929	24,633

CUB-200-2011. CUB-200-2011 [29], extended on CUB-200 [32], has 11788 images in 200 categories and is diffusely used by most advanced models. Since the amount of images in each training class is only up to 30, it is a challenge to train a model in CUB-200-2011.

Stanford Dogs. This dataset covers 120 class labels and 20,580 images. Although dogs are easier to be distinguished compared to birds, it is also hard to recognize accurately in such a dataset owe to the complex background in images which disturbs operators to extract features.

Stanford Cars. Stanford Cars consists of 196 class labels and 16185 images. Various angles of cars are shoot in the images which enhance the difficulties in training. The production year and the model of cars are able to be identified through the labels.

FGVC-Aircraft. There are 10,000 images and 100 classes contained in the aircraft set. The ratio of the train set and the test set is approximately $1:2$. FGVC-Aircraft was once used as part of the fine-grained recognition challenge FGComp 2013.

NABirds. NABirds, a dataset about North American birds, is the largest one in the datasets we exploit, which occupies 48562 images and 555 categories. There are abundant bird postures in the images (e.g. bending down to capture meats and pecking on the tree), which notably increase the classification difficulty. Each class label is noted by number.

4.2 Implementation

In our experiments, images are first resized to 448×448 and then disrupted randomly in training. Meanwhile, we apply the random cropping on the birds dataset such as cubbirds and nabirds in training while exerting the center cropping in testing. All experiments are conducted in 45 epochs with a batch of 32 images as input. The base learning rate is set to 0.01 in all datasets excluding stdogs which is set to 0.001 and decayed by 0.1 for every 15 epochs. Moreover, the momentum of the optimization function is set to 0.9 while the gamma is 0.1. The value of the balance weight λ in Eq. 5 is depended on the dataset we opt and the amount of excitations in OSME module is set to 2 in the whole experiments.

Table 2. Results on CUB-200-2011. "Anno." means extra annotations (bounding box or part) utilized in training. "Acc." stands for the top-1 accuracy in probabilities.

Method	Anno.	Acc.
DeepLAC [17]	√	80.3%
Part-RCNN [34]	√	81.6%
MG-CNN [30]	×	81.7%
ResNet-50 [6]	×	81.7%
PA-CNN [14]	√	82.8%
RAN [31]	×	82.8%
MG-CNN [30]	√	83.0%
B-CNN [18]	×	84.1%
ST-CNN [12]	×	84.1%
FCAN [20]	×	84.3%
MAMC-OSME-ResNet-50 [27]	×	**8**6.3%
Ours (SE-ResNet-50)	×	82.8%
Ours (OSME-ResNet-50)	×	83.5%
Ours (OSME-ResNet-50+C^3S)	×	86.0%

To certify our C^3S constraint is effective and powerful, we undertake experiments on five datasets. Our baseline, OSME-ResNet-50, is on the basis of ResNet and the performance of the net can be promoted through our C^3S constraint prominently. To better illustrate the superiority of our novel method, we reimplement SE-ResNet as a contrast. At the same time, We compare our outcomes with other distinctive network cited from authors' papers, the according results are elucidated in tables. All the experiments testify that our method has the capacity to enhance the performance of the baseline.

4.3 Comparison with State-of-the-Arts

Quantitative experimental results are shown in Table 2, 3, 4, 5 and 6.

Results on CUB-200-2011. The experimental results are shown in Table 2. It's observed that with our C^3S, OSME-ResNet-50 excels most of state-of-the-art models. While the OSME-ResNet-50 exceeds the SE-resnet-50 by 0.7%, our method improve the performance of OSME-ResNet-50 by 2.5% significantly. What's more, for supervised methods with extra bounding box, such as MG-CNN [30], PA-CNN [14] and Part-RCNN [34], our method outperforms them by 3.0%, 3.2% and 4.4%, respectively. And compared to methods trained without extra annotations, our model surpasses FACN [20] and ST-CNN [12] by 1.7% and 1.5%. At the same time, our result is closed to the accuracy of MAMC-OSME-ResNet50 [27] which is running in the identical baseline as us and transcends ours by 0.3%. However our method is easier and more general to achieve compared

Table 3. Results on Stanford Dogs. "Anno." means extra annotations (bounding box or part) utilized in training. "Acc." stands for the top-1 accuracy in probabilities.

Method	Anno.	Acc.
PDFR [35]	×	72.0%
ResNet-50 [6]	×	81.1%
DVAN [36]	×	81.5%
RAN [31]	×	83.1%
FCAN [20]	×	84.2%
MAMA-ResNet-50 [27]	×	84.8%
MAMA-ResNet-101 [27]	×	85.2%
Ours (SE-ResNet-50)	×	83.7%
Ours (OSME-ResNet-50)	×	85.9%
Ours (OSME-ResNet-50+C^3S)	×	**87.3%**

Table 4. Results on Stanford Cars. "Anno." means extra annotations (bounding box or part) utilized in training. "Acc." stands for the top-1 accuracy in probabilities.

Method	Anno.	Acc.
DVAN [36]	×	87.1%
RAN [31]	×	91.0%
B-CNN [18]	×	91.3%
FCAN [20]	√	91.3%
MACNN [37]	×	92.8%
MAMC-ResNet-50 [27]	×	92.8%
MAMC-ResNet-101 [27]	×	93.0%
Ours (SE-ResNet-50)	×	91.7%
Ours (OSME-ResNet-50)	×	91.9%
Ours (OSME-ResNet-50+C^3S)	×	**93.7%**

to the MAMC for the reason that our model needn't compose three types of triplets to learn. Although CUB-200-2011 contains relatively fewer samples in the training set, our method still presents the good performance.

Results on Stanford Dogs. Table 3 exhibits the results on Stanford Dogs. We introduce the accuracy of methods operating without additional annotations as a contrast. The condition where our method compared to FCAN [20] which is on the basis of semantic segmentation and costs much computing resource, the outcomes claim explicitly that our method is effective (87.3% vs 84.4%). Especially in this dataset, our method surpasses the method of MAMC [27] greatly by 2.5% on the ResNet-50 as well as 2.1% on the ResNet-101. As for the original OSME-ResNet-50, our method can improve it by 1.4%.

Table 5. Results on Stanford FGVG-Aircraft. "Anno." means extra annotations (bounding box or part) utilized in training. "Acc." stands for the top-1 accuracy in probabilities.

Method	Anno.	Acc.
B-CNN [18]	×	84.1%
RA-CNN [5]	×	88.2%
HIHCA [2]	×	88.3%
Boost-CNN [22]	×	88.5%
MACNN [37]	×	89.9%
NTS-Net (K=2) [33]	×	90.8%
NTS-Net (K=4) [33]	×	91.4%
Ours (SE-ResNet-50)	×	89.8%
Ours (OSME-ResNet-50)	×	90.2%
Ours (OSME-ResNet-50+C^3S)	×	**91.9%**

Table 6. Results on NABirds. "Anno." means extra annotations (bounding box or part) utilized in training. "Acc." stands for the top-1 accuracy in probabilities.

Method	Anno.	Acc.
Branson et al. [1]	√	35.7%
PC-ResNet-50 [4]	×	68.2%
GoogLeNet [28]	×	70.7%
Lp + GoogLeNet [3]	×	72.0%
Van et al. [7]	√	75.0%
DenseNet-161 [9]	×	79.4%
Ours (SE-ResNet-50)	×	82.2%
Ours (OSME-ResNet-50)	×	82.8%
Ours (OSME-ResNet-50+C^3S)	×	**83.0%**

Results on Stanford Cars. Our experiments in cars are presented in Table 4 and the prediction accuracy of our method in Stanford Cars is the highest one among the experiments of five datasets we exerted due to the relatively obvious features in cars. In this dataset, our method also surpasses the MAMC-ResNet-101 [27] by 0.7%. On the contrast with B-CNN [18], a model commanding an extra net to capture features, our model is not only more outstanding but also much simpler and more exercisable(93.7% vs 91.3%).

Results on FGVG-Aircraft. Table 5 shows the outcomes about FGVG-Aircraft. Although NTS-Net [33], a method designed in a novel learning way that produces multiple proposal regions and ranks the regions to capture the features, carries out the great performance in FGIR. However, our method also has a capability to excels the NTS-Net by 0.5%.

Results on NABirds. The consequences conducted on NABirds are presented in Table 6. It is hard for networks to learn NABirds thanks to the diversity of attitude. Even DenseNet-161 [9] only achieves 79.4%, which executes in a radical dense linking mechanism that all layers are interconnected. Yet in such a complicated dataset, our C^3S maintains the excellent property to achieve 3.6% higher than DenseNet-161.

5 Conclusion

In this paper, we proposed a method, termed cross-category cross-semantic (C^3S), which exploits the relationships between object parts from different images to regularize the fine-grained features learning for FGIR. C^3S encourages the features from the same excitation module to have strong correlations between each other while decorrelating the features from different excitation modules as much as possible. Our method can be trained end-to-end in one stage without any bounding box or part annotations. Experiments on five benchmark datasets demonstrated the effectiveness and the state-of-the-art performance of our method.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61702197, in part by the Natural Science Foundation of Guangdong Province under Grant 2017A030310261, in part by the program of China Scholarship Council.

References

1. Branson, S., Horn, G.V., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952 (2014)
2. Cai, S., Zuo, W., Zhang, L.: Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: ICCV (2017)
3. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N.: Training with confusion for fine-grained visual classification. CoRR (2017)
4. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N.: Pairwise confusion for fine-grained visual classification. In: ECCV (2018)
5. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: CVPR (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Horn, G.V., et al.: Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In: CVPR (2015)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
9. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
10. Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194 (2001)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)

12. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS (2015)
13. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: CVPR (2011)
14. Krause, J., Jin, H., Yang, J., Feifei, L.: Fine-grained recognition without part annotations. In: CVPR (2015)
15. Krause, J., Stark, M., Deng, J., Li, F.F.: 3D object representations for fine-grained categorization. In: 4th IEEE Workshop on 3D Representation and Recognition at ICCV (2013)
16. Larochelle, H., Hinton, G.E.: Learning to combine foveal glimpses with a third-order Boltzmann machine. In: NIPS (2010)
17. Lin, D., Shen, X., Lu, C., Jia, J.: Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: CVPR (2015)
18. Lin, T., Roychowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: International Conference on Computer Vision, pp. 1449–1457 (2015)
19. Liu, W., et al.: SSD: single shot multibox detector. In: ECCV (2016)
20. Liu, X., Xia, T., Wang, J., Yang, Y., Zhou, F., Lin, Y.: Fully convolutional attention networks for fine-grained recognition. arXiv preprint arXiv:1603.06765 (2016)
21. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: NIPS (2014)
22. Moghimi, M., Belongie, S.J., Saberian, M.J., Yang, J., Vasconcelos, N., Li, L.: Boosted convolutional neural networks. In: BMVC (2016)
23. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: ICML (2010)
24. Olshausen, B.A., Anderson, C.H., Essen, D.C.V.: A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. **13**(11), 4700–4719 (1993)
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR (2016)
26. Simon, M., Gao, Y., Darrell, T., Denzler, J., Rodner, E.: Generalized orderless pooling performs implicit salient matching. In: ICCV (2017)
27. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: ECCV (2018)
28. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR (2015)
29. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-UCSD birds-200-2011 dataset. Tech. rep. California Institute of Technology (2011)
30. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: ICCV (2015)
31. Wang, F., et al.: Residual attention network for image classification. In: CVPR (2017)
32. Welinder, P., et al.: Caltech-UCSD Birds 200. Tech. rep. CNS-TR-2010-001. California Institute of Technology (2010)
33. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: ECCV (2018)
34. Zhang, N., Donahue, J., Girshick, R.B., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: European Conference on Computer Vision (2014)
35. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: CVPR (2016)

36. Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S.: Diversified visual attention networks for fine-grained object classification. IEEE Trans. Multimedia **19**(6), 1245–1256 (2017)
37. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: ICCV (2017)
38. Zhou, F., Lin, Y.: Fine-grained image classification by exploring bipartite-graph labels. In: CVPR (2016)