REAPS: Towards Better Recognition of Fine-grained Images by Region Attending and Part Sequencing

Peng Zhang[1], Xinyu Zhu[2], Zhanzhan Cheng[1], Shuigeng Zhou[2], and Yi Niu[1]

[1] Hikvision Research Institute, China
{zhangpeng23,chengzhanzhan,niuyi}@hikvision.com
[2] Fudan University, Shanghai, China
{16210720101,sgzhou}@fudan.edu.cn

Abstract. Fine-grained image recognition has been a hot research topic in computer vision due to its various applications. The-state-of-the-art is the part/region-based approaches that first localize discriminative parts/regions, and then learn their fine-grained features. However, these approaches have some inherent drawbacks: 1) the discriminative feature representation of an object is prone to be disturbed by complicated background; 2) it is unreasonable and inflexible to fix the number of salient parts, because the intended parts may be unavailable under certain circumstances due to occlusion or incompleteness, and 3) the spatial correlation among different salient parts has not been thoroughly exploited (if not completely neglected). To overcome these drawbacks, in this paper we propose a new, simple yet robust method by building part sequence model on the attended object region. Concretely, we first try to alleviate the background effect by using a region attention mechanism to generate the attended region from the original image. Then, instead of localizing different salient parts and extracting their features separately, we learn the part representation implicitly by applying a mapping function on the serialized features of the object. Finally, we combine the region attending network and the part sequence learning network into a unified framework that can be trained end-to-end with only image-level labels. Our extensive experiments on three fine-grained benchmarks show that the proposed method achieves the state of the art performance.

Keywords: Fine Grained · Region Attending · Part Sequencing.

1 Introduction

Fine-grained image recognition has attracted much research interest of the computer vision community [1–3, 35], which tries to distinguish sub-ordinate categories such as car models [18], bird species [4, 14, 29], dog breeds [16] and flower categories [26] etc. Though much effort [9, 14, 15, 23, 30, 39] has been devoted to solving this problem, recognizing fine-grained images is still a challenging task due to their relatively small inter-class difference and large intra-class variation.

Roughly speaking, there are two kinds of popular frameworks for handling fine-grained categorization: *key region localization and amplification* (*abbr.* RLA) and *discriminative part learning* (*abbr.* PL). In general, RLA tries to attend and amplify the key region for capturing detailed visual representation while avoiding background disturbance. On the other hand, PL usually first localizes discriminative parts via some sophisticated part selection mechanisms such as part attentions [9, 32, 39] and convolutional responses [33, 38], and then extracts the visual representations of the selected parts by using multiple independent feature extractors. Fig. 1 (a) and (b) illustrate these two frameworks. Though previous studies proved their effectiveness, they have several inherent drawbacks. For example, RLA-based methods may miss some salient parts when progressively attending the key region, as shown in Fig. 2 (a). PL-based methods usually fix the number of salient parts to be extracted, which is unreasonable and inflexible because in certain scenarios some of the intended parts may be unavailable due to image occlusion and incompleteness, as shown in Fig. 2 (b). Furthermore, learning independent extractor for each salient part neglects the spatial correlation among these different parts, which should be useful for image recognition if properly exploited.

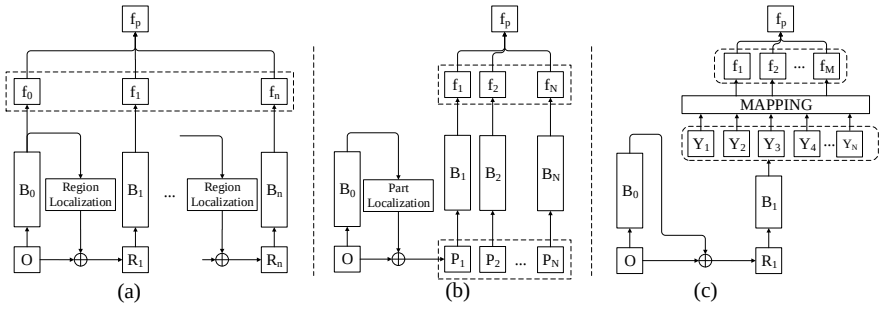

Fig. 1. An illustrative comparison between our framework and two popular existing fine-grained recognition frameworks. (a), (b) and (c) represent RLA, PL and our proposed framework respectively. O, P_i, R_i and B_i correspond to the original image, a selected part, an attended region and a backbone network respectively. $\bigoplus$ represents the operation of crop and zoom in.

These drawbacks mentioned above motivate us to develop a new method for fine-grained image recognition by simultaneously taking region attending and part sequence learning into consideration. We call the new method ***REAPS*** —— an abbreviation of **RE**gion **A**ttending and **P**art **S**equence learning. The framework of REAPS is shown in Fig. 1 (c), and its detailed architecture is illustrated in Fig. 3, which consists of two major components:

Region attending network (RAN): Inspired by previous works [9, 12], we apply the class activation mapping (CAM) mechanism [40] to constructing the

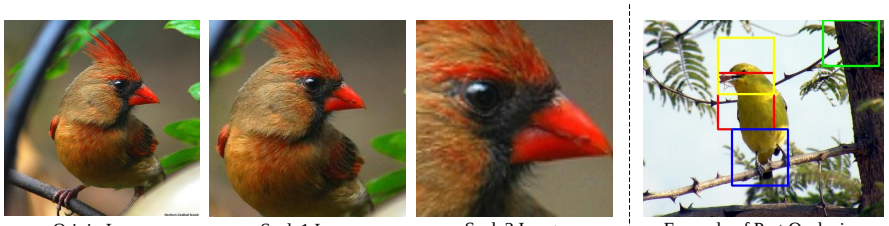

Fig. 2. Drawbacks of the RLA and PL frameworks. (a) RLA focuses on a detailed region progressively, while neglecting some other salient parts (the feet and wings of the bird disappear in scale1 and scale2 images). (b) PL detects a fixed number of preset parts and may get into trouble when some intended parts miss or be occluded (the back of the bird cannot be seen). [Best viewed in color]

region attending network for generating the region attention. The attended region cropped and amplified from the origin image are fed into the part sequence-learning network (PSN). RAN can effectively depress the impact of the background noise, while PSN can depict the detailed visual features well.

Part sequence-learning network (PSN): As mentioned above, the traditional part-based methods usually fix the number of parts and the representation of each part is explicitly learned by independent extractors. So we call them 'hard-part'-based methods. In addition to the drawbacks we previously mentioned, 'hard-part'-based methods adopt independent feature extractor (*e.g.* VGG19) for each part, which incurs high computational cost. In order to overcome the above drawbacks, we propose the 'soft-part' concept, which is implemented by mapping the serialized visual features into a group of implicit discriminative part representation and capturing the spatial correlation among different salient parts simultaneously.

In REAPS, we integrate the region attending network and the part sequence learning network into a unified framework, which can be trained end-to-end with only image-level annotations.

Our contributions are as follows: 1) We propose the novel 'soft-part' concept, and implement this concept by designing a part sequence learning network (PSN), which learns implicit discriminative part representation and captures the spatial context simultaneously. 2) We apply the region attending network to localizing the object region and alleviating the interference of complicated background to fine feature representation. 3) We integrate the region attending network and the part sequence learning network into a unified framework, and train it end-to-end without any part-level annotation. 4) We conduct extensive experiments on three challenging datasets (Stanford Cars, FGVC-Aircraft and CUB Birds), which demonstrate the superiority of our method over the existing ones.

2 Related Work

Fine-grained recognition (or categorization) is a challenging problem that has been extensively studied. Related works can be grouped in two dimensions: representation learning and part localization.

2.1 Representation Learning

Discriminative representation learning is crucial for fine-grained recognition. Thanks to their strong encoding capability, most existing fine-grained recognition algorithms [8, 9, 25, 39] employ deep convolutional networks for feature representation, which have achieved much better performance than traditional descriptors and hand-crafted features [27, 36].

To better handle the subtle inter-class difference and large intra-class variation in fine-grained recognition tasks, [21] proposes a bilinear structure to model second-order interactions of local convolutional features in a translationally invariant manner. This idea was later extended by [20] and other variants [10, 17]for better recognition performance. Recently, [5, 8] further exploit higher-order integration of convolutional activations that can yield more discriminative representation, and achieve impressive performance.

Besides, some approaches (e.g. [28], [37]) try to learn more robust representations via distance metric learning. [38] unifies deep CNN features with spatially weighted Fisher vectors to capture important details and eliminate background disturbance. [25] incorporates deep CNNs into a generic boosting framework to combine the strength of multiple weaker learners, which improves the classification accuracy of a single model and simplifies the network design.

2.2 Part Localization

Previous works have studied the localization impact of discriminative parts on capturing subtle visual difference. Early part-based approaches rely on extra annotations of bounding boxes or part landmarks to localize pre-defined semantic parts. [11, 22] assume that annotations are available in both training and testing. Some later works [14, 18, 19] use annotations only in training. However, the cost-prohibitive manually-labeled annotations prevent the application of these algorithms to large-scale real problems. Therefore, most of recent works focus on weakly-supervised task-driven part localization with only category labels. Attention-based models have been widely used to automatically discover salient parts. [33] proposes a two-level attention model, where one object-level filter-net selects relevant patches for a certain object, another part-level domain-net localizes discriminative parts. Since deep filter responses from CNNs are able to significantly and consistently respond to specific visual patterns, [38] proposes to learn part detectors by picking distinctive filters, while [30] identifies discriminative regions according to channel activations. [15] takes one step further and

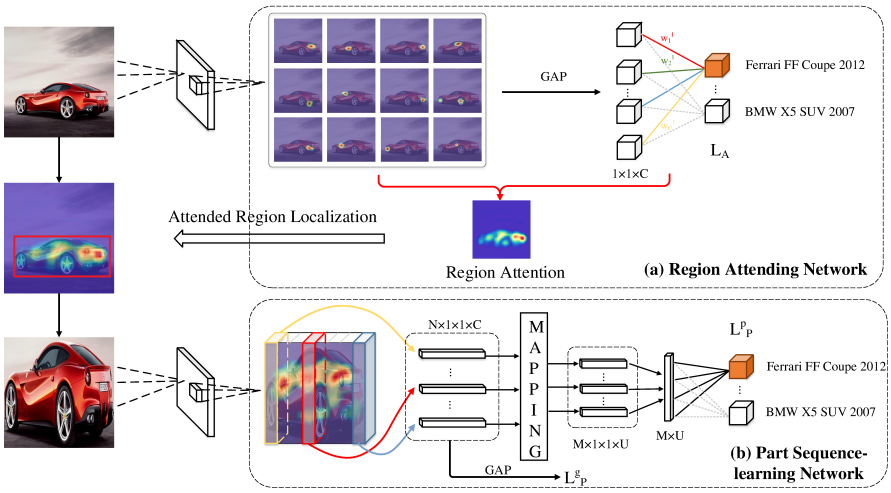

Fig. 3. The REAPS architecture. The region attending network (RAN) takes an original image as input and produces the region attention by weighting the last convolutional feature maps with the parameters of softmax layer. The attended region is cropped out and zoomed in, and fed into the part sequence learning network (PSN) where part representation is learned in an implicit way by applying the mapping function on the serialized features. The whole network can be trained end-to-end under the supervision of three softmax loss functions in Eq. (5). [Best viewed in color]

proposes a spatial transform module based on the differentiable attention mechanism, which enables CNN to learn better invariance to classification and all kinds of warping.

The latest works use hidden filter responses of deep CNNs as part detectors. [23] proposes a fully convolutional attention network to optimally glimpse local discriminative regions by reinforcement learning. [9] introduces a recurrent attention convolutional neural network (RA-CNN) that recursively learns discriminative region attention from coarse to fine by an attention proposal network. [39] develops a multi-attention convolutional neural network (MA-CNN) that generates multiple part attentions by clustering, weighting and pooling from spatially-correlated channels, and achieves the state-of-art performance.

3 The REAPS Method

3.1 Overview

The architecture of our REAPS method is shown in Fig. 3, which consists of two major components: a *region attending network* (RAN) and a *part sequence learning network* (PSN). RAN leverages deep convolutional responses to generate the discriminative region attention. Then the attended region is cropped out

and zoomed in as the input of PSN. While in PSN, the deep visual features are extracted by a backbone network and further serialized to a sequence of vectors, each of which describes a rectangle region in the raw image. A mapping function is learned to map the sequence of vectors to discriminative part representation.

3.2 Region Attending Network

The RAN is based on a standard classification network, where global average pooling is applied on the last convolutional feature maps, followed by a SoftmaxLoss layer. Given an input image $\mathcal{I}$, we can get its last convolutional feature maps, denoted as $\mathcal{F}$. In order to eliminate the background disturbance, we apply the CAM [40] mechanism to generating the region attention in a "self-guided" way. Concretely, let $f^k_{(x,y)}$ denote the activation value of unit k at (x, y) in $\mathcal{F}$, the global average pooling (*abbr.* GAP) can be represented as $\mathcal{P}_k = \sum_{x,y} f^k_{(x,y)}$. For a specific class c, the probability yielding c is $\sum_k W_{k,c}\mathcal{P}_k$, where $W_{k,c}$ is the weight of the last inner-product layer and can be learned under the supervision of SoftmaxLoss $\mathcal{L}_A$. $W_{k,c}$ also acts as the weight indicating the importance of unit k for class c. Therefore, we can compute the region attention of c at (x, y) by

$$RA^c_{x,y} = \sum_k W^c_k f^k_{(x,y)}. \tag{1}$$

Since each unit k responds to a certain type visual pattern (*e.g.* circle), the region attention can represent all the discriminative visual patterns at different locations by conducting weighted-sum over all units. Furthermore, we locate the attended region of class c by

$$AR^c = f_+(BB_\tau(RA^c), \mathcal{I}), \tag{2}$$

where BB_τ is the operation of calculating rectangular bounding box over the binary mask based on the preset threshold τ, and f_+ represents the operation of crop and zoom in.

3.3 Part Sequence Modeling

Let $X \in \mathbb{R}^{(H\times W\times C)}$ denote the deep representation through the deep convolutional neural network (the backbone network in Fig. 3), where H, W and C respectively refer to the height, width and the number of channels of X. Instead of localizing a fixed number of parts and extracting their representation separately, we learn the part representation in a soft way. Concretely, we evenly decompose X into a sequence of N vectors by

$$Y = [Y_1, Y_2, ..., Y_N] = seq(X), \tag{3}$$

where $Y_i \in \mathbb{R}^{(1\times C)}$ describes a rectangle region in the raw image and *seq* is the pooling operation with kernel size of $[H \times \frac{W}{N}]$. We abstract the sequence of visual feature vectors into M implicit parts with the learned mapping function

$$P_P = [P_1, P_2, ..., P_M] \simeq mapping(Y), \tag{4}$$

where $M \leq N$ refers to the rough number of learnt discriminative parts.

The mapping function here should keep the discriminative features and depress the useless ones. Several sequence learning techniques meet this requirement, *e.g.* the recurrent neural networks (GRU [6] and LSTM [13]) and attention-based sequence models [7, 34]. Attention-based models can fulfill the mapping from a source sequence of length N to a target sequence of length M. With sequential labels, they can effectively learn the alignment between labels and their corresponding salient representation of features [7, 34]. Recurrent neural networks (GRU [6] and LSTM [13] can strengthen the sequence representation with the sequence length being unchanged. In our case, classification is performed in a weakly-supervised fashion and the only available supervision information is the image-level category, so we apply a bi-directional LSTM on the sequenced vectors and concatenate all the hidden states as the part representation $P_P \in \mathbb{R}^{(N \times U)}$, followed by a fully connected layer and a SoftmaxLoss $\mathcal{L}_P^p$.

To accelerate the training process, another SoftmaxLoss $\mathcal{L}_P^g$ is attached to the global representation $P_g \in \mathbb{R}^C$ of the backbone network of PSN.

3.4 Training and Joint Representation

Instead of alternative optimization, REAPS can be trained end-to-end straightforwardly by

$$\mathcal{L} = \lambda_1 \mathcal{L}_A + \lambda_2 \mathcal{L}_P^g + \lambda_3 \mathcal{L}_P^p, \tag{5}$$

where $\lambda_j (j = 1, 2, 3)$ is the corresponding loss weight. Once the training process converges, the joint representation F of the input image $\mathcal{I}$ can be represented by a set of descriptors, followed by a fully-connected layer with softmax function for final classification:

$$F = \{A_g, P_g, P_p\}, \tag{6}$$

where $A_g \in \mathbb{R}^C$ denotes the global representations of the backbone network of RAN.

4 Performance Evaluation

4.1 Datasets and Implementation Details

We conduct extensive experiments on three benchmark datasets, including Stanford Cars[18], FGVC-aircraft[24] and CUB-200-2011[29], which are widely used to evaluate fine-grained image recognition. Table 1 shows the detailed statistics of the three datasets.

For fair comparison, all compared methods employ similar backbone network. Specifically, we start with the 19-layer VGGNets pre-trained on ImageNet and fine-tune it on the three fine-grained datasets. The parameters of RAN and PSN are initialized with the same pre-trained model. Input images and the cropped attended regions are both resized to 448×448, where high resolution highlights details and benefits recognition. We use SGD with momentum 0.9 to minimize

Table 1. Statistics of the fine-grained benchmark datasets used in this paper.

Dataset	Category	Training	Testing
Stanford Cars[18]	196	8,144	8,041
FGVC-aircraft[24]	100	6,667	3,333
CUB-200-2011[29]	200	5,994	5,794

the loss function $\mathcal{L}$ in Eq. (5), where λ_1, λ_2, and λ_3 are all set to 1. The threshold τ in Eq. (2) is set to 0.1. Following the common practice of learning rate decaying schedule, the initial learning rate is set to 0.001 and multiplied by 0.1 every 60 epoches.

4.2 Experiment and Analysis

Effectiveness of Region Attending Network. We first visualize some results of our CAM-based region attending network in Fig. 4 for qualitative analysis. We can see that the discriminative regions of the input images are highlighted. The attended regions that are cropped from the raw images and then amplified p-reserve the object-level structure, eliminate background interference and enrich local visual details. We evaluate the effectiveness of RAN in terms of the single scale classification accuracy. In Table 2, we compare the result of our PSN *without* part modeling branch with that of two other attention based approaches. To be fair, we select the *single-attention* based performance from FCAN [23], and *the second scale* result from RA-CNN [9]. RA-CNN[9] is the most relevant work to ours considering the region attention concept and the way to use it. We can see that our method outperforms FCAN [23] with a clear margin (7.1% relative gain) and RA-CNN[9] with 1.3% accuracy improvement

Effectiveness of part branch in PSN. As pointed out in Section 3.3, we take a bi-directional LSTM as the mapping function to map the serialized features to discriminative part representation.To evaluate its effectiveness, we present the classification results of two models: *PSN wo/w part*, which are the classification accuracy based on the P_g *without/ with* the part modeling branch. The results are shown in Table 2. We can see that the model adopting part sequence modeling branch achieves a relative performance gain of 1.0%. Some illustrations are given in Figure 5. It can be observed that the part branch encourages a more compact distribution on the feature maps, further enhances the crucial part areas and depresses the unless ones.

Fine-grained Categorization. We compare our REAPS framework with several existing methods and the results are summarized in Table 3. Our method achieves better performance than those [4, 18, 30–32] using ground-truth bounding boxes or part annotations during training or testing time on three datasets. Compared with BCNN-based methods [8, 20, 25] our method obtains comparable or even better performance due to accurate attention localization and part

Table 2. Performance comparison of attention localization on the Stanford Cars dataset.

Approach	Accuracy
FCAN (single-attention) [23]	84.2
RA-CNN (scale 2) [9]	90.0
PSN *wo part*	91.3
PSN	92.3

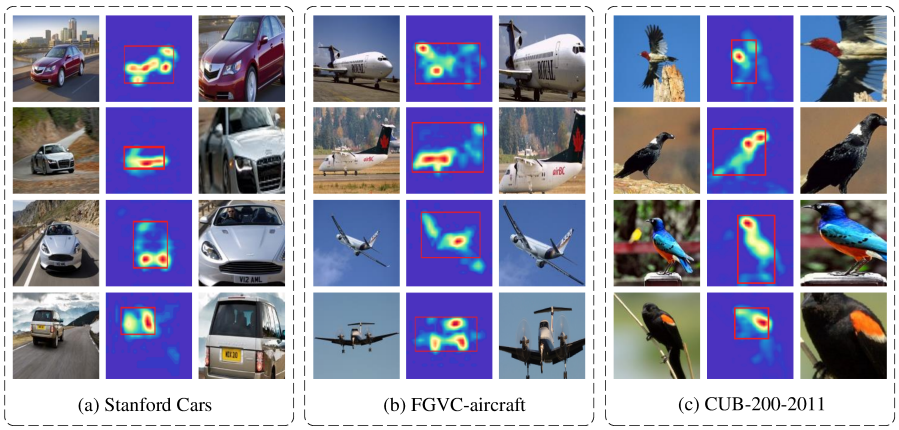

Fig. 4. Region attention localization results of RAN for some examples from (a) Stanford Cars, (b) FGVC-aircraft, and (c) CUB-200-2011. Pictures from left to right in (a-c) are the raw image, the attention mask with bounding box indicating the area of top attention response, and the cropped object-level region respectively.

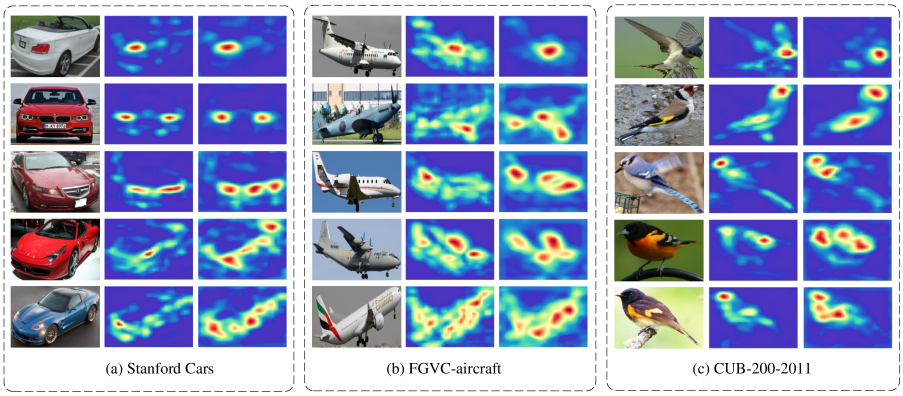

Fig. 5. Visualization of feature maps for some examples from (a) Stanford Cars, (b) FGVC Aircraft, (c) CUB-200-2011. Pictures from left to right in (a-c) are the raw image, the feature map generated by PSN *without* part branch and the feature map generated by PSN *with* part branch, respectively.

sequence modeling. To further enhance the capacity of REAPS, as RA-CNN [9] does, we incorporate one more PSN into our framework, and the 2nd PSN is based on the attended region of the 1st PSN. We call the resulting network REAPS+. Note that, REAPS+ obtains the best performance on three datasets. Especially on FGVC Aircraft dataset, our proposed REAPS+ obtains the best accuracy of 92.6%, surpassing state-of-the-art MA-CNN [39] by a relative 2.7% gain. The significant improvement suggests that the proposed part sequence modeling network works as expected to leverage spatial information of parts, and does even better when the objects to be recognized have strong sequential structures.

Table 3. Performance comparison on the Stanford Cars, FGVC Aircraft and CUB-200-2011 datasets. *(*)* indicates whether bounding box or part annotation is used in training.

Approach	Stanford Cars	FGVC Aircraft	CUB200-2011
PA-CNN [18]	92.8 (*)	–	82.8 (*)
MDTP [31]	92.5 (*)	88.4 (*)	–
MG-CNN [30]	–	86.6 (*)	83.0 (*)
PN-CNN [4]	–	–	85.4 (*)
Mask-CNN [32]	–	–	85.4 (*)
STNs [15]	–	–	84.1
FCAN [23]	91.5	–	84.3
PDFR [38]	–	–	84.5
Improved B-CNN [20]	92.0	88.5	85.8
BoostCNN [25]	92.1	88.5	86.2
KP [8]	92.4	86.9	86.2
RA-CNN(scale 1+2+3) [9]	92.5	–	85.3
MA-CNN [39]	92.8	89.9	**86.5**
REAPS *wo PSN*	92.0	89.8	81.3
REAPS	**93.1**	**91.8**	86.0
REAPS+	**93.5**	**92.6**	**86.8**

5 Conclusion

In this paper, we propose a novel framework REAPS for fine-grained recognition, which consists of a region attending network and a part sequence-learning network. The proposed framework does not need bounding box/ part annotations for training and can be trained end-to-end. We conduct extensive experiments on three fine-grained benchmark datasets, and the experimental results show that REAPS outperforms the existing methods.

References

1. Angelova, A., Zhu, S.: Efficient Object Detection and Segmentation for Fine-Grained Recognition. In: CVPR. pp. 811–818 (2013)
2. Berg, T., Belhumeur, P.N.: POOF: Part-Based One-vs.-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation. In: CVPR. pp. 955–962 (2013)
3. Berg, T., Liu, J., Lee, S.W., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds. In: CVPR. pp. 2019–2026 (2014)
4. Branson, S., Horn, G.V., Belongie, S.J., Perona, P.: Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. CoRR **abs/1406.2952** (2014)
5. Cai, S., Zuo, W., Zhang, L.: Higher-Order Integration of Hierarchical Convolutional Activations for Fine-Grained Visual Categorization. In: CVPR. pp. 511–520 (2017)
6. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In: EMNLP. pp. 103–111 (2014)
7. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: EMNLP. pp. 1724–1734 (2014)
8. Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y., Belongie, S.J.: Kernel Pooling for Convolutional Neural Networks. In: CVPR. pp. 3049–3058 (2017)
9. Fu, J., Zheng, H., Mei, T.: Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In: CVPR. pp. 4476–4484 (2017)
10. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact Bilinear Pooling. In: CVPR. pp. 317–326 (2016)
11. Gavves, E., Fernando, B., Snoek, C.G.M., Smeulders, A.W.M., Tuytelaars, T.: Fine-Grained Categorization by Alignments. In: ICCV. pp. 1713–1720 (2013)
12. He, X., Peng, Y., Zhao, J.: Fast Fine-Grained Image Classification via Weakly Supervised Discriminative Localization. IEEE Trans. Circuits Syst. Video Techn. **29**(5), 1394–1407 (2019)
13. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997)
14. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-Stacked CNN for Fine-Grained Visual Categorization. In: CVPR. pp. 1173–1182 (2016)
15. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial Transformer Networks. In: NIPS. pp. 2017–2025 (2015)
16. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: CVPR Workshop on FGVC. vol. 2 (2011)
17. Kong, S., Fowlkes, C.C.: Low-Rank Bilinear Pooling for Fine-Grained Classification. In: CVPR. pp. 7025–7034 (2017)
18. Krause, J., Jin, H., Yang, J., Li, F.: Fine-grained recognition without part annotations. In: CVPR. pp. 5546–5555 (2015)
19. Lin, D., Shen, X., Lu, C., Jia, J.: Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: CVPR. pp. 1666–1674 (2015)
20. Lin, T., Maji, S.: Improved Bilinear Pooling with CNNs. In: BMVC (2017)
21. Lin, T., Roy Chowdhury, A., Maji, S.: Bilinear CNN Models for Fine-Grained Visual Recognition. In: ICCV. pp. 1449–1457 (2015)

22. Liu, J., Kanazawa, A., Jacobs, D.W., Belhumeur, P.N.: Dog Breed Classification Using Part Localization. In: ECCV. pp. 172–185 (2012)
23. Liu, X., Xia, T., Wang, J., Lin, Y.: Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition. CoRR **abs/1603.06765** (2016)
24. Maji, S., Rahtu, E., Kannala, J., Blaschko, M.B., Vedaldi, A.: Fine-Grained Visual Classification of Aircraft. HAL - INRIA (2013)
25. Moghimi, M., Belongie, S.J., Saberian, M.J., Yang, J., Vasconcelos, N., Li, L.: Boosted Convolutional Neural Networks. In: BMVC (2016)
26. Nilsback, M., Zisserman, A.: Automated Flower Classification over a Large Number of Classes. In: Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India. pp. 722–729 (2008)
27. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: ECCV. pp. 143–156 (2010)
28. Qian, Q., Jin, R., Zhu, S., Lin, Y.: Fine-grained visual categorization via multi-stage metric learning. In: CVPR. pp. 3716–3724 (2015)
29. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, 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. pp. 2399–2406 (2015)
31. Wang, Y., Choi, J., Morariu, V.I., Davis, L.S.: Mining Discriminative Triplets of Patches for Fine-Grained Classification. In: CVPR. pp. 1163–1172 (2016)
32. Wei, X., Xie, C., Wu, J., Shen, C.: Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition **76**, 704–714 (2018)
33. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: CVPR. pp. 842–850 (2015)
34. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: ICML. pp. 2048–2057 (2015)
35. Zhang, N., Farrell, R., Darrell, T.: Pose pooling kernels for sub-category recognition. In: CVPR. pp. 3665–3672 (2012)
36. Zhang, N., Farrell, R., Iandola, F.N., Darrell, T.: Deformable Part Descriptors for Fine-Grained Recognition and Attribute Prediction. In: ICCV. pp. 729–736 (2013)
37. Zhang, X., Zhou, F., Lin, Y., Zhang, S.: Embedding Label Structures for Fine-Grained Feature Representation. In: CVPR. pp. 1114–1123 (2016)
38. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking Deep Filter Responses for Fine-Grained Image Recognition. In: CVPR. pp. 1134–1142 (2016)
39. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In: ICCV. pp. 5219–5227 (2017)
40. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. In: CVPR. pp. 2921–2929 (2016)