2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

A Multi-part Convolutional Attention Network for Fine-Grained Image Recognition

Weilin Zhong, Linfeng Jiang, Tao Zhang, Jinsheng Ji, Huilin Xiong
School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, China
Institute for Sensing and Navigation, Shanghai Jiao Tong University, China
{zhongweilin, fine0228, sjtu--zt, jinshengji, hlxiong}@sjtu.edu.cn

***Abstract*—The goal of fine-grained image recognition is to recognize hundreds of sub-categories affiliating to the same basic-level category (*e.g.*, bird species). It is a highly challenging task due to the large intra-class variance and small inter-class variance. Existing approaches deal with the subtle difference among object classes via learning and localizing discriminative parts. However, most of the part localization methods follow a step-to-step manner that first localizes larger parts and then generates smaller parts from the larger ones, which is not efficient. In this paper, we present a Multi-part Convolutional Attention Network (M-CAN), which simultaneously focuses on the discriminative image parts at multiple scales. In specific, a convolutional attention based part localization network is presented to localize multi-scale parts from different layers of the deep Convolutional Neural Networks (CNN). Importantly, our part localization network requires no part annotations but only the image labels, which avoids the heavy labor of complex part labeling. We conduct comprehensive experiments and the experimental results show that, our method outperforms the state-of-the-art approaches on three challenging fine-grained datasets, including CUB-Birds, Stanford-Dogs and Stanford-Cars.**

I. Introduction

Fine-grained image recognition aims to distinguish sub-ordinate categories belonging to the same species. The task is very challenging due to the small variance in object appearances. Thus, the discriminative features for recognizing similar classes are typically extracted from the subtle differences, which are usually positioned at some regions of the object, such as the heads, wings and tails in bird species [1]. Conventional methods mostly solve this problem by utilizing the manually labeled parts [2], [3], [4]. There are several drawbacks in depending on human pre-defined parts: 1) acquiring precise parts is a labor intensive work. 2) not all semantic parts are beneficial and indispensable for discriminating sub-categories but may degrate recognition accuracy [5]. Hence, it is desirable to develop part localization methods based on weakly supervised learning, which only requires the image labels.

Recent years have witnessed the outstanding performance of deep Convolutional Neural Networks (CNN) in various computer vision tasks including fine-grained image recognition [6], [7], [8], [9], [10], [11]. Particularly, CNN based methods have made impressive progress in learning discriminative features for fine-grained recognition, with no dependencies on bounding box or part annotations [7], [12], [13], [10], [14].

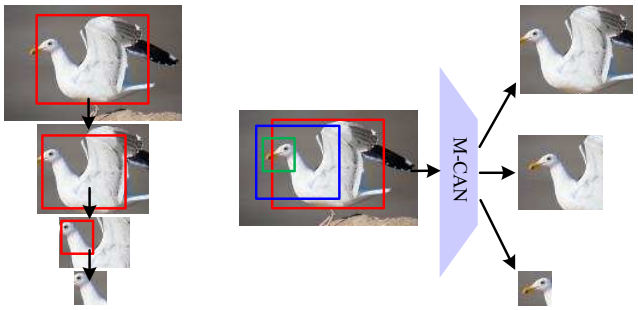

Fig. 1. Illustration of the discriminative part localization process in previous methods and our proposed method. (a) The part localization process of previous methods. The finer scale parts are conditioned on the coarse scale parts. (b) The part localization process of the proposed M-CAN. Our method simultaneously localizes multi-scale image parts in one attention network.

This significantly increases the usability and scalability of fine-grained recognition applied in real-world scenarios. Some of those methods directly extract feature representations from the pre-trained deep network. Some of them first localize the object parts, and then re-train the deep network using the localized parts.

Among them, the methods based on part localization have attracted increasing attention due to the good explanation and easy implementation [15], [16], [10]. As shown in Figure 1(a), they usually localize the larger regions, and then, select smaller parts over previous larger ones [13], [17]. The coarse-fine part localization methods significantly improve the performance of fine-grained recognition. However, if some highly discriminative parts are lost in the larger part localization process, it is hard to recover in the subsequent smaller part selection. Besides, the step-to-step part localization method usually lacks efficiency.

To tackle these challenges mentioned above, in this paper, we present a multi-part localization network to simultaneously identify the highly discriminative parts at different scales. Our part localization process is as shown in Figure 1(b). Intuitively, simulating human visual systems that usually focus on relevant regions when comparing different objects, an attention network, called Multi-part Convolutional Attention Network (M-CAN), is presented to localize the discriminative parts. Specifically, our M-CAN simultaneously localizes multi-

978-1-5386-3787-6/18/$31.00 ©2018 IEEE

scale parts from different convolutional layers of the deep network. Note that, the corresponding region size in the input images of the CNN unit, called receptive field, is larger in the deeper layers. Hence, we localize the larger parts from top layers of the deep network.

Furthermore, a lot of background noises will be included in the larger parts if we directly localize the discriminative parts from the original images. Thus, we incorporate a object detection network to localize the image objects in advance, and then, perform part localization. Extensive experiments are conducted on three challenging fine-grained datasets, including CUB-Birds, Stanford-Dogs and Stanford-Cars, and the experimental results demonstrate the effectiveness of our proposed method.

II. Related Work

The research on fine-grained image recognition can be grouped into two categories, namely fine-grained feature learning and part localization. A large number of methods have been developed to learn representative features for fine-grained image recognition. In particular, with the great success of deep learning, most of the recognition frameworks are based on the powerful feature learning ability of the deep Convolutional Neural Networks (CNN). Branson et al. [3] propose a pose normalized deep convolutional network which integrates the features from lower-level layers and higher-level layers. Lin et al. [8] propose a bilinear structure to capture the image local differences by combining the convolutional features from two models in a translation invariant manner. Qi et al. [18] propose a multi-stage distance metric to project the images into a distance space, where the distance of the images from the same class are pulled close, and meanwhile, the distance of images from different classes are pushed far apart. In terms of large-scale fine-grained car classification, Wang et al. [16] propose a object-centric sampling scheme that implicitly incorporates the object detection into the representation learning process.

As for the part localization, there are numerous studies focusing on localizing discriminative parts using weakly supervised approaches, which means they only need image labels. Xiao et al. [19] propose a two-level attention model to localize the objects and parts, where the part templates are achieved by clustering the internal hidden representations of CNN. Max et al. [10] propose a spatial transformer network which automatically detects the discriminative parts for fine-grained visual categorization. Simon et al. [20] propose to learn part features in a completely unsupervised manner, by finding constellations of neural activation patterns. Zhao et al. [12] propose a diversified visual attention network to localize the objects or parts from coarse to fine granularity. Pierre et al. [17] utilize a recurrent neural attention model to recurrently localize the local parts. Zhang et al. [9] propose to learn multiple part detectors by analyzing neural activations computed from the deep CNN. Liu et al. [15] propose a Fully Convolutional Attention Networks (FCANs) that optimally glimpses local discriminative regions using a reinforcement learning framework.

In this paper, we present a multi-part based attention network for fine-grained image recognition that simultaneously localizes multi-scale object parts in a unified framework with high efficiency. The most relevant methods to ours come from [17], [12], [15]. All of them learn candidate parts using a attention network. Different from [17] and [12], our method explores the multi-level spatial contextual information from different layers. Similar to FCANs [15], our attention model is based on convolutional network, which is a fast and effective approach to produce dense prediction. However, compared to FCANs [15] that learn the discriminative parts with complex greedy reward strategy, our part localization network is trained by using the simple image category classification loss.

III. Proposed Model Architecture

Figure 2 illustrates the overall network architecture. There are three streams in our proposed method, that is, object detection stream, multi-part localization stream and part-specific feature learning stream. First, the object detection network that takes the full-size images as input is incorporated to localize the image objects (Sec. III-A). Second, the detected objects are sent to the part localization attention network to localize discriminative parts at multiple resolutions (Sec. III-B). Finally, all the localized parts are re-sacled to a larger resolution, and then they are forward to the part-specific feature learning network (Sec. III-C).

A. Object Detection Network

Previous object detection methods typically need the ground-truth object bounding box to train a object detection model. Here, our object detection network only depends on the image labels. Specifically, we generate the image objects based on the convolutional feature maps computed from the deep CNN. Inspired by CAM [21], we remove the last three fully-connected layers and the last max-pooling layer in VGGNet [22], and then, replace them with a Global Average Pooling (GAP) layer. The pooled features are followed by a fully-connected softmax layer to output the class confidence score.

For the $k\text{-}th$ unit of the fully-connected softmax layer, let $f_k(x, y)$ represent its activation in the last convolutional layer at spatial location (x, y). Then the Class Activation Map (CAM) for class c can be defined as follows:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y), \tag{1}$$

where $M_c(x, y)$ indicates the importance of the activation at spatial location (x, y) for classifying an image to class c, and w_k^c denotes the weight used to generate the CAM by summing the activation $f_k(x, y)$.

After upsampling the CAM to the size of the original images, the most discriminative image regions relevant to the particular category can be identified. For the task of object detection, we need to generate a bounding box associated to the image category. Hence, we first threshold the CAM, and then, take the bounding box that covers the largest connected area as the results of object detection.

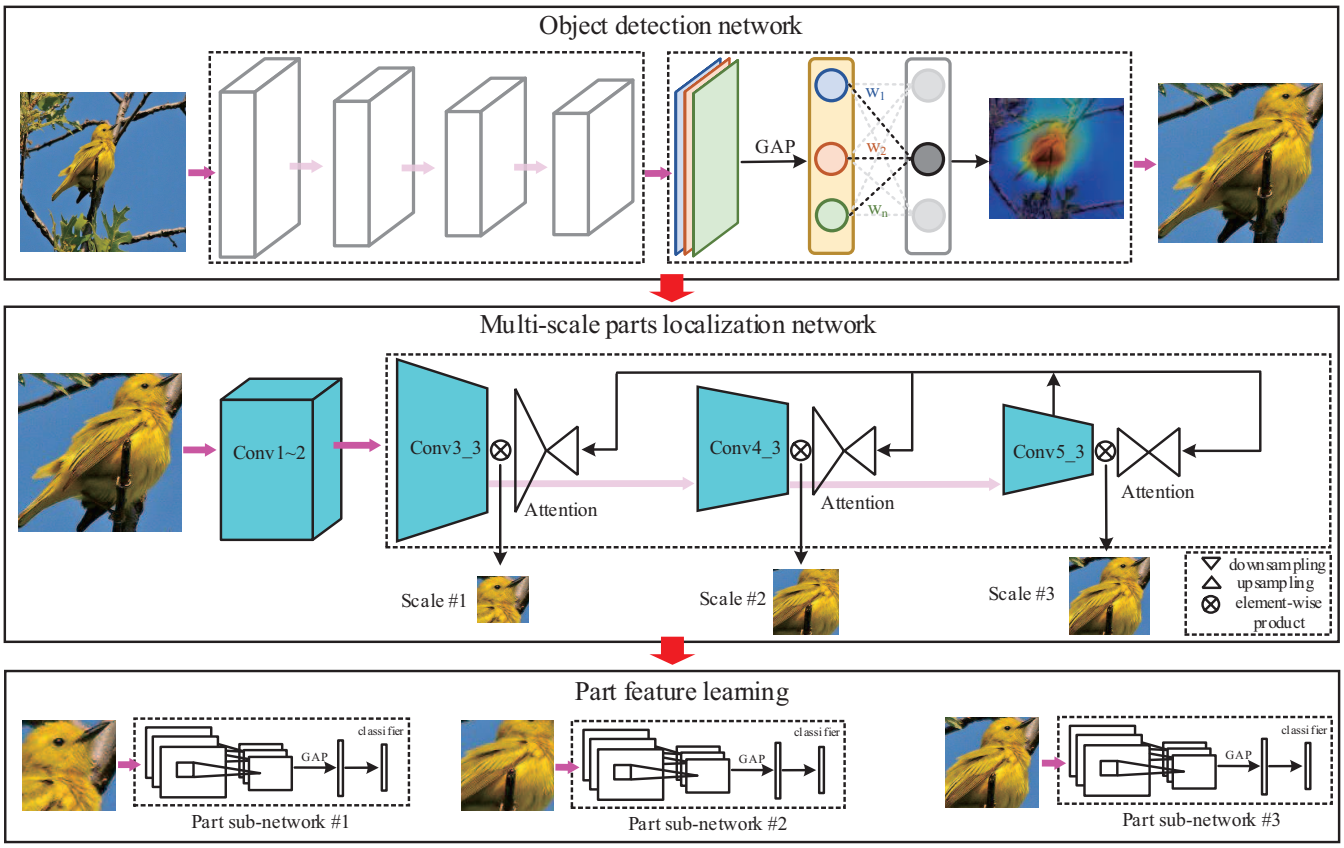

Fig. 2. The sketch of the presented method that contains three stream networks, namely objection detection network, multi-scale parts localization network and part feature learning network. The original images are first sent to the objection detection network to roughly detect the objects. Then, the multi-scale attention network is proposed to localize multiple discriminative parts on the detected objects. Finally, each localized part is taken as input by the part feature learning network.

B. Multi-part Attention Network

As shown in the second stream in Figure 2, the attention network simultaneously localizes multiple parts from different convolutional layers of the deep CNN. The critical components of our part-localization are illustrated in Figure 3.

We adopt a `conv-deconv` module to generate the attention map for the input feature maps. In specific, our part localizer first downsamples the `conv5_3` features from the VGGNet using a convolutional layer whose kernel size is $3 \times 3 \times 20$ and stride is 2. And then, we use the deconvolution operation [23] followed by one ReLU layer to generate the single channel attention map. The size of the attention map is identical to the size of input feature maps. On the attention map, the spatial location with the highest value is selected as the part location. Each location from different layers corresponds to different sizes of the input image, called receptive field. We localize the discriminative parts from different layers, and thus, we can simultaneously generate multiple parts with different resolutions. In the part localization process on different layers, the downsampling `conv` weights are shared, which can in some extent avoid learning arbitrary attention maps.

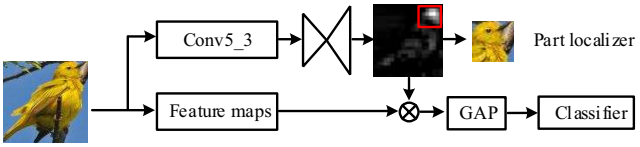

Fig. 3. Demonstration of the attention localization network. The part localizer generates an attention map, which is used for weighting the input feature maps. The weighted feature maps are pooled by the GAP operation, and then, the pooled features are sent to the image category classifier.

The part localization process can be expressed as:

$$h_l = \text{GAP}(M_l \odot X_l), l \in \{1, \ldots, L\}, \tag{2}$$

where M is the generated attention map, X is the input feature maps and h is the pooled feature. Moreover, $\odot$ denotes the element-wise product and L means the number of layers taken to the part localizer.

In this paper, we adopt the soft-max classification loss to train the localization network. Suppose we sample n images into a training mini-batch. For each image, the pooled feature h_l is forward to a softmax classifier that outputs the class probability $\hat{p}_l$. Then, we aim to minimize the following cross-

TABLE I
THE STATISTICS OF THREE FINE-GRAINED DATASETS.

Datasets	Category	Train	Test	Bounding box	Part annotation
CUB-200-2011 [1]	200	5,994	5,794	✓	✓
Stanford Dogs [24]	120	12,000	8,580	✓	
Stanford Cars [25]	196	8,144	8,041	✓	

entropy in the localization network:

$$\mathcal{L}_{cla} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L} y^i \log \hat{p}_l^i, \tag{3}$$

where y^i is the ground truth image label.

C. Part Feature Learning

In the part localization network, the discriminative parts and the object based features are simultaneously learned. It may be still difficult to distinguish the categories with subtle visual differences only depending on the object-level features. In previous methods [15], [7], [4], the effectiveness of region amplification is broadly validated in further boosting the performance of fine-grained image recognition. Following the same strategy, we first amplify the localized parts to larger resolutions, and then, extract part features from the amplified parts.

Our part feature learning process is shown in the third stream in Figure 2. First, we crop the localized parts from the detected object. These cropped parts are then amplified to a larger resolution. Each amplified part is taken as the input of a part sub-network. The part sub-network aims to classify the part into image-level categories. The structure of the part sub-network is same as the object detection network, in which the pooled features are sent to the soft-max classifier.

In the testing phase, as in [15], [10], the final classification result for each testing image is the average of all the classification results from three streams, namely the object detection stream and part localization stream as well as part feature learning stream.

IV. EXPERIMENT RESULTS

We conduct experiments on three benchmark datasets, including CUB-200-2011 [1], Stanford Dogs [24], Stanford Cars [25]. The statistics of the three datasets are shown in Table I.

A. Implementation Details

Our method is implemented based on the open source PyTorch[1] library. The CNN architecture of the three streams in our method are all tailored from the VGG-16 [22], in which the fully-connected layers are removed. For the object detection network, the input image size is 227×227 and the size of the output feature maps is $14 \times 14 \times 512$. It is expected that the detected bounding box can cover the object. Hence, unlike the implementation in [21], we threshold the CAM with 60% of the max value in CAM. In the part localization network, the receptive fields in the last layers are much larger than those of previous layers, which may be too large to model an object part. For example, given 227×227 input image, the receptive field of `conv5_3` is 212. Hence, we localize the discriminative parts from another two layers, namely `conv4_3` and `conv3_3`, whose receptive fields are 100 and 44, respectively. The size of our localized parts are identical to that of the receptive fields. In the part feature learning module, we first upsample the localized parts to a larger resolution of 227×227 and forward them to the truncated VGG-16 network.

[1]http://pytorch.org/

TABLE II
RECOGNITION PERFORMANCE COMPARISON ON CUB-200-2011.

Methods	Annotation	Accuracy(%)	CNN Features
CAM [21]	BBox	70.5	GoogLeNet
PoseNorm [3]	BBox+Parts	75.7	AlexNet
Part-RCNN [4]	BBox+Parts	76.4	AlexNet
PA-CNN [5]	BBox	82.8	VGGNet
FCANs [15]	BBox	84.2	GoogLeNet
B-CNN [8]	BBox	**85.1**	VGGNet
CAM [21]	-	67.8	GoogLeNet
TLAN [19]	-	77.9	VGGNet
DVAN [12]	-	79.0	VGGNet
NAC [20]	-	81.0	VGGNet
FCANs [15]	-	82.0	GoogLeNet
PDFR [9]	-	82.6	VGGNet
STN [10]	-	84.1	GoogLeNet
B-CNN [8]	-	84.1	VGGNet
Ours	-	**84.8**	VGGNet

TABLE III
RECOGNITION PERFORMANCE COMPARISON ON STANFORD DOGS WITHOUT EXTRA BOUNDING BOX OR PART ANNOTATIONS.

Methods	Accuracy(%)	CNN Features
NAC [20]	68.6	AlexNet
PDFR [9]	72.0	AlexNet
Ours	**72.6**	AlexNet
VGG-16 baseline [22]	76.2	VGGNet
DVAN [12]	81.5	VGGNet
FCANs [15]	84.2	VGGNet
Ours	**85.3**	VGGNet

B. Comparison with State-of-the-Art Methods

We compare our method with recent state-of-the-art approaches. All the methods can be grouped into two categories depending on whether they use human-defined object bounding box (BBox) or part annotations.

Experiments on CUB-200-2011. The performance comparison on CUB-200-2011 dataset is summarized in Table II. The column of "CNN Features" means which CNN model is adopted as the backbone network. We observe that our method obtains comparative results with the methods using human-defined objects or parts bounding box. Specifically, B-CNN [8] and FCANs [15] achieve 85.1% and 84.2% recognition accuracy, respectively. Even though, our method under the weakly supervised setting achieves 84.8% recognition accuracy which is slightly lower than the methods

TABLE IV
RECOGNITION PERFORMANCE COMPARISON ON STANFORD CARS.

Methods	Annotation	Accuracy(%)	CNN Features
FCANs [15]	BBox	89.1	GoogLeNet
PA-CNN [5]	BBox	**92.8**	VGGNet
DVAN [12]	-	87.1	VGGNet
FCANs [15]	-	89.1	GoogLeNet
B-CNN [8]	-	91.3	VGGNet
Ours	-	91.5	VGGNet

using objects or parts annotations. Compared with the weakly supervised methods, such as PDFR [9], STN [10], and B-CNN [8], our approach achieves improvement by at least 0.7%. FCANs [15] and DVAN [12] are two most related methods to ours, which are also based on spatial attention model. As seen from Table II, our method achieves 0.6%, 2.8% and 5.8% relative improvement compared with FCANs [15] (with part annotations), FCANs [15] (without part annotations) and DVAN [12], respectively.

Experiments on Stanford Dogs. The classification results on Stanford Dogs dataset are illustrated in Table III. On this dataset, we implement our method based on two types of CNN features, namely AlexNet [26] and VGGNet [22]. Note that, for AlexNet, the input images are re-scaled to 300×300 and we localize the parts from the last max-pooling layer and two convolutional layers, generating three parts with size of 195, 163, and 131, respectively. It can be observed that, among the methods based on AlexNet, our approach achieves the best recognition accuracy. Moreover, as seen from Table III, the performance of our method can be further boosted by using the deeper network VGGNet. Among the methods using VGGNet, our approach achieves 1.1%, 3.8% and 9.1% relative improvement compared with FCANs [15], DVAN [12] and VGG-16 baseline, respectively.

Experiments on Stanford Cars. The recognition results on Stanford Cars are shown in Table IV. It can be observed that PA-CNN [5], which learns to detect the discriminative parts from the manually segmented objects, obtains the best performance on this dataset. Although the performance of PA-CNN [5] is better than our method, the advantage of our method is that it is free of labor intensive object or part annotations. Compared with the weakly supervised methods DVAN [12] and FCANs [15], our method achieves better recognition accuracy, and meanwhile, our method simultaneously localizes the discriminative parts which is more efficient.

C. Analysis of Proposed Model

Here, we perform a detailed analysis of our method on CUB-200-2011 dataset. The results are summarized in Table V. It can be observed that the fine-tuned VGG-16 model takes the original images as input and obtains 76.1% recognition accuracy. In the part localization network, the third-scale M-CAN from `conv5_3` achieves 79.2% recognition accuracy, outperforming the VGG-16 baseline with 3.1% improvement. We find that the performance of our M-CAN can be further enhanced by taking the amplified parts into the part feature learning network. In specific, the combination of single scale M-CAN and part feature learning achieves 82.9% recognition accuracy. When we combine the first-scale and second-scale localized parts from `conv3_3` and `conv4_3`, the recognition accuracy is promoted to 83.2% and 83.7%, respectively.

TABLE V
RECOGNITION RESULTS ON THE THE CUB-200-2011 DATASET WITH DIFFERENT SETTINGS.

Methods	Object localization	Accuracy(%)
VGG-16 baseline	-	76.1
M-CAN (scale 3)	-	79.2
M-CAN (scale 3) + Part feature	-	82.9
M-CAN (scale 2+3) + Part feature	-	83.2
M-CAN (scale 1+2+3) + Part feature	-	83.7
VGG-16 baseline	✓	78.3
M-CAN (scale 3)	✓	81.1
M-CAN (scale 3) + Part feature	✓	83.6
M-CAN (scale 2+3) + Part feature	✓	84.5
M-CAN (scale 1+2+3) + Part feature	✓	84.8

Fig. 4. Five examples of the bird objects detected by the object detection network. Compared with the original images, the detected regions with less background are more discriminative to the corresponding categories.

In our method, we also incorporate a object detection network to localize the image objects in advance so as to avoid too much background noise in the subsequent part localization stream. As shown in Table V, we can observe that both the VGG-16 baseline and our M-CAN can obtain better recognition accuracy after object detection. Also, the recognition accuracy of the third-scale M-CAN is raised to 81.1%. The combination of object detection, M-CAN and part feature learning can achieve the best performance, with 84.8% recognition accuracy on CUB-200-2011 dataset.

We visualize the detected objects in Figure 4. It can be observed that most of detected regions can contain the image objects and get rid of irrelevant background. We also find that some of the detected regions only contain a large portion of the objects, and meanwhile, ignore some less discriminative parts such as the tails.

Figure 5 shows the visualization of the localized parts from our M-CAN. We can observe that our M-CAN learns to focus on the discriminative regions in the image such as the heads and bodies. Although our part localization method is not in a coarse-to-fine manner, we find that the smaller localized parts are usually in the range of the larger parts. Moreover, the first-scale parts with size of 44 only contain a small portion of the image objects, which are mainly the heads in the birds dataset.

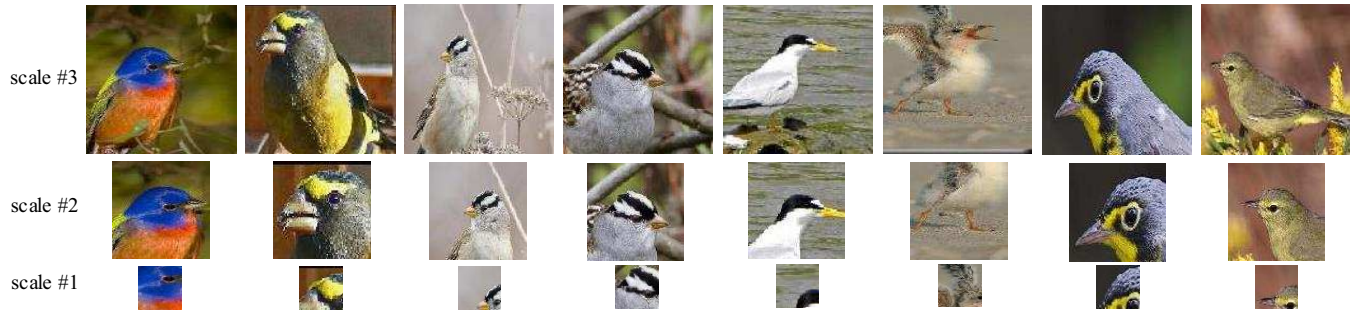

Fig. 5. Examples of the localized parts from the Multi-part Convolutional Attention Network (M-CAN) in CUB-200-2011 birds dataset. The three scale parts are with size of 212, 100 and 44, respectively. It can be seen that most of the part regions contain a portion of the image objects, which are beneficial for image recognition.

V. Conclusion

In this paper, we present a multi-scale attention network for fine-grained recognition, which simultaneously localizes the discriminative parts at multiple scales. Our part localization method is free of human defined object or part annotations and can be trained from end to end. We conduct experiments on three fine-grained recognition datasets and the experiment results demonstrate the effectiveness of our proposed method. In the future, we will explore the joint training of part localization and part feature learning so as to further increase the training efficiency and boost the performance of fine-grained recognition.

Acknowledgment

This work is partially supported by the National Science Foundation of China under Grant 61673274.

References

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds200-2011 dataset," 2011. 1, 4

[2] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. CVPR*, 2016, pp. 1124–1133. 1

[3] S. Branson, G. V. Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," *arXiv preprint arXiv:1406.2952*, 2014. 1, 2, 4

[4] Z. Ning, D. Jeff, G. Ross, and D. Trevor, "Part-based r-cnns for fine-grained category detection," in *Proc. ICML*, 2014, pp. 834–849. 1, 4

[5] J. Krause, H. Jin, J. Yang, and F. F. Li, "Fine-grained recognition without part annotations," in *Proc. CVPR*, 2015, pp. 5546–5555. 1, 4, 5

[6] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as hsnet search for informative image parts," in *Proc. CVPR*, 2017, pp. 6497–6506. 1

[7] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. CVPR*, 2017, pp. 4476–4484. 1, 4

[8] T. Y. Lin, A. Roychowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. ICCV*, 2015, pp. 1449–1457. 1, 2, 4, 5

[9] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proc. CVPR*, 2016, pp. 1134–1142. 1, 2, 4, 5

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025. 1, 2, 4, 5

[11] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked cnn for fine-grained visual categorization," in *Proc. CVPR*, 2016, pp. 1173–1182. 1

[12] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017. 1, 2, 4, 5

[13] J. Fang, Y. Zhou, Y. Yu, and S. Du, "Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture," *Trans. Intell. Transp. Syst.*, vol. PP, no. 99, pp. 1–11, 2017. 1

[14] Z. Heliang, F. Jianlong, M. Tao, and L. Jiebo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. ICCV*, 2017. 1

[15] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition," *arXiv preprint arXiv:1603.06765*, 2016. 1, 2, 4, 5

[16] X. Wang, T. Yang, G. Chen, and Y. Lin, "Object-centric sampling for fine-grained image classification," *arXiv preprint arXiv:1412.3161*, 2014. 1, 2

[17] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," *arXiv preprint arXiv:1412.7054*, 2014. 1, 2

[18] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *Proc. CVPR*, 2015, pp. 3716–3724. 2

[19] 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," in *Proc. CVPR*, 2015, pp. 842–850. 2, 4

[20] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proc. ICCV*, 2015, pp. 1143–1151. 2, 4

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2015, pp. 2921–2929. 2, 4

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 2, 4, 5

[23] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. CVPR*, 2010, pp. 2528–2535. 3

[24] A. Khosla, J. Nityananda, Y. Bangpeng, and F. F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop*, 2011. 4

[25] J. Krause, M. Stark, D. Jia, and F. F. Li, "3d object representations for fine-grained categorization," in *Proc. CVPR Workshop*, 2014, pp. 554–561. 4

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105. 5