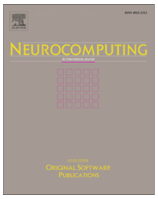
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Multi-level dictionary learning for fine-grained images categorization with attention model

Jinsheng Ji [a], Yiyou Guo [b], Zhen Yang [c], Tao Zhang [d], Xiankai Lu [e],*

[a] Shanghai Key Lab. of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, 200240 Shanghai, China
[b] College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China
[c] School of Communication and Electronics, Jiangxi Science and Technology Normal University, Nanchang 330013, China
[d] Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China
[e] School of Software, Shandong University, Ji'nan 250101, China

ARTICLE INFO

Article history:
Received 11 April 2020
Revised 16 June 2020
Accepted 9 July 2020
Available online xxxx

Keywords:
Fine-grained
Visual attention
Multi-level
Pre-trained
Dictionary learning

ABSTRACT

Fine-grained image categorization is a challenging task due to the difficulty of localizing the discriminative regions for different sub-categories. Previous works mainly focus on using the manual annotations or the attention algorithm to localize these regions, which is demanding and complex in practical applications. This paper proposes a method of using a multi-level attention model (MLA-CNN) which has been trained on the full-size image train set of current tasks to localize the most discriminative regions. Intuitively, three typical receptive field sizes are selected for the multi-level attention maps. Then, multi-level dictionary learning is introduced to extract discriminative features from these localized regions. Our method explores a new thought about how to use the neural activations to generate multi-scale regions which are helpful for the fine-grained categorization. The method can be achieved in two steps. The first step is to select the neurons that have the max activation in the selected three feature maps. These feature maps are the outputs of the pre-trained CNN model by feeding the full-size images into the model. Then, we generate the discriminative regions according to the receptive field size of the selected neurons. The second step is to train the subtle networks with these multi-scale regions. One scaled discriminative region can be regarded as one typical dictionary feature. Then these results are integrated for final prediction. We evaluate our method on three challenging fine-grained image datasets, CUB-200-2011, Stanford Dogs, and Stanford Cars. The experimental results demonstrate that our method outperforms many state-of-the-art methods, using extra object/parts annotations and attention-based methods.

1. Introduction

Fine-grained image categorization is to classify these images with subtle distinctions which usually have many subclasses belonging to the same class like birds [1–5], dogs [6], plants [7,8], cars [9,10], etc. Different from the traditional general image categorization, fine-grained image categorization should focus on these regions that can distinguish the similar subclasses. As shown in Fig. 1, it may be difficult to classify the two classes just using the learned feature representations from the whole image due to the similar appearance between them. Therefore, it requires us to locate these discriminative regions that can easily distinguish similar subclasses and learn more fine-grained feature representations from these regions.

Most part-based methods [11–16] use the manually defined parts to locate the discriminative regions for the task of fine-grained image categorization. These manual defined parts, such as "head", "back" and "tail", can improve the performance of the classification model because these parts capture more detailed information about the targets. Therefore, many methods explored object level or even part level annotations to improve the performance of classification model [17–19]. But, these annotations are demanding in practical applications and the human-defined regions may not be optional for all fine-grained categorization tasks [20]. Especially when dealing with the task which even the experts couldn't give the accurate locations of these discriminative regions, it is necessary to find a method that can locate the discriminative regions by the inner structure of the CNN model.

* Corresponding author.
E-mail address: carrierlxk@gmail.com (X. Lu).

https://doi.org/10.1016/j.neucom.2020.07.147
0925-2312/© 2021 Elsevier B.V. All rights reserved.

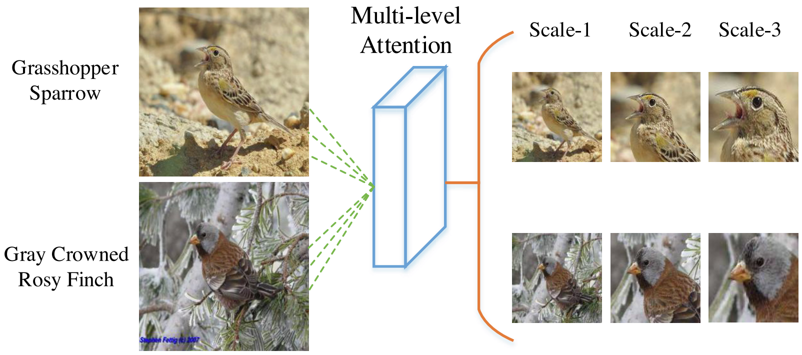


Fig. 1. Overview of the proposed module which localizes the discriminative regions by the pre-trained CNN model. We use the multi-level attention model as the neural activations generator and select the neurons which have the max activations from feature maps. According to the receptive field size of these selected neurons, we crop the multi-scale discriminative regions.

Recently, many works focus on the visual attention algorithm [21–29] to locate these discriminative regions instead of relying on the objects/parts annotations and have achieved impressive performance on these challenging fine-grained datasets.

Recently, many researchers [22,23,30,31] explored the inner structure of CNN and used the information for fine-grained image recognition. As we do not use the manual annotations of objects or parts for the input image, the challenge for our method is how to make full use of the full-size image and its label information to train a model that can classify these similar subclasses. The challenge can be divided into two aspects: how to localize these discriminative regions and how to get the fine-grained feature representations. Previous works, including methods using predefined objects/parts and relying on the visual attention algorithm to localize discriminative regions, have made impressive progress on the fine-grained image categorization. Although these methods have made promising results, there are still several limitations to localizing these regions. First, due to the limitations of acknowledge the subtle difference among these subclasses, these human-defined annotations may not be optional for all tasks of fine-grained categorization. Second, the subtle differences in the regions of the input image are sometime difficult to locate and the existing visual attention algorithm is time costing and needs an elaborate design for the certain task of fine-grained categorization. Additionally, many works also proposed the dictionary learning [32,33] for the task of feature fusion. we are also inspired by these methods to learn and fuse multi-level features from different parts of the object.

To deal with these challenges mentioned above, we propose the multi-level attention model(MLA-CNN), which has been pre-trained on the full-size image train set of the current task, to localize the most discriminative regions as shown in Fig. 1. Our method is free of the objects/parts annotations both during the training and testing stage. Specifically, our method consists of two parts: one part is the attention module for localizing the most discriminative regions from the fine-grained images; another part is the multi-level dictionary learning module which fuses features of the localized multi-scale regions. Considering the fact that a CNN model pre-trained on the ImageNet [34] usually performances better results on many categorization tasks, the proposed MLA-CNN mainly uses the model which has been trained on the train set of the fine-grained dataset to localize the most discriminative regions. Compared with previous works, this can easily get multi-scale discriminative regions for the fine-grained categorization and reduce the complexity of localizing them significantly. As these multi-scale regions cover certain parts of the image ranging from the object level to a more subtle level, a single network could not achieve the best accuracy. Thus, we feed these multi-scale discriminative regions into three subtle networks and ensemble their result together. By integrating the outputs of these subtle networks, we get higher accuracy on the fine-grained categorization.

It can be found that the proposed MLA-CNN does not rely on the manual annotations to localize the regions both during the training and testing stage. Specially, we train the CNN model with the full-size image train set in advance and use the pre-trained model as our neural activations generator. It should be noted that the regions generated by the proposed method from the original image can help the network to learn more discriminative features for the fine-grained categorization. Our contributions can be summarized as follows:

(1) We use the multi-level attention model which has been pre-trained on the full-size image train set of the fine-grained dataset to localize multi-scale discriminative regions according to these selected neural activations.
(2) We propose a method that makes full use of the receptive field size of the selected neurons. The neurons that have larger receptive field size can capture overall information while those having smaller receptive field size usually can capture more details about the image. The proposed attention module fuses the activations of multiple convolutional layers for better recognition performance.
(3) To get better performance, we feed these multi-scale discriminative regions into several subtle networks to train finer networks. By integrating the outputs of these subtle networks, the proposed method achieves higher accuracy than these subtle networks.
(4) We also conduct comprehensive experiments on these three challenging fine-grained datasets: CUB-200-2011, Stanford Dogs, and Stanford Cars. The experimental results evaluate the effectiveness of the proposed method.

The rest of this paper is organized as follows: Section 2 describes the related works on the fine-grained categorization. The details of the multi-level attention network and multi-level dictionary feature learning are elaborated in Section 3. In Section 4, we show the experimental results and comparisons with other methods. The conclusion and future work are presented in Section 5.

2. Related work

2.1. Fine-grained categorization

Many methods using the convolutional network to learn better feature representations for the task of fine-grained image categorization have been reported recently [35–37,34,38]. These methods are divided into four groups by Zhao et al. [39] according to the additional information or the human inference. Those approaches that use the convolutional network to localize the discriminative regions and those that use multiple networks and the visual attention algorithm are the most used in these previous works. Zhang et al. [40] utilize a part-based CNN module to detect the objects/-parts and achieve an impressive performance by learning the pose of these detected objects and parts.

Most previous works focus on the front part of the network and aim to get more discriminative features. Lin et al. [41] propose a bilinear architecture to compute the pairwise features through two independent CNN models. This brings us the thought that we also can do a lot of work on these features. Similarly, many methods that use the convolutional layers to localize the parts have been proposed. Liu et al. [42] use the fully convolutional network to locate these most discriminative regions. Branson et al. [12] claim that integrating the low-level layer and high-level layer can get more discriminative features. A promising direction of the fine-grained categorization is that using the neural activations to localize these regions and learning the more discriminative features.

2.2. Discriminative regions localization and dictionary learning

Some previous part-based works [43–45,30,46] use the manually defined objects/parts annotations to localize the important regions. It is demanding in practical applications because of the professional background and huge human resources. Due to the limitations of acknowledge about the task of fine-grained object classification, these annotations may not bring improvement to every fine-grained image categorization. Recently, there have been several methods that aim to localize the most discriminative regions by visual attention algorithm or the learned region proposal module. The attention technology has also been used to extract high level semantic information for the task of recognition and detection [30,22,47–54].

Generating a large number of multi-scale region proposals by object proposal and selective search, Zhang et al. [30] cluster these region proposals and then feed them into the network to learn the fine-grained feature representations. Similarly, Zhang et al. [55] pick the neural responses and get these part filters for the discriminative part localization. Zhang et al. [30] propose to learn discriminative data representations that can simultaneously shrink the off-block-diagonal components and highlight the block-diagonal representation. Liu et al. [42] employ a fully convolutional network to localize these regions, and by zooming in on these localized regions, they get finer discriminative regions which improve the performance of fine-grained image categorization significantly. To localize these discriminative regions precisely, Fu et al. [22] propose a recurrent attention convolutional network which recursively learns these regions and the region-based feature representation at multiple scales. However, it is still difficult for these methods to localize important regions succinctly due to the complexity of learning a part detector.

To better construct the inner relationship among the multiple features, many works also proposed the dictionary learning methods [33,56–60] for the task of feature fusion. Li et al. [33] propose a cross-view dictionary learning model which takes advantages of the view-consistency information, and adaptively learns pairs of dictionaries to generate robust and compact representations for pedestrian images. To better fuse the multiple features, Wang et al. [32] develop a non-linear feature fusion scheme that better combines object and motion features.

3. Proposed method

In this section, we will introduce the proposed multi-level attention and dictionary learning model for the fine-grained image categorization. Our strategy consists of two parts, one is the multi-level attention network which indicates where is an important region for the fine-grained categorization by the value of neural activation. The second is the multi-level dictionary learning model which learns discriminative features from the localized regions (see Fig. 2).

3.1. Multi-level attention network

In deep convolutional networks, the filters in the deeper layers have larger receptive field size, thus these filters can learn more high-level semantic information for fine-grained image classification [41,42]. Most previous works focus on the detection algorithm or the attention algorithm to locate the discriminative regions, and they use these regions to train a more generalized network for the fine-grained categorization. Some of them use annotation information to improve the performance of localizing the discriminative regions. But, all these methods do not take full advantage of the network itself. Inspired of the fact that the pre-trained networks from ImageNet [34] usually get better performance on many categorization tasks, we propose the method of selecting neural activations from the network which has been trained on the full-size image set, such as the CUB200-2011, Stanford Dogs, according to the task.

The deeper convolutional layers usually can capture high-level information, while the shallower layers can get more details about the targets [61]. But, we found that if only use one convolutional layer to get the region proposal could not get satisfying performance. Because although the deeper layer can get higher and more stable features compared with the shallower layer, it also has the drawback that it cannot capture enough detail for the fine-grained categorization. On the contrary, the shallower layer is easily attracted by certain edges or shapes which may affect the categorization task badly. So, we use these last three layers to select the neural activations which can bring us both the stable information of the object and enough details about it.

Compared with the network pre-trained on the ImmageNet, our network can capture more specific information which is crucial important for our method to localize the discriminative regions. Different from previous methods, our pre-trained network is well trained on the full-size image train set using (1). We also take fully use of the feature maps of the last three layers to get the multi-scale regions.

$$\mathscr{L}_{cl} = -\sum_{i=1}^{C} y_i log(\hat{y}_i), \tag{1}$$

where $\mathscr{L}_{cl}$ is the label loss, the y_i is the ground truth label, $\hat{y}_i$ is the predicted result.

Given an input image X, we firstly resize it to 448×448 and then feed it into the attention network to get the neural activations, which are also called as feature maps. The feature maps of $conv_{4\_3}$ and $conv_{5\_3}$ are unsampled using bilinear interpolation technology to the same size of $conv_{3\_3}$, and accumulated as follows to calculate the multi-level attention map.

$$M = \sigma\left\{\sum_{i=1}^{C_{3\_3}} M^i_{3\_3} + \sum_{i=1}^{C_{4\_3}} \widehat{M}^i_{4\_3} + \sum_{i=1}^{C_{5\_3}} \widehat{M}^i_{5\_3}\right\}, \tag{2}$$

where σ is the sigmoid function, C_{3_4}, C_{4_4} and C_{5_4} denote the channel numbers of the 3_3th, 4_3th, 5_3th convolutional layers

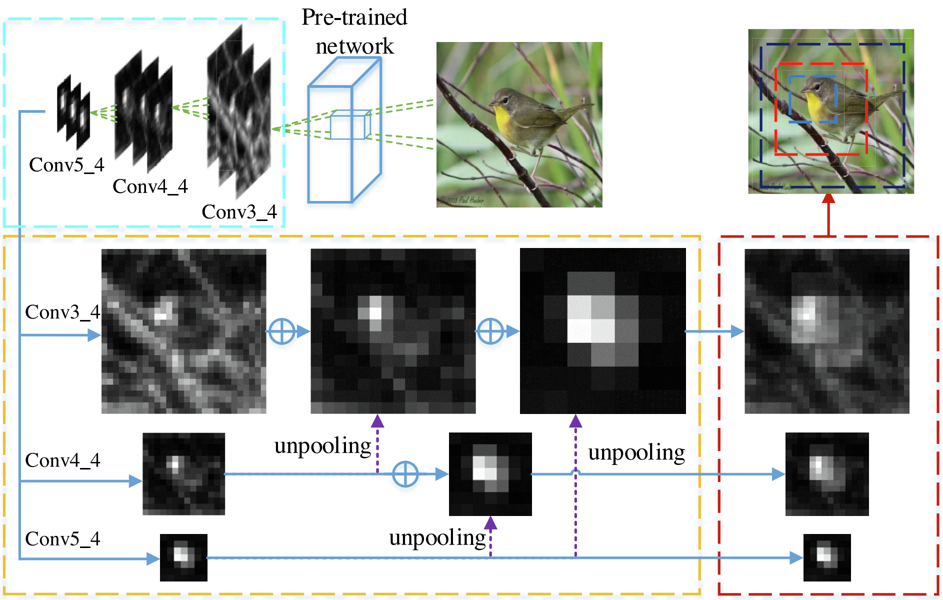


Fig. 2. The framework of proposed multi-level attention network. The inputs are full-size images and they are fed into the pre-trained network to get the neural activations of the last three convolutional layers, $conv_{3\_3}$, $conv_{4\_3}$ and $conv_{5\_3}$ respectively. The feature maps of $conv_{4\_3}$ and $conv_{5\_3}$ are unsampled using bilinear interpolation technology to the same size of $conv_{3\_3}$, and accumulated as follows to calculate the multi-level attention map.

respectively, M and $\widehat{M}$ denote the feature map and unsampled feature map respectively.

As demonstrated in Fig. 3, the attention maps have significant higher activations on the area of targeted object rather than the background. That's because our attention network has learned the domain specific information during the training stage, e.g., cars for the Stanford Cars dataset [62] and dogs for the Stanford Dogs dataset [6]. Filters of our attention network are trained together with the fully connected layers, each of them is constrained for recognizing the targets during the training stage, e.g., dogs or cars. Thus, filters with higher activations for the discriminative regions are reserved. Based on this point, our multi-level attention maps can be used to localize the right object in image.

3.2. Multi-level dictionary learning network

As we have selected the neurons that have the max response to some certain regions of the original image, the next step is localizing them.

When a pre-trained CNN model is applied on an input image X, its internal layers can capture abundant features. Assume the output of one convolutional(pooling) layer is a matrix of $C*N*N$ dimension where the size of the feature maps is $N*N$ and the channel number of this layer is C. Thus, each spatial cell of a convolutional (pooling) layer is computed by the region with the size of its receptive field. The size of the receptive field is determined by the kernel size and the stride of a convolutional layer. The receptive field size can be computed by Eq. (3).

$$R_i = s*(R_{i-1}-1)+k, \tag{3}$$

where R_{i-1} is the receptive field size of its input layer and R_i denotes the receptive field size of the output layer, s and k are the stride and kernel size respectively. Thus, one cell in these layers we select corresponds to $44*44$, $100*100$, and $212*212$ receptive filed in the input image of $224*224$ respectively.

Based on the different receptive field size of the convolutional (pooling) layers, we can map the selected neurons to the original image and get the regions as shown in Fig. 3. We can see that the mapping regions in the input image are in three different sizes ranging from small to large. This can simplify the procedure of localizing the discriminative regions compared with the traditional attention algorithm significantly. Especially when there is more than one object in the image, our CNN model can locate the target accurately in the fine-grained categorization task. We can see in Fig. 3, our method can locate the right target rather than the irrelevant background.

Let $x_i \in \mathbb{R}^m$ denote one training sample belonging to C classes, where N is the number of samples and m is the original dimensionality. Then, the sparsest representation α_i based on x_i can be obtained using the reconstructive dictionary D and the dictionary learning model can be to obtain from the following optimization problem:

$$\min_{D,\alpha_i}\sum_{i=1}^{N}||x_i - D\cdot\alpha_i||_2^2 + \lambda||\alpha_i||_p, \tag{4}$$

where $||x_i - D\cdot\alpha_i||_2^2$ is the reconstruction error, and λ is a positive scalar constant. $||\alpha_i||_p$ is the l_p-norm regularization, p corresponds to the widely-used l_0-norm/l_1-norm. By optimizing the Eq. (4), the α_i can be obtained for ensemble features extracted from the multi-scale regions.

Given a dictionary, the sparse coding is to represent a signal as a linear combination of a few atoms of it. For a set of feature $\{x_i\}_{i=1,2,\dots,m}$, finding a dictionary D that each feature in the dictionary D can be represented using the atoms of D for just a sparse linear combination. Therefore, we can optimize this task by using an iterative approach which lies in two steps: update step on fixed α for optimizing the dictionary D and calculate the sparse coding step on a fixed D for the α respectively. As shown in Table 2, the algorithm of the proposed MLA-CNN is presented.

Different from these methods, we use the end-to-end method by training several subtle networks and integrate their outputs. This can bring a significant advantage of simplicities and efficiency, and it can also get satisfying accuracy. As each subtle network is fed with those regions with a certain scale, they can learn more scale-specific information about the targets from the input images. Although any one of them can't reach the best accuracy, the result can be improved significantly by integrating them. As

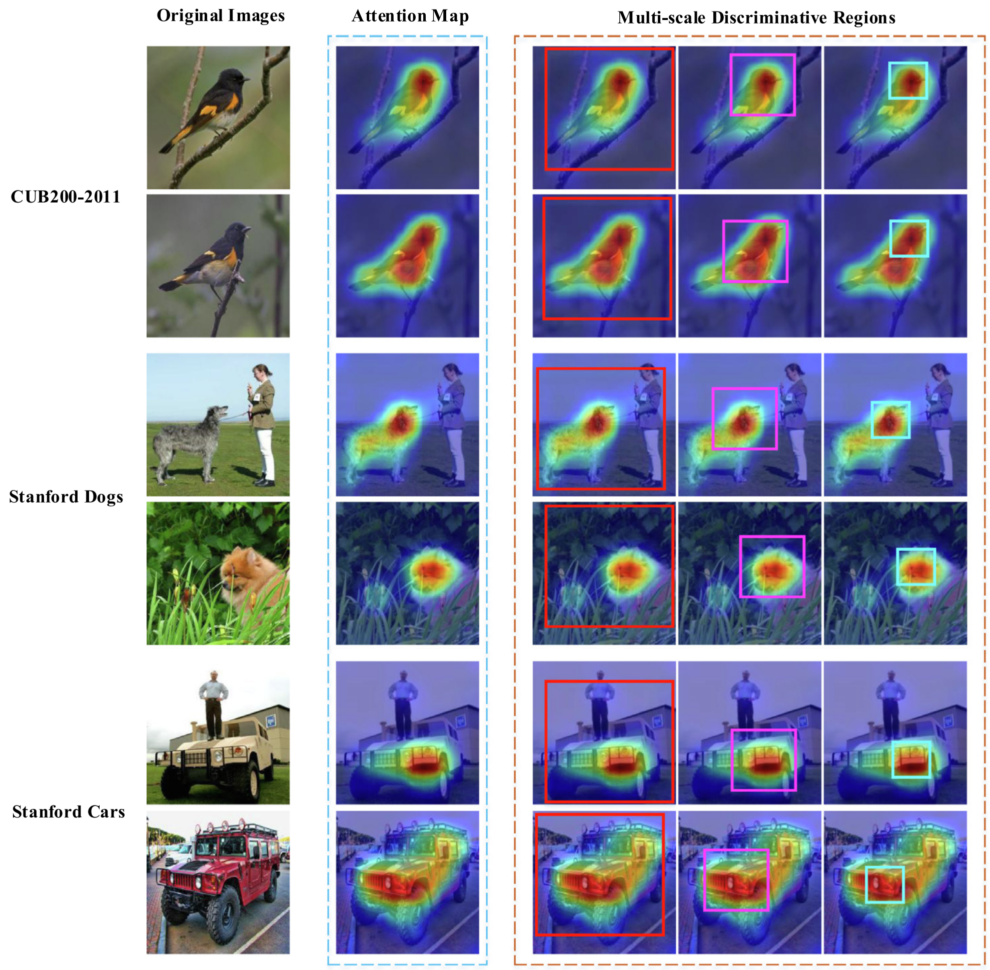


Fig. 3. The original images and multi-scale discriminative regions from the three benchmark of birds, dogs, and cars. The image in each row represents the original image and three scaled regions generated based on the neural activations in the feature maps. These regions are discriminative for the task of categorization of the birds, dogs, and cars.

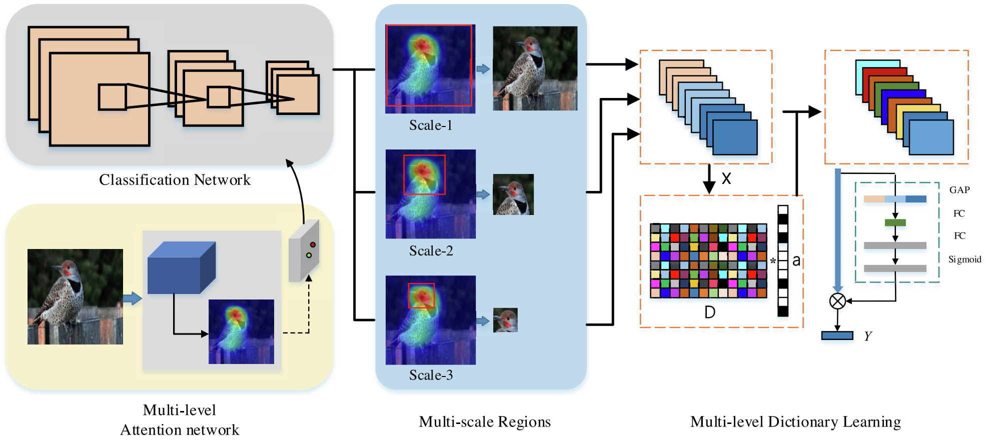


Fig. 4. The overall of proposed multi-level attention and dictionary learning network. Different discriminative regions are generated according the neural activations of the attention network. Multi-level features are then fused together for better recognition.

shown in Fig. 4, we use three CNN networks that are trained independently by these multi-scale regions. For one test image, we can get the predicted label by summing the results of these subtle networks.

To evaluate the efficiency of the proposed method, experiments on the CUB-200-2011 dataset [1] are conducted as shown in Table 1. The deep learning framework is Pytorch. Specifically, Table 1 shows the overall computation time. We can see that during the training phase, RA-CNN and ours take 6.3 and 5.1 h, respectively. During the testing phase, the inference speed of RA-CNN [22] and our proposed MLA-CNN are 18.7 and 21.2 fps (frame per second) respectively (see Fig. 5).

4. Experiments

In this section, we will present our experimental results on three widely used fine-grained image classification datasets and the comparison with these state-of-the-art methods. We also discuss the result of different networks which trained with different scaled regions.

4.1. Datasets and network

We evaluate our method on these three fine-grained benchmarks: CUB-200-2011 [1], Stanford Dogs [6] and Stanford Cars [62]. We use the default train and test spilt and the detailed statistics are summarized in Table 3.

We utilize the VGG-19 [63] as the backbone of our multi-level attention and dictionary learning network. All experiments are run on a computer with Intel Xeon E5 CPU, 64 GB main memory, and four Nvidia Titan GPUs with 48 GB memory in total. Our implementation is based on Pytorch, which is a popular framework for studying the deep neural network.

4.2. Implementation details

We usually use the CNN model which has been trained on the ImageNet [34] and fine-tune [64,65] it on the fine-grained image dataset. Similarly, for the Caltech-UCSD Birds-200-2011 dataset [1], we use the CNN model which has been trained on the full-size image train set as our region generator. As the CNN model has been trained on the image train set, it can capture a wealth of information about the subclasses of the dataset which is helpful for the localizing of these discriminative regions. The experimental results have shown that the pre-trained network can localize the regions effectively which can improve the accuracy of the fine-grained classification.

To make full use of these characteristics of the deep and shallow layers, we summarize these feature maps of deeper and shallow layers together for calculating the multi-level attention map. In detail, feature maps of $conv_{4\_3}$ and $conv_{5\_3}$ are unsampled using bilinear interpolation technology to the same size of $conv_{3\_3}$, and accumulated to calculate the multi-level attention map. As demonstrated in Fig. 3, multi-scale discriminative regions are obtained from the input image using Eq. (3). The receptive field sizes of one neuron are 44, 100, and 212 respectively. This means that one neuron of the multi-level attention can map a certain sized region from the input image.

In practice, we extend the receptive field size of one layer manually instead of just using its original receptive field size like $44 * 44$, $100 * 100$, and $212 * 212$. As the size of input image is the fixed size of $224 * 224$, we enlarge and lessen the receptive field size by about 20%. Then, we get three new different receptive field sizes of each layer.

Table 1
Computation comparison on CUB-200-2011 dataset.

Method	Training times (hours)	Testing speed (fps)
RA-CNN [22]	6.3	18.7
our MLA-CNN	5.1	21.2

4.3. Results and comparisons

Since the proposed method uses the neural activations of the pre-trained network to localize the most discriminative regions for the task of fine-grained classification, we mainly conduct the comparison of several methods that use the attention algorithm. To demonstrate the effectiveness of the attended regions by our model, we further perform the experimental comparison with the methods which are trained on samples annotated with extra region labels. In Fig. 3, we can observe that these regions localized by the pre-trained CNN model can focus on the discriminative regions of the image at different scales. By fusing the multi-level features of the localized discriminative regions, the classification network can achieve significantly better performance.

CUB-200-1-2011. The classification accuracy on CUB-200-2011 dataset is shown in Table 4. The baseline that just uses the VGG19 network has achieved 77.8% recognition accuracy. First, Part-RCNN [40], PA-CNN [66], and FCAN [42] are based on the attention algorithm and they also use the annotation information during the procedure of training and testing. These methods achieve 81.6%, 82.8%, and 84.3% respectively. Compared with these methods, our method can achieve significantly better accuracy of 85.4%. Meanwhile, because of the simplicity of our method, it can get the discriminative regions directly by these selected neurons without relying on the performance of the attention algorithm. Second, similar to the RA-CNN [22], we also use several subtle networks to improve recognition performance. But, instead of designing the complex loss functions elaborately, we use the pre-trained CNN model to localize the most discriminative regions and this can reduce the complexity and time of localizing these regions significantly. The experimental results have shown that our method can achieve the best accuracy of 85.4% compared with the RA-CNN [22]. We achieve about 3.4% gain compared with the FCAN [42] that without using the annotation information. As shown in Table 4, by integrating the results of these subtle networks, we can get higher accuracy than just using a single network.

Stanford Dogs. We show the categorization accuracy for the Stanford Dogs dataset in Table 5. The DVAN [20] emphasizes the importance of diversity of the attention maps and utilizes an LSTM (Long Short-Term Memory) recurrent unit to learn attentiveness and discrimination of the attention canvases. FCAN [42] utilizes a fully convolutional attention network to localize the regions and gets an accuracy of 84.2%. VGG-GAP [68] trains the network with zooming in the images iteratively and utilizes the class activation maps to generate discriminative regions, achieving the accuracy of 86.2%. We conduct comparisons on these methods which use attention to localize the regions as shown in Table 5. By integrating features of three subnetworks, our method can achieve the accuracy of 86.8% on the Stanford Dogs Dataset. It can be noted that our method has achieved about 5.3% and 2.6% gain than DVAN [20] and FCAN [42]. It can be noted that the RA-CNN [22] achieves slightly higher than our method (86.8% vs. 87.3%) on the Standford Dogs dataset. However, the architecture of our proposed method is much simpler than RA-CNN. Specifically, the RA-CNN architecture employs extra subnetworks called *branch network* to predict the categories recurrently. Moreover, RA-CNN framework adopts more elaborate training skills such as setting more hyper-parameters to achieve relatively high accuracy.

Similar to the experiment on birds, we also record the results of these subtle networks as shown in Table 5. We can observe that these subtle networks can achieve about 6% gain than the baseline, and the integration of them has reached 86.8%. This demonstrates that these regions localized by our method are indeed discriminative for the task of categorization.

Stanford Cars. We conduct the experiments on the Stanford Cars Dataset for completeness and summarize the comparisons

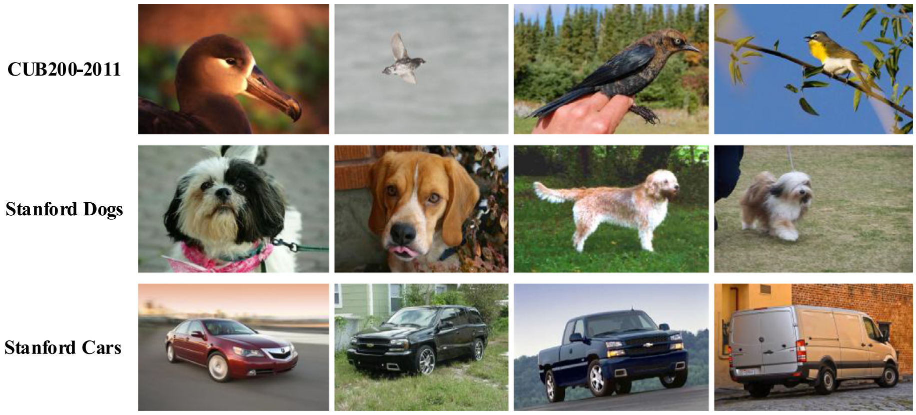


Fig. 5. Some example images of fine-grained image datasets used in the experiments. The background and scale of objects vary greatly among these images of the same datasets.

Table 2
Algorithm of MLA-CNN for fine-grained image categorization.

```
Algorithm of MLA-CNN for fine-grained image categorization

Input: Training images $X$, testing images $T$.
Output: Predicting label result.
Training:
1: Generate multiple discriminative patches $I_p$ according to the multi-level attention network.
2: Construct the multi-level deep features $X$ from the localized discriminative patches $I_p$.
3: Learn dictionaries $D$ and sparse matrix $\alpha$ based on the given features.
Testing:
4: Generate discriminative patches $I_p$ using the attention network.
5: Extract multi-level deep features $X$ from the localized patches.
6: Encode the features $X$ using the dictionary $D$.
7: Obtain the predicting label result.
```

Table 3
The train and test spilt of the datasets used in this paper.

Datasets	Category	Train	Test
CUB200-2011 [1]	200	5994	5794
Stanford Dogs [6]	120	12000	8580
Stanford Cars [62]	196	8144	8041

Table 4
Categorization performance comparisons on CUB-200-2011 dataset. The ✓ and n represent the methods using object/part annotations or not.

Method	Train/Test Anno.	Acc (%)
Part-RCNN [40]	✓/✓	81.6
PA-CNN [66]	✓/✓	82.8
FCAN [42]	✓/✓	84.3
Mask-CNN [67]	✓/✓	85.4
VGG-19 [63]	n/n	77.8
VGG-GAP [68]	n/n	79.5
MG-CNN [69]	n/n	81.7
TPPL [70]	n/n	81.7
FCAN [42]	n/n	82.0
RA-CNN [22]	n/n	85.3
MLA-CNN (scale 1)	n/n	78.9
MLA-CNN (scale 2)	n/n	81.8
MLA-CNN (scale 3)	n/n	80.4
MLA-CNN (integration)	n/n	**85.7**

Table 5
Categorization performance comparisons on Stanford Dogs dataset.

Method	Acc (%)
NAC(AlexNet) [71]	68.6
PDFR(AlexNet) [31]	71.9
VGG-19 [63]	79.1
DVAN [20]	81.5
FCAN [42]	84.2
VGG-GAP [68]	86.2
RA-CNN [22]	**87.3**
MLA-CNN (scale 1)	80.2
MLA-CNN (scale 2)	83.3
MLA-CNN (scale 3)	82.9
MLA-CNN (integration)	86.8

Table 6
Categorization performance comparisons on Stanford Cars dataset. The ✓ and n represent the methods using object/part annotations or not.

Method	Train/Test Anno.	Acc (%)
R-CNN [64]	✓/✓	88.4
PA-CNN [66]	✓/✓	92.8
VGG-19 [63]	n/n	80.2
FCAN [42]	n/n	89.1
RA-CNN [22]	n/n	**92.5**
MLA-CNN (scale 1)	n/n	72.1
MLA-CNN (scale 2)	n/n	85.4
MLA-CNN (scale 3)	n/n	85.6
MLA-CNN (integration)	n/n	91.2

with these previous works which use attention to localize the discriminative regions. Table 6 has shown the comparison results. We can observe that our method can also achieve 91.2% on the Stanford Cars dataset without using any manual annotations. Our method achieves a comparable performance to the PA-CNN [66] and RA-CNN [22]. It is noted that our method is much simpler than these counterparts without extra supervision.

5. Conclusion

In this paper, we present a framework for fine-grained recognition which uses the CNN model pre-trained on the full-size image train set of current task to localize the discriminative regions and trains subtle networks. By learning multi-level dictionary features and integrating the results of these subtle networks, we can get the state-of-the-art accuracy on three benchmarks. Our method is free of any object/part annotations at both training and testing stages. Our basic idea is to use the inner structure of the CNN network to localize the most discriminative regions and train a more generalized and robust model for the fine-grained categorization task. We mainly claim two contributions. First, the multi-level attention model is proposed to localize the discriminative regions for fine-grained image recognition. Second, we integrate the multi-scale discriminative regions by training the subtle networks and learning multi-level dictionary features from the localized discriminative regions.

Extensive experimental results have shown that our method can significantly improve the performance of fine-grained classification. Future works would focus on improving the performance of the pre-trained CNN model and finding a more effective method of integrating these subtle networks.

CRediT authorship contribution statement

Jinsheng Ji: Conceptualization, Methodology, Writing - original draft. **Yiyou Guo:** Data curation, Software. **Zhen Yang:** Visualization, Investigation. **Tao Zhang:** Validation, Writing - review & editing. **Xiankai Lu:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work was supported by the National Science Foundation of China (NSFC), under contract No. 61866016, the Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing.

References

[1] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001..

[2] T. Berg, J. Liu, S. Woo Lee, M.L. Alexander, D.W. Jacobs, P.N. Belhumeur, Birdsnap: Large-scale fine-grained visual categorization of birds, in: Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2011–2018.

[3] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, S. Belongie, Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection, in: Computer Vision and Pattern Recognition (CVPR), 2015, pp. 595–604.

[4] L. Xie, Q. Tian, M. Wang, B. Zhang, Spatial pooling of heterogeneous features for image classification, IEEE Transactions on Image Processing 23 (5) (2014) 1994–2008.

[5] A. Iscen, G. Tolias, P.-H. Gosselin, H. Jégou, A comparison of dense region detectors for image search and fine-grained classification, IEEE Transactions on Image Processing 24 (8) (2015) 2369–2381.

[6] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization: Stanford dogs, in: Computer Vision and Pattern Recognition Workshops, vol. 2, 2011, p. 1..

[7] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: ICVGIP, IEEE, 2008, pp. 722–729.

[8] J.D. Wegner, S. Branson, D. Hall, K. Schindler, P. Perona, Cataloging public objects using aerial and street-level images-urban trees, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6014–6023.

[9] M. Stark, J. Krause, B. Pepik, D. Meger, J.J. Little, B. Schiele, D. Koller, Fine-grained categorization for 3d scene understanding, International Journal of Robotics Research 30 (13) (2011) 1543–1552.

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

[11] J. Liu, A. Kanazawa, D. Jacobs, P. Belhumeur, Dog breed classification using part localization, in: European Conference on Computer Vision, Springer, 2012, pp. 172–185.

[12] S. Branson, G.V. Horn, S.J. Belongie, P. Perona, Bird species categorization using pose normalized deep convolutional nets, CoRR (abs/1406.2952.).

[13] N. Zhang, E. Shelhamer, Y. Gao, T. Darrell, Fine-grained pose prediction, normalization, and recognition, arXiv preprint arXiv:1511.07063..

[14] R. Farrell, O. Oza, N. Zhang, V.I. Morariu, T. Darrell, L.S. Davis, Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance, in: International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 161–168.

[15] N. Zhang, R. Farrell, T. Darrell, Pose pooling kernels for sub-category recognition, in: Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 3665–3672..

[16] L. Bourdev, S. Maji, T. Brox, J. Malik, Detecting people using mutually consistent poselet activations, in: European Conference on Computer Vision (ECCV), Springer, 2010, pp. 168–181.

[17] Y. Chai, V. Lempitsky, A. Zisserman, Symbiotic segmentation and part localization for fine-grained categorization, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 321–328.

[18] L. Xie, Q. Tian, R. Hong, S. Yan, B. Zhang, Hierarchical part matching for fine-grained visual categorization, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1641–1648.

[19] N. Zhang, R. Farrell, F. Iandola, T. Darrell, Deformable part descriptors for fine-grained recognition and attribute prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 729–736.

[20] B. Zhao, X. Wu, J. Feng, Q. Peng, S. Yan, Diversified visual attention networks for fine-grained object classification, IEEE Transactions on Multimedia 19 (6) (2017) 1245–1256.

[21] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: Computer Vision and Pattern Recognition (CVPR), 2015, pp. 842–850.

[22] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4438–4446.

[23] H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5209–5217.

[24] W. Wang, J. Shen, Deep visual attention prediction, IEEE Transactions on Image Processing 27 (5) (2017) 2368–2378.

[25] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (7) (2018) 1531–1544.

[26] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: unsupervised video object segmentation with co-attention siamese networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3623–3632.

[27] W. Wang, J. Shen, X. Lu, S.C. Hoi, H. Ling, Paying attention to video object pattern understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) (to be published)..

[28] J. Shen, X. Tang, X. Dong, L. Shao, Visual object tracking by hierarchical attention siamese network, IEEE transactions on cybernetics (2019) (to be published)..

[29] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1448–1457.

[30] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, M.N. Do, Weakly supervised fine-grained categorization with part-based image representation, IEEE Transactions on Image Processing 25 (4) (2016) 1713–1725.

[31] X. Zhang, H. Xiong, W. Zhou, W. Lin, Q. Tian, Picking neural activations for fine-grained recognition, IEEE Transactions on Multimedia 19 (12) (2017) 2736–2750.

[32] M. Wang, C. Luo, B. Ni, J. Yuan, J. Wang, S. Yan, First-person daily activity recognition with manipulated object proposals and non-linear feature fusion, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2017) 2946–2955.

[33] S. Li, M. Shao, Y. Fu, Person re-identification by cross-view multi-level dictionary learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (12) (2017) 2963–2977.

[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105..

[35] J. Fu, T. Mei, K. Yang, H. Lu, Y. Rui, Tagging personal photos with transfer deep learning, in: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2015, pp. 344–354..

[36] J. Fu, J. Wang, Y. Rui, X.-J. Wang, T. Mei, H. Lu, Image tag refinement with view-dependent concept representations, IEEE Transactions on Circuits and Systems for Video Technology 25 (8) (2015) 1409–1422.

[37] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, Y. Rui, Relaxing from vocabulary: robust weakly-supervised deep learning for vocabulary-free image tagging, in: International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1985–1993.

[38] G.-S. Xie, X.-Y. Zhang, W. Yang, M. Xu, S. Yan, C.-L. Liu, Lg-cnn: From local parts to global discrimination for fine-grained recognition, Pattern Recognition 71 (2017) 118–131.

[39] B. Zhao, J. Feng, X. Wu, S. Yan, A survey on deep learning-based fine-grained object classification and semantic segmentation, International Journal of Automation and Computing (2017) 1–17.

[40] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based r-cnns for fine-grained category detection, in: European Conference on Computer Vision, Springer, 2014, pp. 834–849.

[41] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2017) 1309–1322.

[42] X. Liu, T. Xia, J. Wang, Y. Lin, Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition, CoRR, abs/1603.06765..

[43] T. Berg, P.N. Belhumeur, Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation, in: Computer Vision and Pattern Recognition, 2013, pp. 955–962.

[44] S. Yang, L. Bo, J. Wang, L.G. Shapiro, Unsupervised template learning for fine-grained object recognition, in: Advances in Neural Information Processing Systems, 2012, pp. 3122–3130..

[45] E. Gavves, B. Fernando, C.G. Snoek, A.W. Smeulders, T. Tuytelaars, Fine-grained categorization by alignments, in: ICCV, 2013, pp. 1713–1720.

[46] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, L. Shao, Attentive region embedding network for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9384–9393.

[47] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99..

[48] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[49] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[50] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[51] Z. Zhang, Y. Xu, L. Shao, J. Yang, Discriminative block-diagonal representation learning for image recognition, IEEE Transactions on Neural Networks and Learning Systems 29 (7) (2017) 3111–3125.

[52] G.-S. Xie, X.-Y. Zhang, S. Yan, C.-L. Liu, Sde: A novel selective, discriminative and equalizing feature representation for visual recognition, International Journal of Computer Vision 124 (2) (2017) 145–168.

[53] G.-S. Xie, Z. Zhang, L. Liu, F. Zhu, X.-Y. Zhang, L. Shao, X. Li, Srsc: Selective, robust, and supervised constrained feature representation for image classification, IEEE Transactions on Neural Networks and Learning Systems..

[54] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, L. Shao, Attentive region embedding network for zero-shot learning, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9384–9393.

[55] X. Zhang, H. Xiong, W. Zhou, W. Lin, Q. Tian, Picking deep filter responses for fine-grained image recognition, in: Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1134–1142.

[56] Y. Sun, Z. Zhang, W. Jiang, Z. Zhang, L. Zhang, S. Yan, M. Wang, Discriminative local sparse representation by robust adaptive dictionary pair learning, IEEE Transactions on Neural Networks and Learning Systems (2020) (to be published)..

[57] G.-S. Xie, X.-Y. Zhang, S. Yan, C.-L. Liu, Hybrid cnn and dictionary-based models for scene recognition and domain adaptation, IEEE Transactions on Circuits and Systems for Video Technology 27 (6) (2015) 1263–1274.

[58] Z. Zhang, Y. Sun, Y. Wang, Z. Zhang, H. Zhang, G. Liu, M. Wang, Twin-incoherent self-expressive locality-adaptive latent dictionary pair learning for classification, IEEE Transactions on Neural Networks and Learning Systems (2020) 1–15.

[59] Z. Zhang, W. Jiang, J. Qin, L. Zhang, F. Li, M. Zhang, S. Yan, Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier, IEEE Transactions on Neural Networks and Learning Systems 29 (8) (2018) 3798–3814.

[60] Z. Zhang, W. Jiang, Z. Zhang, S. Li, G. Liu, J. Qin, Scalable block-diagonal locality-constrained projective dictionary learning, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, ijcai.org, 2019, pp. 4376–4382..

[61] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.

[62] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization, in: International Conference on Computer Vision Workshops (ICCVW), IEEE, 2013, pp. 554–561.

[63] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014) 1409–1556..

[64] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.

[65] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717–1724.

[66] J. Krause, H. Jin, J. Yang, L. Fei-Fei, Fine-grained recognition without part annotations, in: Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 5546–5555..

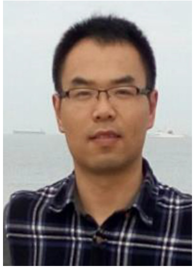
[67] X.-S. Wei, C.-W. Xie, J. Wu, C. Shen, Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization, Pattern Recognition 76 (2018) 704–714.

[68] A. Rosenfeld, S. Ullman, Visual concept recognition and localization via iterative introspection, in: Asian Conference on Computer Vision, Springer, 2016, pp. 264–279.

[69] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, Z. Zhang, Multiple granularity descriptors for fine-grained categorization, in: International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 2399–2406.

[70] C. Huang, Z. He, G. Cao, W. Cao, Task-driven progressive part localization for fine-grained object recognition, IEEE Transactions on Multimedia 18 (12) (2016) 2372–2383.

[71] M. Simon, E. Rodner, Neural activation constellations: Unsupervised part model discovery with convolutional networks, in: International Conference on Computer Vision (ICCV), 2015, pp. 1143–1151.



Jinsheng Ji received the B.S. degree in automation from Nanjing Agricultural University in and the M.S. degree in control science and engineering from Shanghai Jiao Tong University, China. He is currently pursuing the Ph. D. degree in Department of Automation at Shanghai Jiao Tong University, China. His research interests are computer version and machine learning.

Yiyou Guo received B.S. in Geomatics Engineering and M.S. degrees in GIS from Wuhan University, China, in 2005 and 2010, respectively. He obtained his Ph.D degree from Shanghai Jiao Tong University, China, in 2020. He worked with Fujian Surveying and Mapping Institute, Fuzhou. He is currently a Post-Doctoral Researcher with the College of Surveying and Geo-Informatics, Tongji University. His research interests include feature reduction, deep learning, image processing, and object tracking.

Zhen Yang received the B.S. degree in Automation from Changchun Institute of Technology in 2008, the M.S. degree in Automation from Qingdao University of Science and Technology in 2011, and Ph.D. degree from Shanghai Jiao Tong University in 2016. Recently, he is a professor of school of communication and electronics in Jiangxi Science and Technology Normal University, Nanchang, China. His research interests include computer vision, machine learning, object recognition and image classification.

Tao Zhang received the B.S. degree in electronic information engineering from Huainan Normal University in 2011, the M.S. degree in communication and information system from Sichuan University in 2014, and the Ph.D degree in control science and engineering from Shanghai Jiao Tong University in 2019. Currently, he studies as a Post-Doctor in Tsinghua University, China. His research work focuses on PolSAR image processing and machine learning.

Xiankai Lu is a Tenure-track Professor in the School of Software, Shandong University. From 2018 to 2020, he was a research associate with Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He received the Ph.D. degree from Shanghai Jiao Tong University in 2018. His research interests include computer vision and deep learning.