The Visual Computer (2022) 38:811–820
https://doi.org/10.1007/s00371-020-02052-8

ORIGINAL ARTICLE

Fine-grained visual classification via multilayer bilinear pooling with object localization

Ming Li[1] · Lin Lei[1] · Hao Sun[1] · Xiao Li[1] · Gangyao Kuang[1]

Accepted: 21 December 2020 / Published online: 9 January 2021

Abstract
Fine-grained visual classification is a challenging task in the computer vision field. How to explore discriminative features is vital for classification. As one crucial step, exactly object localization is able to eliminate the background noises and highlight interesting objects at the same time. However, some current methods usually use bounding boxes to locate objects, that are not suitable when the poses of objects change. Furthermore, it has been demonstrated that deep features have strong feature representation capability, especially the bilinear pooling features, which achieved superior performance in fine-grained visual classification tasks. However, the bilinear features, which captured only from the last convolutional layer, have limited discriminability, especially when dealing with small-scale objects. In this paper, we propose a multilayer bilinear pooling model combined with object localization. First, a flexible and scalable object localization module is utilized to locate the interesting object in an image instead of using bounding boxes. Then the refined features are obtained by highlighting object region and suppressing background noises. While the multilayer bilinear pooling, which exploits the complementarity between different layers, is used for further extracting more discriminative features. Experiment results on three public datasets show that our proposed method can achieve competitive performance compared with several state-of-the-art methods.

Keywords Fine-grained visual classification · Multilayer bilinear pooling (MLBP) · Object localization · Convolutional neural networks (CNNs)

1 Introduction

Fine-Grained Visual Classification (FGVC) task aims to classify subordinate-level categories of a common visual category, such as bird species [1], car [2], or aircraft [3] models, and so on. Although cosiderable success has been achieved for traditional coarse-grained classification, such as ImageNet [4], using deep convolutional neural networks (DCNNs). There are still some challenges in FGVC, which can be summarized as follows: (1) Subtle inter-class differences and large intra-class diversity among images. As shown in Fig. 1, the California Gull and the Glaucous-winged Gull birds may look very similar in global appearance except for some subtle differences, e.g., the color style of a birds beak. However, for the same bird species, such as the California Gull, the images vary a lot due to different poses, various viewpoints, and different illumination conditions. Therefore, extracting discriminative features from subtle local regions are critical for solving fine-grained visual classification. (2) Multi-scale characteristics. The scale of the objects in images varies a lot due to the effect of distance, viewpoint, and different poses of objects. As shown in Fig. 1, most of the birds or aircraft cover the entire images, however, there are still some birds that only occupy a small part of images because of long-distance imaging. If the multi-scale changes cannot be dealt with well in the classification process, the final classification performance may decrease. (3) Expensive human annotation costs. Compared with coarse-grained annotation, labeling fine-grained categories usually needs specialized knowledge

✉ Lin Lei
alaleilin@163.com

Ming Li
liming17@nudt.edu.cn

Hao Sun
clhaosun@gmail.com

Xiao Li
lxcherishm@126.com

Gangyao Kuang
Kuangyeats@hotmail.com

[1] College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

with the corresponding field and requires a large amount of annotation time.

Recently, the development of deep learning technology has been further boosted the classification performance. In general, existing CNN-based FGVC approaches can be roughly categorized into three main groups [5]: part/attention-based methods which aim to locate discriminative part regions and then learn region-based feature representation; end-to-end discriminative feature learning methods which hope to learn better visual representation directly from the original image; and external information-based methods which attempt to use auxiliary information, such as text data [6], Web data [7], for further improving classification accuracy. Due to the third mainstream approaches require additional information assistance beyond the fine-grained image itself to enhance the discrimination of feature representation, we mainly review the first two mainstreams as follows.

The main idea of the first mainstream is that part regions of an object usually play an important role in differentiating sub-categories. Therefore, many researchers in the fine-grained community pay more attention to locate the foreground or semantic parts of the object. Early works [8–10] usually rely on additional annotations, e.g., using available bounding boxes or part annotations to locate objects/parts. And then, constructing discriminative feature representation which corresponding to these parts for further classification. Zhang et al. [8] proposed a Part-based R-CNN method which utilized R-CNN to detect objects and parts under a geometric prior at first, and then gave the predicted label. Lin et al. [9] proposed a feedback-control framework to localization by using back-propagate alignment and classification errors. Wei et al. [10] proposed the Mask CNN algorithm for fine-grained classification by using part-level and image-level annotations. However, these methods have obvious limitations: first, defining the key parts of the object and annotating these parts in images need relevant export knowledge of the field. Simultaneously, the procedure of annotation is time-consuming and laborious. Therefore, some methods [11–13] that only require image-level labels draw more and more attention. Liu et al. [11] presented a fully convolutional attention networks (FCANs) with reinforcement learning. Fu et al. [12] proposed RA-CNN to predict the discriminative areas by using recurrent attention mechanism and then extract the corresponding features from coarse-scale to fine-scale. In [12], three different scale features are combined to predict the final category. To locate several attention regions simultaneously, MA-CNN [13] used channel grouping loss to generate multiple parts by clustering. Still, the number of parts is limited, which might decrease the classification accuracy. In summary, all of the above methods still utilized the bounding boxes to locate the objects or parts. When the poses of objects changes, or the shapes of objects like strip, the classification performance will be limited due to lack of flexibility.

Fig. 1 Examples of CUB-200-2011 and FGVC-Aircraft datasets. The blue dotted box represents the general image classification task (Coarse-Grained), while the red and green dotted box represent the fine-grained image classification, respectively

Different from the first mainstream, the second group methods attempt to directly learn more discriminative features by designing powerful deep network architectures for classification. Compared with the raw pixel information, the second-order statistics have been shown impressive performance [14,15]. Inspired by the second-order descriptors, a simple but efficient method called bilinear convolutional neural networks (B-CNN) has been proposed [16], which achieved state-of-the-art classification performance on various fine-grained datasets. It used the outer-product operation to pool the pairwise-correlated local descriptors, captured from two parallel CNNs, into a global representation. However, the bilinear features representation usually suffers from high-dimension, which increase computation burdens and requires a large number of training samples to fit. To address these limitations, some improvement works [17–20] have been proposed, i.e., compact bilinear pooling [17], low-rank bilinear pooling [18]. Although significant achievements have been made, the methods as mentioned above still have certain limitations due to the scale variation of the objects in images. When the object only occupies a small area of background, e.g., the California Gull in the center position of Fig. 1. The features of small-scale objects are easily be obliterated by background during the feature learning.

In this paper, we propose a multilayer bilinear model combined with object localization for fine-grained visual classification to address the limitations mentioned above. The proposed method mainly consists of two modules: object localization block, and multilayer bilinear pooling block. The first block uses to eliminate the background noises and obtain more pure features about the main object region. The second block utilizes to capture more discriminative features through multilayer bilinear pooling operation. Our main contributions are summarized as follows. (1) Instead of using bounding boxes, we use the object edge to precisely locate the main objects, which is more flexible. This way, we can suppress the background noises as much as possible and highlight the small scale objects simultaneously. (2) With

the number of layers of network gradually deepening, the bilinear features, which obtained only from the last convolutional layer, may have the limited discriminative capability, especially for small-scale objects in images. Considering the feature complementarity between different convolutional layers. A multilayer bilinear pooling operation adopts to enhance the discriminative of features further. Experiments and comparisons on three widely used fine-grained datasets demonstrated the effectiveness and competitiveness of our proposed method.

The rest of this paper is organized as follows. Section 2 introduces some related works. In Sect. 3, the overall architecture and corresponding sub-modules are represented in detail. Experiments setup and results are analyzed in Sect. 4. Section 5 is the conclusion and future work.

2 Related work

2.1 Object localization

Detecting and locating the target in images is an important but challenging task in computer vision. Currently, with the wide application of deep learning, a large amount of literature has been proposed in this field, which can be classified as follows: fully-supervised-based methods, weakly-supervised-based methods, and unsupervised-based methods. Fully-supervised-based methods usually utilize annotations as supervised information, e.g., bounding boxes, key points, to achieve precise localization results, e.g., Faster-RCNN [23] is a classical two-stage object detection method, which utilized region proposal network (RPN) for generating regions of interest (ROI) at first, and then used the R-CNN to classify and locate the object in the region. SSD [24] and YOLO [25] series of methods are popular one-stage methods for object detection. Besides, some anchor-free techniques are also proposed to address the anchor problem, e.g., CenterNet [26]. However, all of the above methods need large, detailed, and accurate annotations, which are expensive, time-consuming, and may not suitable for large-scale practical applications. Unlike the first category, weakly-supervised-based methods aim to solve expensive annotation problems by only considering image-level labels for localization. Zhou et al. [27] proposed a class activation mapping (CAM) technique for discriminative localization without using any bounding box annotations. Later, J. Choe et al. [28] utilized an attention-based dropout layer (ADL) to make the CNN classifier learn the whole region of the object, which more useful than CAM.

While unsupervised-based methods are more challenging because it does not rely on any auxiliary information rather than a given image. Some works [27,29] demonstrated that the convolutional activations have the ability to represent spatial and semantic information simultaneously and have remarkable localization ability. Zhang et al. [30] proposed an automatic fine-grained classification method without using any object/part annotation at both training and testing stages. Wei et al. [31] proposed a selective convolutional descriptor aggregation method to selectively fuse multilayer convolutional features of a pre-trained VGG-16 model. And then, a mean-threshold strategy was adopted to locate the main object in fine-grained images. To reduce annotation cost and avoid complex model training and optimization. In this paper, we refer the method in [31], which can not only avoid to introduce extra parameters, but also be well embedded into the model.

2.2 Bilinear pooling model

As a simple but effective approach for fine-grained visual classification task, bilinear convolutional neural network (B-CNN) achieved state-of-the-art in various fine-grained datasets. B-CNN utilized outer-product operation to pool the pairwise-correlated local descriptors into a global representation, i.e., bilinear features. Compared with first-order pooling methods, such as global average pooling (GAP), bilinear pooling can model second-order statistics information and obtain more discriminative features. However, the representation power of bilinear pooling features comes at the cost of very high-dimensional, e.g., $512 \times 512 \approx 262k$ dimension for VGG-16 model, leading to high computational burdens and risk of overfitting. To address these problems, some improved methods have been proposed. Random Maclaurin [32] or Tensor Sketching [33]-based compact model was investigated in [17] and obtained similar classification performance with relatively few parameters. Kong et al. [18] proposed a low-rank bilinear method to solve the computation dilemma problem and also obtained similar performance. Currently, a squeezed bilinear pooling method [19] was proposed, which aimed to decrease both the computation cost and the feature dimension simultaneously. In order to explore higher-order features, Cui et al. [20] proposed a deep kernel pooling approach by constructing compact explicit feature mapping. Beyond that, some other works [21]-[22] that model the interaction between different convolutional layers by using bilinear pooling further improved the classification performance. Sun et al. [21] proposed the hyperlayer bilinear pooling (HLBP) approach to exploit the information inherent in different convolutional layers. Yu et al. [22] proposed the hierarchical bilinear pooling (HBP) to enhance the feature representation capability by integrating multiple cross-layer bilinear features and achieved state-of-the-art performance. Inspired by the above methods, different from the HBP, we first calculate the bilinear features by using the last convolutional feature map. Then the features from preceding layers

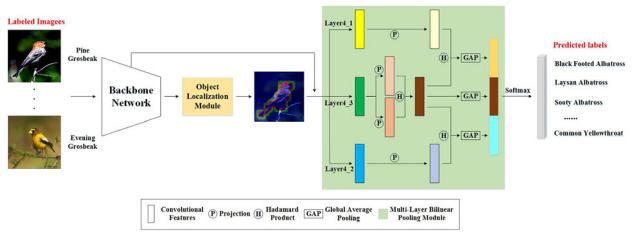


Fig. 2 Overall architecture of the proposed method

are used as complementary information further to enhance the representation capability of the bilinear features.

3 Proposed method

In this section, we explain our proposed method in detail. As shown in Fig. 2, the framework of our method is composed of two main components: (1) The object localization module utilized to locate the interesting object in given images. (2) The multilayer bilinear pooling module used to extract more discriminative feature representation.

3.1 Object localization module

By analyzing the convolutional feature activation response of each channel. Wei et al. [31] proposed a selective convolutional descriptor aggregation (SCDA) algorithm to select useful convolutional features that correspond to the main object in an image and simultaneously abandon the background or noise regions. Inspired by SCDA, we attempt to improve the positioning performance through a series of measures. In this paper, we select ResNet-34 [34] network as the backbone. Suppose $F \in \mathbb{R}^{H\times W\times C}$ denote the last convolutional feature map with C channels and spatial size $H \times W$ of an input image X. At first, activation map A can be obtained by adding up the feature maps F through the channel direction. After that, a threshold T, obtained by calculating the mean value of A, is used to determine whether the activation response in position (i, j) is selected or discarded. If the activation response value of position (i, j) is higher than T, the position will be retained. As represented in Eq. (1),

$$M_{(i,j)} = \begin{cases} 1, \ if \ A_{(i,j)} > T \\ 0, \quad otherwise \end{cases} \tag{1}$$

Thus, we will obtain a coarse mask map M. Considering that different convolutional feature map has different activation responses to the object, [31] also demonstrated that the positioning performance obtains significant improvement by fusing multiple convolutional layers. Suppose $layer4\_1$, $layer4\_2$, and $layer4\_3$ represent the convolutional feature maps of the last layer of ResNet-34, respectively, we can obtain the initial coarse mask map, M_{4_1}, M_{4_2}, and M_{4_3} of $layer4\_1$, $layer4\_2$ and $layer4\_3$ by repeating Eq. (1), respectively. Then, a more accurate mask M will be generated by pixel-wise multiplication operation. To visualize the result, M is resized to the size of the input image X and then overlay it on X. The entire framework and visual result are shown in Fig. 3. It can be seen that the above method can obtain accurate positioning performance without introducing any training parameters. In the following section, we will evaluate its positioning performance through qualitative analysis with experiments.

3.2 Multilayer bilinear pooling module

Supposed $X \in \mathbb{R}^{H\times W\times C}$ is the last convolutional feature map of a backbone CNN, e.g., ResNet-34, H, W, and C are the height, width, and the number of channels, respectively. The C-dimensional matrix at a spatial location on X can be denoted as $\boldsymbol{x} = [x_1, x_2, \cdots, x_c]^T$, each row in matrix $\boldsymbol{x}$ represent a local feature vector. Then the features $B(\boldsymbol{x})$ of fully bilinear pooling can be calculated by

$$B(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} W_i \boldsymbol{x} \tag{2}$$

where $W_i \in \mathbb{R}^{c\times c}$ represents the weight matrix. However, the $B(\boldsymbol{x})$ generated by bilinear pooling generally has a large dimension which increases the computation cost and also easy to over-fitting. In order to address these limitations, according to matrix factorization in [35], the weight matrix W_i can be factorized into two one-rank vectors, denoted $U_i \in \mathbb{R}^c$ and $V_i \in \mathbb{R}^c$. So Eq. (2) can be rewritten as

$$B(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} W_i \boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}} U_i V_i^{\mathrm{T}} \boldsymbol{x} = \left(U_i^{\mathrm{T}} \boldsymbol{x}\right) \circ \left(V_i^{\mathrm{T}} \boldsymbol{x}\right) \tag{3}$$

where $\circ$ is the Hadamard product operation. When redefining $U, V \in \mathbb{R}^{c\times d}$ as low-rank project matrices, d is a hyperparameter. Equation (3) becomes

$$B(\boldsymbol{x}) = P^{\mathrm{T}}\left(U^{\mathrm{T}}\boldsymbol{x} \circ V^{\mathrm{T}}\boldsymbol{x}\right) = SumPooling\left(U^{\mathrm{T}}\boldsymbol{x} \circ V^{\mathrm{T}}\boldsymbol{x}\right) \tag{4}$$

where $SumPooling$ is a pooling operation which calculates the sums of all spatial localizations in each feature map.

Considering different convolutional layers of CNN can represent different characteristics of input image. Some works [21,22] attempted to capture more discriminative features by modeling the interaction between different convolutional layers, which significantly improved the classification accuracy. Inspired by these ideas, unlike the HBP approach, we construct a new multilayer bilinear pooling module to model the interaction of different layers for further enhancing the discriminative of features.

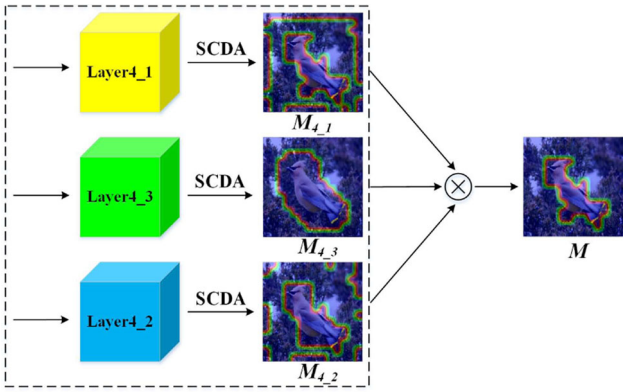


Fig. 3 Pipeline of the object localization module

As shown in Fig. 3, assuming X, Y, and Z denote the convolutional feature maps of $layer4\_3$, $layer4\_1$, and $layer4\_2$ from the last convolutional layer of ResNet-34, respectively. As we all know, the features from deeper layers can represent more semantic information which are more related to the task. Thus, the bilinear features which generated by utilizing the last convolutional layer usually more informative. Based on Eq. (4), we first calculate the bilinear feature $B(X)$ from X, and then the bilinear feature $B(B(X), Y)$ between $B(X)$ and Y, bilinear feature $B(B(X), Z)$ between $B(X)$ and Z are calculated, respectively. In actually, these operations can be treated as using other's information to compensate the basic bilinear features. Therefore, Eq. (4) can be extended as

$$\begin{aligned} &B(X, Y, Z) = P^{\mathrm{T}} Concat\left(B(X), B(B(X), Y), B(B(X), Z)\right) \\ &where \\ &B(X) = SumPooling\left(U^{\mathrm{T}}X \circ V^{\mathrm{T}}X\right) \\ &B(B(X), Y) = SumPooling\left(\left(U^{\mathrm{T}}X \circ V^{\mathrm{T}}X\right) \circ \left(S^{\mathrm{T}}Y\right)\right) \\ &B(B(X), Z) = SumPooling\left(\left(U^{\mathrm{T}}X \circ V^{\mathrm{T}}X\right) \circ \left(W^{\mathrm{T}}Z\right)\right) \end{aligned} \tag{5}$$

where U, V, S, and W are the projection matrices of the feature maps, respectively. $P \in \mathbb{R}^{d \times n}$ is a project matrix of feature embedding, here we set d is 8192 by reference [22] in this paper. $Concat$ denotes the concatenation operation. By modeling the interaction between different layers, more discriminative features can be obtained which are significant useful for classification. The overall architecture of multilayer bilinear pooling procedure is shown in Fig. 2.

4 Experiments

In this section, we evaluate the performance of our proposed method on three widely and challenging benchmark datasets, including Caltech-UCSD Birds (CUB-200-2011) [1], FGVC-Aircraft [3] and Stanford-Cars [2]. Note that our experiments only use image-level labels instead of utilizing any bounding box/part annotations.

4.1 Datasets and implementation details

Datasets: The CUB-200-2011 dataset is a classic fine-grained benchmark dataset that contains 11,788 images of 200 different bird species. Almost all the FGVC methods choose it for evaluating classification performance. The FGVC-Aircraft dataset contains 100 different aircraft models with roughly 100 images for each class. The Stanford-Cars is a 196 class dataset of 8144 training images and 8041 test images. Detail information about three datasets are shown in Table 1.

Parameters Set: The pre-trained of ResNet-34 is adopted as the backbone network. We remove the average pooling layer to obtain the last convolutional layer features. For the input of the network, all images are resized to 600×600 pixels by bilinear interpolation at the first and then horizontal flip. The final images are 448×448 pixels, which random crop in the training stage and center crop in the testing stage, respectively. We use a two-step training strategy to train the network. First, we initially train only the classifiers by fixing the backbone network and then fine-tune the whole network using stochastic gradient descent (SGD) with a batch size of 32, momentum of 0.9, weight decay of 5×10^{-4}. In the first and second steps, the initial learning rates are set 1.0 and 0.1, respectively. After every 60 epochs, the learning rate decay is 10. Besides, the overall accuracy (OA) and confusion matrix (CM) are considered as evaluators to analyze the classification performance.

Hardware and Software: All experiments are completed by using Pytorch 1.3, a deep learning library framework, on PC using 32-GB memory, i7-8700K CPU processor and a single NVIDIA GeoForce GTX 1080ti GPU with 11 GB memory (The source code will be released at www.github.com/ww-hh/).

4.2 Comparison with state-of-the-art methods

In this section, we compare our proposed method with some state-of-the-art methods on CUB-200-2011, FGVC-Aircraft, and Stanford-Cars datasets. The results are shown in Tables 2, 3 and 4, respectively. In our method, we do not use any part/bounding box annotations, considering that some of the compared methods use part/bounding box annotations or image-level labels. We use Yes or No in Tables 2, 3 and 4 to indicate whether the annotation information is used.

For CUB-200-2011 dataset, we split Table 2 into four-part over the rows: the first illustrates the results of the annotation-based methods (using part or object bounding boxes annotations); the second includes the attention-based

Table 1 The detail information about three benchmark datasets

Dataset	Class	Train	Test	Annotation		
				Label	BBox	Part
CUB-200-2011	200	5994	5794	Yes	Yes	Yes
FGVC-Aircraft	100	6667	3333	Yes	Yes	No
Stanford-Cars	196	8144	8041	Yes	Yes	No

Table 2 Comparison with state-of-the-art methods on CUB-200-2011 dataset

Methods	Annotation (Yes/No)	Backbone	Accuracy (%)
Part-RCNN [8]	Yes	AlexNet	76.37
DeepLAC [9]	Yes	AlexNet	80.30
FCAN [11]	Yes	ResNet-50	84.70
Mask-CNN [10]	Yes	ResNet-50	87.30
FCAN [11]	No	ResNet-50	84.30
RA-CNN [12]	No	VGG-19	85.30
MA-CNN[13]	No	VGG-19	86.50
B-CNN [16]	No	VGG-16	84.10
Compact B-CNN [17]	No	VGG-16	84.00
Low-Rank B-CNN [18]	No	VGG-16	84.20
Kernel-Pooling [20]	No	VGG-16	86.20
Kernel-Pooling [20]	No	ResNet-50	84.70
HLBP [21]	No	VGG-16	84.60
HBP [22]	No	VGG-16	87.10
Ours	No	ResNet-34	**87.75**

Table 3 Comparison with state-of-the-art methods on FGVC-Aircraft dataset

Methods	Annotation (yes/no)	Backbone	Accuracy (%)
RA-CNN [12]	No	VGG-19	88.20
MA-CNN [13]	No	VGG-19	89.90
B-CNN [16]	No	VGG-16	86.90
Low-Rank B-CNN [18]	No	VGG-16	87.30
Kernel-Pooling [20]	No	VGG-16	86.90
Kernel-Pooling [20]	No	ResNet-50	85.70
HLBP [21]	No	VGG-16	88.10
HBP [22]	No	VGG-16	90.30
Ours	No	ResNet-34	**91.10**

methods (using attention models to explore part information in image); the third summarizes some bilinear pooling methods and the last lists the result of bilinear pooling methods which modeling the interaction between different convolutional layers. From the results in Table 1, our proposed method outperforms all part-based methods and obtains 3.05% accuracy improvement compared with FCAN, which using ResNet-50 as the backbone network. Besides, we get 0.45% accuracy improvement in comparison with Mask-CNN. Our method also surpasses three attention-based classification methods, with 3.45%, 2.45%, and 1.25% accuracy improvements in FCAN, RA-CNN, and MA-CNN, respectively. Compared with fully bilinear pooling and its corresponding improved methods, we also get excellent performance because of using object localization module and multilayer bilinear pooling module. In addition, we also surpass two other bilinear pooling methods using the information of different convolutional layers to enhance feature discrimination.

For FGVC-Aircraft, as shown in Table 3, our proposed method obtains the best classification over the compared methods. Similarly, we also get good performance, which compared with B-CNN and a series of improvement methods. In addition, we outperform RA-CNN and MA-CNN

Table 4 Comparison with state-of-the-art methods on Stanford-Cars dataset

Methods	Annotation (yes/no)	Backbone	Accuracy (%)
FCAN [11]	Yes	ResNet-50	91.30
FCAN [11]	No	ResNet-50	89.10
RA-CNN [12]	No	VGG-19	92.50
MA-CNN [13]	No	VGG-19	92.80
B-CNN [16]	No	VGG-16	91.30
Low-Rank B-CNN [18]	No	VGG-16	90.10
Kernel-Pooling [20]	No	VGG-16	92.40
Kernel-Pooling [20]	No	ResNet-50	91.10
HBP [22]	No	VGG-16	93.70
Ours	No	ResNet-34	**93.84**

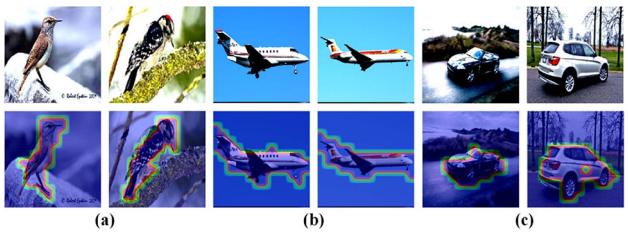


Fig. 4 Some visualization results on three benchmark datasets. **a** CUB-200-2011. **b** FGVC-Aircraft. **c** Stanford-Cars

Table 5 Comparison with different baselines on CUB-200-2011 dataset

Methods	Accuracy (%)
ResNet-34	83.55
ResNet-34+OL	85.68
ResNet-34+OL+MLBP (ours)	87.75

with 2.90% and 1.20% accuracy improvement, respectively. Finally, we also perform well compared with HLBP and HBP methods, which demonstrate that our approach can achieve competitive performance.

For Stanford-Cars, our method achieves 2.54% and 4.74% classification accuracy improvement compared with the FACN method in Table 4, which uses or without annotations to guide attention learning, with ResNet-50 backbone network. In comparison with other methods, whether RA-CNN, MA-CNN, or B-CNN with its improvement methods, our model still achieves the highest classification accuracy, highlighting its effectiveness.

4.3 Ablation studies

To further understand our method, we conduct ablation studies on some key components which play an important role in boosting the classification performance.

1. Object localization module. As an essential module in our method, whether or not to accurately locate the object in an image and eliminate background noises as much as possible is very crucial for improving classification performance. In order to verify the object location performance intuitively, we conduct qualitative experiments on three benchmark datasets with the ResNet-34 as backbone network. As shown in Fig. 4, our approach is able to wrap the edge of the object well rather than using bounding boxes which contain some background noises, such as tree branches in the right of Fig. 4a or trees in the right of Fig. 4c. In this way, relative purer features would be obtained to further help improve the representation ability of bilinear features.
2. Compared with the baselines. To explore the contribution of different modules to the classification performance. We first compare our method with the baseline, i.e., the standard ResNet-34 network is selected in this paper without using OL (object localization) module and MLBP (multi-layer bilinear pooling) module. Then, based on the standard baseline, we add the OL module to it. The third is our proposed method. It is worth noting that all experiments are completed only on the CUB-200-2011 data set. As shown in Table 5, by adding the object localization (OL) module in the baseline, we obtain accuracy improvement with 2.13% compared with the baseline. Moreover, compared with the baseline and the second combination, our method's accuracy based on both the OL module and the MLBP module further improved with 4.20% and 2.07% accuracy gain, respectively.
3. Confusion matrix analysis. To visualize classification performance and understand why some samples are easily misclassified. We also give the confusion matrix maps on three benchmark datasets. Figure 5 shows the classification confusion matrices on CUB-200-2011, FGVC-Aircraft, and Stanford-Cars, respectively, where coordinate axes denote different subcategories, and dif-

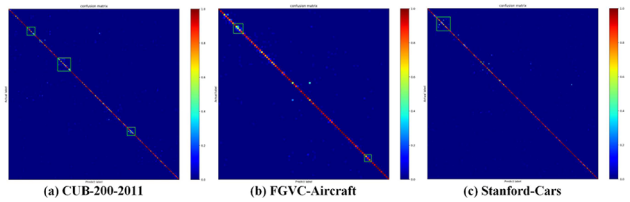


Fig. 5 Classification confusion matrices on CUB-200-2011, FGVC-Aircraft and Stanford-Cars

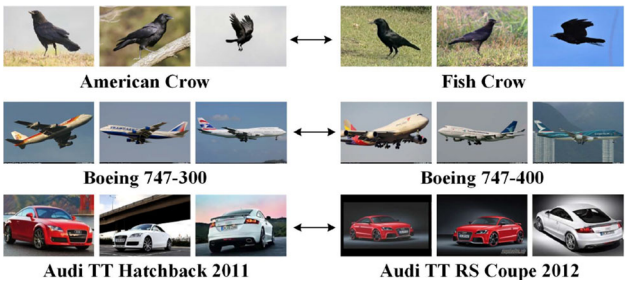


Fig. 6 Some examples of most confused subcategories on three benchmark datasets

ferent colors denote different probabilities of classification. On the diagonal of the confusion matrix, the redder the color is, the higher the probability that the subcategory is correctly classified. And the green boxes in each benchmark dataset indicate some subcategories that are easily confused because of large inter-class similarity between them. By studying the experiment datasets, we can conclude that these similar subcategories may largely belong to the same genus, e.g., American Crow and Fish Crow, or sub-models of the family, e.g., Boeing 747-300 and Boeing 747-400, Audi TT Hatchback 2011 and Audi TT RS Coupe 2012. Some confused subcategories examples are shown in Fig. 6. We can see that the subcategories in the same row are looking almost very similar that humans may not be able to recognize them correctly because they belong to the same genus or sub-models.

5 Conclusion and future work

In this paper, we have presented a new method to promote the performance of fine-grained visual classification with only using class-label information. First, an object localization module, which more flexible than bounding boxes, used to capture relatively pure features without disturbed by background noises. And then, these features are fed to a multilayer bilinear pooling module to further enhance the features discriminative capacity. Experimental results on three benchmark datasets show that our method can significantly improve the classification accuracy and show competitive performance compared with several state-of-the-art methods. In the future, we will attempt to combine object and part information to jointly enhance the discriminability of features.

Acknowledgements This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61971426.

Compliance with ethical standards

Conflict of interest We declare that we have no conflict of interest with other people or organizations.

References

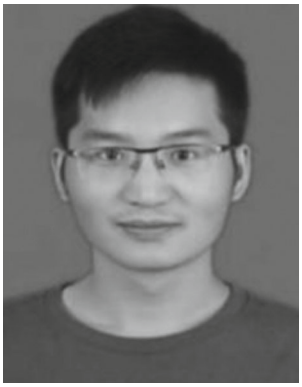
1. Wah C.B.S., Branson S., Welinder P., Perona P.: The Caltech-UCSD Birds-200-2011 Dataset, Computation and Neural Systems Technical Report 2011 (2011)
2. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision (2013)
3. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
5. Wei, X.S., Wu, J., Cui, Q.: Deep learning for fine-grained image analysis: A survey. arXiv preprint arXiv:1907.03069 (2019)
6. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
7. Sun, X., Chen, L., Yang, J.: Learning from web data using adversarial discriminative neural networks for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
8. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Lecture Notes in Computer Science (Including its Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2014)
9. Lin, D., Shen, X., Lu, C., Jia, J.: Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015)
10. Wei, X.S., Xie, C.W., Wu, J., Shen, C.: Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognit. **76**, 704–714 (2018)
11. Liu, X., Xia, T., Wang, J., Yang, Y., Zhou, F., Lin, Y.: Fully convolutional attention networks for fine-grained recognition. arXiv preprint arXiv:1603.06765 (2016)
12. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of 30th IEEE Conference Computer Visible Pattern Recognition, CVPR 2017 (2017)
13. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
14. Yao, H., Zhang, S., Zhang, Y., Li, J., Tian, Q.: Coarse-to-fine description for fine-grained visual categorization. IEEE Trans. Image Process. **25**(10), 4858–4872 (2016)

15. Tuzel, O. Porikli, F., Meer, P.: Region covariance: a fast descriptor for detection and classification. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2006)
16. Lin, T.Y., Roychowdhury, A., Maji, S.: Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1309–1322 (2018)
17. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
18. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (2017)
19. Liao, Q., Wang, D., Holewa, H., Xu, M.: Squeezed bilinear pooling for fine-grained visual categorization. In: Proceedings of the 2019 International Conference on Computer Vision Workshops. ICCVW 2019 (2019)
20. Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y., Belongie, S.: Kernel pooling for convolutional neural networks. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (2017)
21. Sun, Q., Wang, Q., Zhang, J., Li, P.: Hyperlayer Bilinear Pooling with application to fine-grained categorization and image retrieval. Neurocomputing **282**, 174–183 (2018)
22. Yu, C., Zhao, X., Zheng, Q., Zhang, P., You, X.: Hierarchical bilinear pooling for fine-grained visual recognition. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: single shot multibox detector. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2016)
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
26. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: Keypoint triplets for object detection. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
28. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2019)
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. In: 3rd International Conference on Learning Representations. ICLR 2015, Conference Track Proceedings (2015)
30. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking neural activations for fine-grained recognition. IEEE Trans. Multimed **19**(12), 2736–2750 (2017)
31. Wei, X.S., Luo, J.H., Wu, J., Zhou, Z.H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans. Image Process. **26**(6), 2868–2881 (2017)
32. Kar, P., Karnick, H.: Random feature maps for dot product kernels. In: Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR, vol. 22, pp. 583–591 (2012)
33. Pham, N., Pagh, R.: Fast and scalable polynomial kernels via explicit feature maps. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013)
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
35. Li, Y., Wang, N., Liu, J., Hou, X.: Factorized bilinear models for image recognition. In: Proceedings of the IEEE International Conference on Computer Vision (2017)

Ming Li received the M.S. degree from Central South University, Changsha, China, in 2017. He is currently working toward the Ph.D. degree in the College of Electronic Science, National University of Defense Technology, Changsha, China. His research interests include computer vision, remote sensing image processing.

Lin Lei received the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2008. She is currently a Professor with the School of Electronic Science, National University of Defense Technology. Her research interests include computer vision, remote sensing image interpretation, and data fusion.

Hao Sun received the M.S. degree in information and communication engineering and the Ph.D. degree in electronic science and technology from the National University of Defense Technology, Changsha, China, in 2009 and 2011, respectively. He is currently an Assistant Professor with the College of Electrical Science, National University of Defense Technology. His research interests include computer vision, and optical remote sensing image acquisition, and processing in general.



Xiao Li received the B.S. degree in the electrical engineering and automation from the University of Jinan, Jinan, China, in 2015, and the M.S. degree in control science and engineering from Xiangtan University, Xiangtan, China, in 2018. He is currently working toward the Ph.D. degree in information and communication engineering with the National University of Defense Technology, Changsha, China. His research interests include image processing and pattern recognition, representation and dictionary learning, and computational pathology applications.

Gangyao Kuang received the B.S. and M.S. degrees in geophysics from the Central South University of Technology, Changsha, China, in 1988 and 1991, respectively, and the Ph.D. degree in communication and information from the National University of Defense Technology, Changsha, China, in 1995. He is currently a Professor with the school of Electronic Science, National University of Defense Technology. His research interests include remote sensing, SAR image processing, change detection, SAR ground moving target indication, and classification with polarimetric SAR images.