Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition

Chaojian Yu[0000−0002−8182−6570], Xinyi Zhao[0000−0002−8499−0229], Qi Zheng[0000−0002−4351−9537], Peng Zhang[0000−0001−9428−9340], and Xinge You(✉)[0000−0003−0607−1777]

School of Electronic Information and Communications
Huazhong University of Science and Technology, Wuhan, China
{yucj,youxg}@hust.edu.cn

Abstract. Fine-grained visual recognition is challenging because it highly relies on the modeling of various semantic parts and fine-grained feature learning. Bilinear pooling based models have been shown to be effective at fine-grained recognition, while most previous approaches neglect the fact that inter-layer part feature interaction and fine-grained feature learning are mutually correlated and can reinforce each other. In this paper, we present a novel model to address these issues. First, a cross-layer bilinear pooling approach is proposed to capture the inter-layer part feature relations, which results in superior performance compared with other bilinear pooling based approaches. Second, we propose a novel hierarchical bilinear pooling framework to integrate multiple cross-layer bilinear features to enhance their representation capability. Our formulation is intuitive, efficient and achieves state-of-the-art results on the widely used fine-grained recognition datasets.

Keywords: Fine-grained visual recognition · Cross-layer interaction · Hierarchical bilinear pooling

1 Introduction

With the development of artificial intelligence, increasing demand appears to recognize subcategories of objects under the same basic-level category, e.g., brand identification for businessman, plant recognition for botanist. Thus recent years have witnessed great progress in fine-grained visual recognition, which has been widely used in applications such as automatic driving [28], expert-level image recognition [14], etc. Different from general image classification task (e.g., ImageNet classification [25]) that is to distinguish basic-level categories, fine-grained visual recognition is very challenging as subcategories tend to own small variance in object appearance and thus can only be recognized by some subtle or local differences. For example, we discriminate breeds of birds depending on the color of their back or the shape of their beak.

Motivated by the observation that local parts of object usually act a role of importance in differentiating subcategories, many methods [35,2,26,36] for fine-grained classification were developed by exploiting the parts, namely part-based

approaches. They mainly consist of two steps: firstly localize the foreground object or object parts, e.g., by utilizing available bounding boxes or part annotations, and then extract discriminative features for further classification. However, these approaches suffer from two essential limitations. First, it is difficult to ensure the manually defined parts are optimal or suitable for the final fine-grained classification task. Second, detailed part annotations incline to be time consuming and labor intensive, which is not feasible in practice. Therefore, some other approaches employ unsupervised techniques to detect possible object regions. For example, Simon and Rodner [26] proposed a constellation model to localize parts of objects, leveraging convolutional neural network (CNN) to find the constellations of neural activation patterns. Zhang *et al.* [36] proposed an automatic fine-grained image classification method, incorporating deep convolutional filters for both selection and description related to parts. These models regard CNN as part detector and obtain great improvement in fine-grained recognition. Unlike part-based methods, we treat activations from different convolution layers as responses to different part properties instead of localizing object parts explicitly, leveraging cross-layer bilinear pooling to capture inter-layer interaction of part attributes, which is proved to be useful for fine-grained recognition.

Alternatively, some researches [3,6,17,12] introduced bilinear pooling frameworks to model local parts of object. Although promising results have been reported, further improvement suffers from the following limitations. First, most existing bilinear pooling based models only take activations of the last convolution layer as representation of an image, which is insufficient to describe various semantic parts of object. Second, they neglect intermediate convolution activations, resulting in a loss of discriminative information of fine-grained categories which is significant for fine-grained visual recognition.

In this work, we present new methods to address the above challenges. We find that inter-layer part feature interaction and fine-grained feature learning are mutually correlated and can reinforce each other. To better capture the inter-layer feature relations, we propose a cross-layer bilinear pooling approach. The proposed method is efficient and powerful. It takes into account the inter-layer feature interactions while avoiding introducing extra training parameters. In contrast to other bilinear pooling based works which only utilize feature from one single convolution layer, our architecture exploits the interaction of part features from multiple layers, which is useful for fine-grained feature learning. Besides, our framework is highly consistent with the human coarse-to-fine perception, the visual hierarchy segregates local and global features in cortical areas V4 based on spatial differences and builds a temporal dissociation of the neural activity [20]. We find that our cross-layer bilinear model is closer to the unique architecture of cortical areas V4 for processing spatial information.

It is well known that information loss exists in the propagation of CNNs. In order to minimize the loss of information that is useful for fine-grained recognition, we propose a novel hierarchical bilinear pooling framework to integrate multiple cross-layer bilinear features to enhance their representation power. To make full use of the intermediate convolution layer activations, all cross-layer

bilinear features are concatenated before the final classification. Note that the features from different convolution layer are complementary, they contribute to discriminative feature learning. Thus the proposed network benefits from the mutual reinforcement between inter-layer feature interaction and fine-grained feature learning. Our contributions are summarized as follows:

- We develop a simple but effective cross-layer bilinear pooling technique that simultaneously enables the inter-layer interaction of features and the learning of fine-grained representation in a mutually reinforced way.
- Based on cross-layer bilinear pooling, we propose a hierarchical bilinear pooling framework to integrate multiple cross-layer bilinear modules to obtain the complementary information from intermediate convolution layers for performance boost.
- We conduct comprehensive experiments on three challenging datasets (CUB Birds, Stanford Cars, FGVC-Aircraft), and the results demonstrate the superiority of our method.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed method. Section 4 provides experiments as well as result analysis, followed by conclusion in Section 5.

2 Related Work

In the following, we briefly review previous works from the two viewpoints of interest due to their relevance to our work, including fine-grained feature learning and feature fusion in CNNs.

2.1 Fine-Grained Feature Learning

Feature learning plays an important and fundamental role in fine-grained recognition. Since the differences between subcategories are subtle and local, capturing global semantic information with merely fully connected layers limits the representation capacity of a framework, and hence restricts further promotion of final recognition [1]. To better model subtle difference for fine-grained categories, Lin *et al.* [17] proposed a bilinear structure to aggregate the pairwise feature interactions by two independent CNNs, which adopted outer product of feature vectors to produce a very high-dimensional feature for quadratic expansion. Gao *et al.* [6] applied Tensor Sketch [23] to approximate the second-order statistics and to reduce feature dimension. Kong *et al.* [12] adopted low-rank approximation to the covariance matrix and further reduced the computational complexity. Yin *et al.* [4] aggregated higher-order statistics by iteratively applying the Tensor Sketch compression to the features. The work in [22] utilized bilinear convolutional neural network as baseline model and adopted an ensemble learning method to incorporate boosting weights. In [16], matrix square-root normalization was proposed and proved to be complementary to existing normalization. However,

these approaches only consider the feature from single convolution layer, which is insufficient to capture various discriminative parts of object and model the subtle differences among subcategories. The method we propose overcome this limitation via integrating inter-layer feature interaction and fine-grained feature learning in a mutually reinforced manner and is therefore more effective.

2.2 Feature Fusion in CNNs

Due to the success of deep learning, CNNs have emerged as general-purpose feature extractors for a wide range of visual recognition tasks. While feature maps from single convolution layer are insufficient for finer-grained tasks, thus some recent works [3,7,19,33] attempt to investigate the effectiveness of exploiting feature from different convolution layers within a CNN. For example, Hariharan *et al.* [7] considered the feature maps from all convolution layers, allowing finer grained resolution for localization tasks. Long *et al.* [19] combined the finer-level and higher-level semantic feature from different convolution layers for better segmentation. Xie *et al.* [33] proposed a holistically-nested framework where the side outputs are added after lower convolution layers to provide deep supervision for edge detection. The very recent work [3] concatenated the activation maps from multiple convolution layers to model the interaction of part features for fine-grained recognition. However, simply cascading the feature map introduces lots of training parameters and even fails to capture inter-layer feature relations when incorporating with more intermediate convolution layers. Instead, our network treats each convolution layer as attribute extractor for different object parts and models their interactions in an intuitive and effective way.

3 Hierarchical Bilinear Model

In this section, we develop a hierarchical bilinear model to overcome those limitations mentioned above. Before presenting our hierarchical bilinear model, we first introduce the general formulation of factorized bilinear pooling for fine-grained image recognition in Sect. 3.1. Based on this, we propose a cross-layer bilinear pooling technique to jointly learn the activations from different convolution layers in Sect. 3.2, which captures the cross-layer interaction of information and leads to better representation capability. Finally, our hierarchical bilinear model combining multiple cross-layer bilinear modules generates finer part description for better fine-grained recognition in Sect. 3.3.

3.1 Factorized Bilinear Pooling

Factorized bilinear pooling has been applied to visual question answer task, Kim *et al.* [11] proposed factorized bilinear pooling using Hadamard product for an efficient attention mechanism of multimodal learning. Here we introduce the basic formulation of factorized bilinear pooling technique for the task of fine-grained image recognition. Suppose an image I is filtered by a CNN and the

output feature map of a convolution layer is $X \in \mathbb{R}^{h \times w \times c}$ with height h, width w and channels c, we denote a c dimensional descriptor at a spatial location on X as $\mathbf{x} = [x_1, x_2, \cdots, x_c]^T$. Then the full bilinear model is defined by

$$z_i = \mathbf{x}^T W_i \mathbf{x} \tag{1}$$

Where $W_i \in \mathbb{R}^{c \times c}$ is a projection matrix, z_i is the output of the bilinear model. We need to learn $\mathbf{W} = [W_1, W_2, \cdots, W_o] \in \mathbb{R}^{c \times c \times o}$ to obtain a o dimensional output $\mathbf{z}$. According to matrix factorization in [24], the projection matrix W_i in Eq. (1) can be factorized into two one-rank vectors

$$z_i = \mathbf{x}^T W_i \mathbf{x} = \mathbf{x}^T U_i V_i^T \mathbf{x} = U_i^T \mathbf{x} \circ V_i^T \mathbf{x} \tag{2}$$

where $U_i \in \mathbb{R}^c$ and $V_i \in \mathbb{R}^c$. Thus the output feature $\mathbf{z} \in \mathbb{R}^o$ is given by

$$\mathbf{z} = P^T (U^T \mathbf{x} \circ V^T \mathbf{x}) \tag{3}$$

where $U \in \mathbb{R}^{c \times d}$ and $V \in \mathbb{R}^{c \times d}$ are projection matrices, $P \in \mathbb{R}^{d \times o}$ is the classification matrix, $\circ$ is the Hadamard product and d is a hyperparameter deciding the dimension of joint embeddings.

3.2 Cross-Layer Bilinear Pooling

Fine-grained subcategories tend to share similar appearances and can only be discriminated by subtle differences in the attributes of local part, such as color, shape, or length of beak for birds. Bilinear pooling, which captures the pairwise feature relations, is an important technique for fine-grained recognition. However, most bilinear models only focus on learning the features from single convolution layer while completely ignoring the cross-layer interaction of information. Activations of individual convolution layer are incomplete since there are multiple attributes in each object part which can be crucial in differentiating subcategories.

Actually in most cases, we need to simultaneously consider multi-factor of part feature to determine the category for a given image. Therefore, to capture finer grained part feature, we develop a cross-layer bilinear pooling approach that treats each convolution layer in a CNN as part attributes extractor. After that the features from different convolution layers are integrated by element-wise multiplication to model the inter-layer interaction of part attributes. Accordingly, Eq. (3) can be rewritten as

$$\mathbf{z} = P^T (U^T \mathbf{x} \circ V^T \mathbf{y}) \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent local descriptors from different convolution layers at the same spatial location.

It is worth noting that the features from different convolution layers are expanded into high-dimensional space by independent linear mappings. It is expected that the convolution activations and project activations encode global

and local feature of object respectively, as shown in Fig. 3. It is highly consistent with the human coarse-to-fine perception: human and non-human primates often see the global "gist" of an object, or a scene, before discerning local detailed features [20]. For example, neurons in macaque inferotemporal cortex that are active during face perception encode the global facial category is earlier than they begin to encode finer information such as identity or expression.

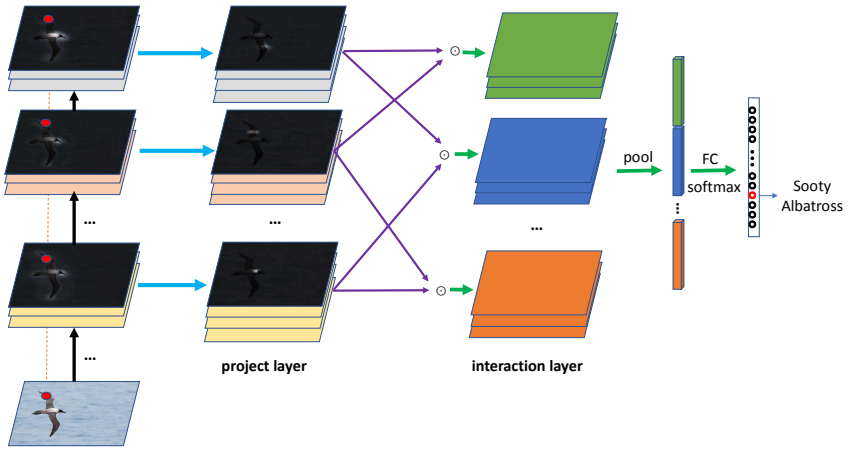

Fig. 1. Illustration of our Hierarchical Bilinear Pooling (HBP) network architecture for fine-grained recognition. The bottom image is the input, and above it are the feature maps of different layers in the CNN. First the features from different layers are expanded into a high-dimensional space via independent linear mapping to capture attributes of different object parts and then integrated by element-wise multiplication to model the inter-layer interaction of part attributes. After that sum pooling is performed to squeeze the high-dimensional features into compact ones. Note that we obtain the visual activation maps above by computing the response of sum-pooled feature vector on every single spatial location.

3.3 Hierarchical Bilinear Pooling

Cross-layer bilinear pooling proposed in Sect. 3.2 is intuitive and effective, as it has superior representation capacity than traditional bilinear pooling models without increasing training parameters. This inspires us that exploiting the inter-layer feature interactions among different convolution layers is beneficial for capturing the discriminative part properties between fine-grained subcategories. Therefore, we extend the cross-layer bilinear pooling to integrate more intermediate convolution layers, which further enhances the representation capacity of features. In this section, we propose a generalized Hierarchical Bilinear

Pooling (HBP) framework to incorporate more convolutional layer features by cascading multiple cross-layer bilinear pooling modules.

Specifically, we divide the cross-layer bilinear pooling module into interaction stage and classification stage, which formulates as follows

$$\mathbf{z}_{int} = U^T\mathbf{x} \circ V^T\mathbf{y} \tag{5}$$

$$\mathbf{z} = P^T\mathbf{z}_{int} \in \mathbb{R}^o \tag{6}$$

To better model inter-layer feature interactions, the interaction feature of the HBP model is obtained by concatenating multiple $\mathbf{z}_{int}$ of the cross-layer bilinear pooling modules. Thus we can derive final output of the HBP model by

$$\mathbf{z}_{HBP} = HBP(\mathbf{x},\ \mathbf{y},\ \mathbf{z},\ \cdots) = P^T\mathbf{z}_{int} \tag{7}$$

$$= P^T concat(U^T\mathbf{x} \circ V^T\mathbf{y},\ U^T\mathbf{x} \circ S^T\mathbf{z},\ V^T\mathbf{y} \circ S^T\mathbf{z},\ \cdots) \tag{8}$$

where P is the classification matrix, $U, V, S, \ldots$ are the projection matrices of convolution layer feature $\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$ respectively. The overall flowchart of the HBP framework is illustrated in Fig. 1.

4 Experiments

In this section, we evaluate the performance of HBP model for fine-grained recognition. The datasets and implementation details of HBP are firstly introduced in Sect. 4.1. Model configuration studies are performed to investigate the effectiveness of each component in Sect. 4.2. Comparison with state-of-the-art methods is provided in Sect. 4.3. Finally in Sect. 4.4, qualitative visualization is present to intuitively explain our model.

4.1 Datasets and Implementation Details

Datasets: We conduct experiments on three widely used datasets for fine-grained image recognition, including Caltech-UCSD Birds (CUB-200-2011) [30], Stanford Cars [15] and FGVC-Aircraft [21]. The detailed statistics with category numbers and data splits are summarized in Table 1. Note that we only use category labels in our experiments.

Table 1. Summary statistics of datasets

Datasets	#Category	#Training	#Testing
CUB-200-2011 [30]	200	5994	5794
Stanford Cars [15]	196	8144	8041
FGVC-Aircraft [21]	100	6667	3333

Implementation Detail: For fair comparison with other state-of-the-art methods, we evaluate our HBP with VGG-16 [27] baseline model pretrained on ImageNet classification dataset [25], removing the last three fully-connected layers and inserting all the components in our framework. It is worth noting that our HBP can be also applied to other network structures, such as Inception [29] and ResNet [8]. The size of input image is 448×448. Our data augmentation follows the commonly used practice, i.e., random sampling (crop 448×448 from $512 \times S$ where S is the largest image side) and horizontal flipping are utilized during training, and only center cropping is involved during inference.

We initially train only the classifiers by logistic regression, and then fine-tune the whole network using stochastic gradient descent with a batch size of 16, momentum of 0.9, weight decay of 5×10^{-4} and a learning rate of 10^{-3}, periodically annealed by 0.5. All experiments are implemented with the Caffe toolbox [10] and performed on a server with Titan X GPUs. The source code and trained model will be made available at https://github.com/ChaojianYu/Hierarchical-Bilinear-Pooling

4.2 Configurations of Hierarchical Bilinear Pooling

Cross-layer bilinear pooling (CBP) has a user-define projection dimension d. To investigate the impact of d and to validate the effectiveness of the proposed framework, we conduct extensive experiments on the CUB-200-2011 [30] dataset, with results summarized in Fig. 2. Note that we utilize $relu5\_3$ in FBP, $relu5\_2$ and $relu5\_3$ in CBP, $relu5\_1$, $relu5\_2$ and $relu5\_3$ in HBP to obtain the results in Fig. 2 and we also provide quantitative experiments about the choice of layers in the following. We focus on $relu5\_1$, $relu5\_2$ and $relu5\_3$ in VGG-16 [27] as they contain more part semantic information compared with shallower layers.

In Fig. 2, we compare the performance of CBP with the general factorized bilinear pooling model, namely FBP. Futhermore, we explore HBP with combination of multiple layers. Finally, we analyze the impact factors of hyperparameter d. We can draw the following significant conclusions from Fig. 2

- First, under the same d, our CBP significantly outperforms FBP, which indicates that the discriminative power can be enhanced by the inter-layer interaction of features.
- Second, HBP further outperforms CBP, which demonstrates the efficacy of activations from intermediate convolution layers for fine-grained recognition. This can be explained by the fact that information loss exists in the propagation of CNNs, thus discriminative features crucial for fine-grained recognition may be lost in intermediate convolution layers. In contrast to CBP, our HBP takes more feature interactions of intermediate convolution layers into consideration and is therefore more robust, since HBP has presented the best performance. In the following experiments, HBP is used to compare with other state-of-the-art methods.
- Third, when d varies from 512 to 8192, increasing d leads to higher accuracy for all models and HBP is saturated with $d = 8192$. Therefore, $d = 8192$

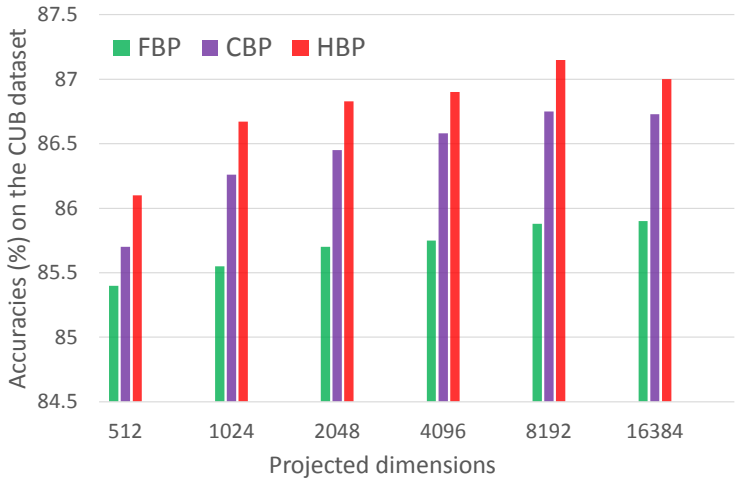

Fig. 2. Classification accuracy on the CUB dataset. Comparison of general Factorized Bilinear Pooling (FBP), Cross-layer Bilinear Pooling (CBP) and Hierarchical Bilinear Pooling (HBP) with various projection dimensions.

is used for HBP in our following experiments in consideration of feature dimension, computational complexity as well as accuracy.

We then provide quantitative experiments on the CUB-200-2011 [30] dataset to analyze the impact factor of layers. The accuracies in Table 2 are obtained under the same embedding dimension ($d = 8192$). We consider the combination of different layers for CBP and HBP. The results demonstrate that the performance gain of our framework comes mainly from the inter-layer interaction and multiple layers combination. As the HBP-3 already presents the best performance, thus we utilize $relu5\_1$, $relu5\_2$ and $relu5\_3$ in all the experiments in Sect. 4.3.

Table 2. Quantitative analysis results on CUB-200-2011 dataset

Method	FBP	CBP			HBP		
	FBP-1[a]	CBP-1[b]	CBP-2[c]	CBP-3[d]	HBP-1[e]	HBP-2[f]	HBP-3[g]
Accuracy	85.70	86.75	86.85	86.67	86.78	86.91	87.15

[a] $relu5\_3 * relu5\_3$.
[b] $relu5\_3 * relu5\_2$.
[c] $relu5\_3 * relu5\_1$.
[d] $relu5\_3 * relu4\_3$.
[e] $relu5\_3 * relu5\_2 + relu5\_3 * relu5\_1$.
[f] $relu5\_3 * relu5\_2 + relu5\_3 * relu5\_1 + relu5\_3 * relu4\_3$.
[g] $relu5\_3 * relu5\_2 + relu5\_3 * relu5\_1 + relu5\_2 * relu5\_1$.

We also compare our cross-layer integration with hypercolumn [3] based feature fusion. For fair comparison, we re-implement hypercolumn as the feature concatenation of $relu5\_3$ and $relu5\_2$, followed by factorized bilinear pooling (denoted as HyperBP) under the same experimental settings. Table 3 shows that our CBP obtains slightly better result than HyperBP with nearly 1/2 parameters, which again indicates that our integration framework is more effective in capturing inter-layer feature relations. This is not surprising since our CBP is consistent with human perception to some extent. On the contrary of the HyperBP, which obtains even worse result when integrating more convolution layer activations [3], our HBP is able to capture the complementary information within intermediate convolution layers and achieves an obvious improvement in recognition accuracy.

Table 3. Classification accuracy on the CUB dataset and model sizes of different feature integrations

Method	Accuracy	Model Size
HyperBP	86.60	18.4M
CBP	86.75	10.0M
HBP	**87.15**	17.5M

4.3 Comparison with State-of-the-art

Results on CUB-200-2011. CUB dataset provides ground-truth annotations of bounding boxes and parts of birds. The only supervised information we use is the image level class label. The classification accuracy on CUB-200-2011 is summarized in Table 4. The table is split into three parts over the rows: the first summarizes the annotation-based methods (using object bounding boxes or part annotations); the second includes the unsupervised part-based methods; the last illustrates the results of pooling-based methods.

From results in Table 4, we can see that PN-CNN [2] uses strong supervision of both human-defined bounding box and ground truth parts. SPDA-CNN [35] uses ground truth parts and B-CNN [17] uses bounding box with very high-dimensional feature representation (250K dimensions). The proposed HBP($relu5\_3 + relu5\_2 + relu5\_1$) achieves better result compared with PN-CNN [2], SPDA-CNN [35] and B-CNN [17] even without bbox and part annotation, which demonstrates the effectiveness of our model. Compared with STN [9] which uses stronger inception network as baseline model, we obtain a relative accuracy gain with 3.6% by our HBP($relu5\_3+relu5\_2+relu5\_1$). We even surpass RA-CNN [5] and MA-CNN [37], which are the recently-proposed state-of-the-art unsupervised part-based methods, with 2.1% and 0.7% relative accuracy gains, respectively. Compared with the baselines of pooling-based model B-CNN [17], CBP [6] and LRBP [12], the superior result that we achieve mainly benefits from

Table 4. Comparison results on CUB-200-2011 dataset. Anno. represents using bounding box or part annotation

Method	Anno.	Accuracy
SPDA-CNN [35]	√	85.1
B-CNN [17]	√	85.1
PN-CNN [2]	√	85.4
STN [9]		84.1
RA-CNN [5]		85.3
MA-CNN [37]		86.5
B-CNN [17]		84.0
CBP [6]		84.0
LRBP [12]		84.2
HIHCA [3]		85.3
Improved B-CNN [16]		85.8
BoostCNN [22]		86.2
KP [4]		86.2
FBP($relu5\_3$)		85.7
CBP($relu5\_3 + relu5\_2$)		86.7
HBP($relu5\_3 + relu5\_2 + relu5\_1$)		**87.1**

the inter-layer interaction of feature and the integration of multiple layers. We also surpass BoostCNN [22] which boosts multiple bilinear networks trained at multiple scales. Although HIHCA [3] proposes similar ideas to model feature interaction for fine-grained recognition, our model can achieve higher accuracy because of the mutual reinforcement framework for inter-layer feature interaction and discriminative feature learning. Note that HBP($relu5\_3+relu5\_2+relu5\_1$) outperforms CBP($relu5\_3 + relu5\_2$) and FBP($relu5\_3$), which indicates that our model can capture the complementary information among layers.

Results on Stanford Cars. The classification accuracy on Stanford Cars is summarized in Table 5. Different car parts are discriminative and complementary, thus object and part localization may play a significant role here [34]. Although our HBP has no explicit part detection, we achieve the best result among state-of-the-art methods. Relying on inter-layer feature interaction learning, we even surpass PA-CNN [13] by 1.2% relative accuracy gains, which uses human-defined bounding box. We can observe significant improvement compared with unsupervised part-based method MA-CNN [37]. Our HBP is also better than pooling-based methods BoostCNN [22] and KP [4].

Results on FGVC-Aircraft. Different aircraft models are difficult to be recognized, due to subtle differences, e.g., one may be able to distinguish them by counting the number of windows in the model. The classification accuracy on FGVC-Aircraft is summarized in Table 6. Still, our model achieves the highest

Table 5. Comparison results on Stanford Cars dataset. Anno. represents using bounding box

Method	Anno.	Accuracy
FCAN [18]	√	91.3
PA-CNN [13]	√	92.6
FCAN [18]		89.1
RA-CNN [5]		92.5
MA-CNN [37]		92.8
B-CNN [17]		90.6
LRBP [12]		90.9
HIHCA [3]		91.7
Improved B-CNN [16]		92.0
BoostCNN [22]		92.1
KP [4]		92.4
HBP		**93.7**

classification accuracy among all the methods. We can observe stable improvement compared with annotation-based method MDTP [32], part learning-based method MA-CNN [37], and pooling-based BoostCNN [22], which highlights the efficacy and robustness of the proposed HBP model.

Table 6. Comparison results on FGVC-Aircraft dataset. Anno. represents using bounding box

Method	Anno.	Accuracy
MG-CNN [31]	√	86.6
MDTP [32]	√	88.4
RA-CNN [5]		88.2
MA-CNN [37]		89.9
B-CNN [17]		86.9
KP [4]		86.9
LRBP [12]		87.3
HIHCA [3]		88.3
Improved B-CNN [16]		88.5
BoostCNN [22]		88.5
HBP		**90.3**

4.4 Qualitative Visualization

To better understand our model, we visualize the model response of different layers in our fine-tuned network on different datasets. We obtain the activation maps by computing the magnitude of feature activations averaged across

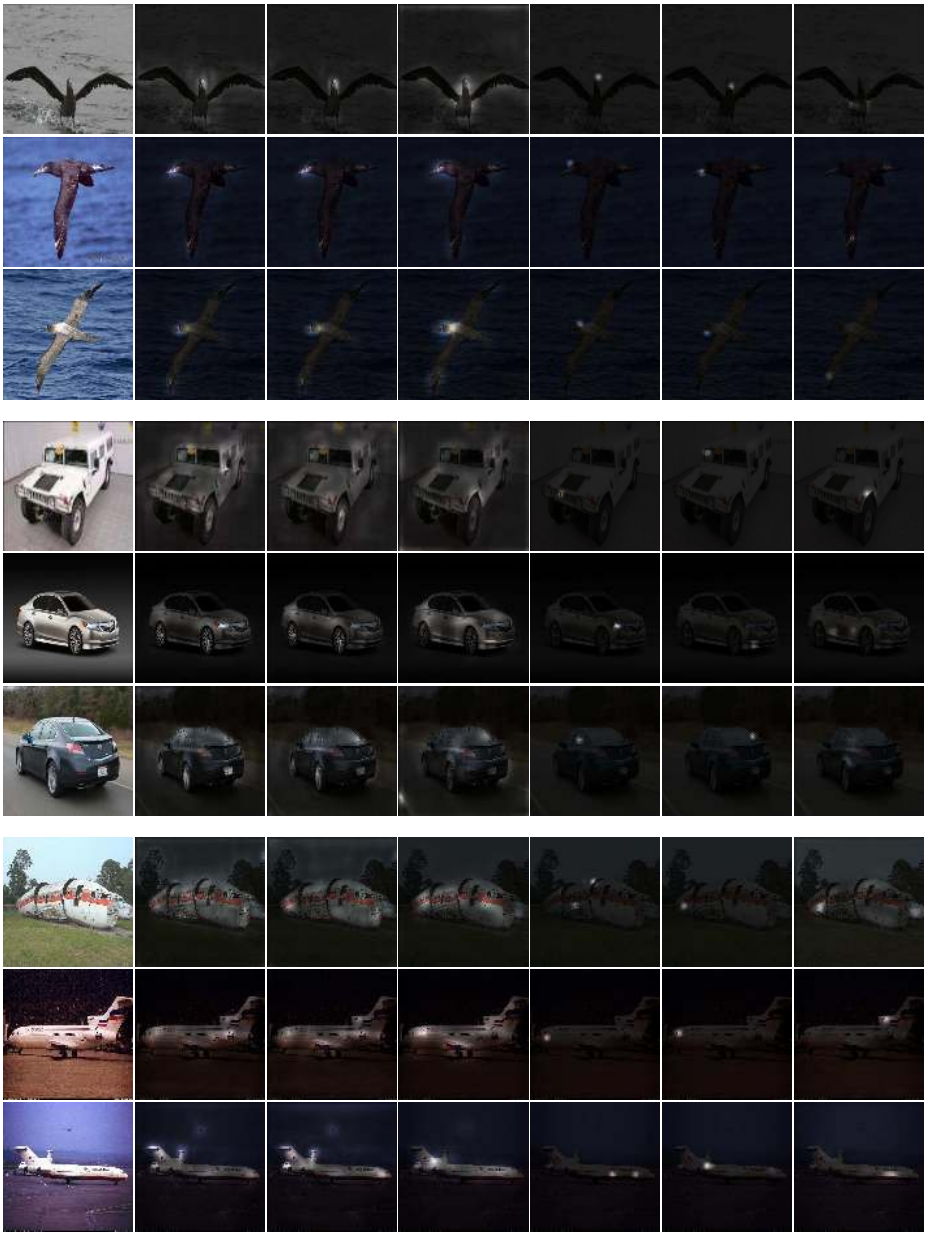

Fig. 3. Visualization of model response of different layers on the CUB, Cars and Aircraft datasets. It can be seen that our model tend to ignore features in the cluttered background and focus on the most discriminative parts of object.

channel. In Fig. 3, we show some randomly selected images from three different datasets and their corresponding visualizations.

The visualizations all suggest that the proposed model is capable of ignoring cluttered backgrounds and tends to activate strongly on highly specific semantic parts. The highlighted activation regions in $project5\_1$, $project5\_2$ and $project5\_3$ are strongly related to semantic parts, such as heads, wings and breast in CUB; front bumpers, wheels and lights in Cars; cockpit, tail stabilizers and engine in Aircraft. These parts are crucial to distinguish the category. Moreover, our model is highly consistent with the human perception that resolve the fine details when perceive scenes or objects. In Fig. 3, we can see that the convolution layers ($relu5\_1, relu5\_2, relu5\_3$) provide a rough localization of target object. Based on this, the projection layers ($project5\_1, project5\_2, project5\_3$) further determine essential parts of the object, which distinguish its category by successive interaction and integration of different part features. The process is consistent with the coarse-to-fine nature of human perception [20] inspired by the Gestalt dictum that the "whole" is prior to the "parts" and it also provides an intuitive explanation as to why our framework can model subtle and local differences between subcategories without explicit part detection.

5 Conclusions

In this paper, we propose a hierarchical bilinear pooling approach to fuse multi-layer features for fine-grained recognition, which combines inter-layer interactions and discriminative feature learning in a mutually-reinforced way. The proposed network requires no bounding box/part annotations and can be trained end-to-end. Extensive experiments on birds, cars and aircrafts demonstrate the effectiveness of our framework. In the future, we will conduct extended research on two directions, i.e., how to effectively fuse more layer features to obtain part representation at multiple scales, and how to merge effective methods for parts localization to learn better fine-grained representation.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (No.61772220, 61571205), in part by National Key Technology Research and Development Program of Ministry of Science and Technology of China (No.2015BAK36B00), in part by the Technology Innovation Program of Hubei Province (No.2017AAA017), in part by the Key Program for International S&T Cooperation Projects of China (No.2016YFE0121200).

References

1. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE international conference on computer vision. pp. 1269–1277 (2015)

2. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952 (2014)
3. Cai, S., Zuo, W., Zhang, L.: Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 511–520 (2017)
4. Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y., Belongie, S.: Kernel pooling for convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR) (2017)
5. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Conf. on Computer Vision and Pattern Recognition (2017)
6. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 317–326 (2016)
7. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 447–456 (2015)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
10. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 675–678. ACM (2014)
11. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
12. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7025–7034. IEEE (2017)
13. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-grained recognition without part annotations. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 5546–5555. IEEE (2015)
14. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision. pp. 301–320. Springer (2016)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. pp. 554–561. IEEE (2013)
16. Lin, T.Y., Maji, S.: Improved bilinear pooling with cnns. arXiv preprint arXiv:1707.06772 (2017)
17. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1449–1457 (2015)
18. 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)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

20. Lu, Y., Yin, J., Chen, Z., Gong, H., Liu, Y., Qian, L., Li, X., Liu, R., Andolina, I.M., Wang, W.: Revealing detail along the visual hierarchy: neural clustering preserves acuity from v1 to v4. Neuron **98**(2), 417–428 (2018)
21. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
22. Moghimi, M., Belongie, S.J., Saberian, M.J., Yang, J., Vasconcelos, N., Li, L.J.: Boosted convolutional neural networks. In: BMVC (2016)
23. Pham, N., Pagh, R.: Fast and scalable polynomial kernels via explicit feature maps. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 239–247. ACM (2013)
24. Rendle, S.: Factorization machines. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 995–1000. IEEE (2010)
25. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
26. Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1143–1151 (2015)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Sochor, J., Herout, A., Havel, J.: Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3006–3015 (2016)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
30. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
31. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 2399–2406. IEEE (2015)
32. Wang, Y., Choi, J., Morariu, V.I., Davis, L.S.: Mining discriminative triplets of patches for fine-grained classification. arXiv preprint arXiv:1605.01130 (2016)
33. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1395–1403 (2015)
34. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3973–3981 (2015)
35. Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., Metaxas, D.: Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1143–1152 (2016)
36. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1134–1142 (2016)
37. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Int. Conf. on Computer Vision (2017)