P-CNN: Part-Based Convolutional Neural Networks for Fine-Grained Visual Categorization

Junwei Han, Xiwen Yao, Gong Cheng, Xiaoxu Feng, Dong Xu

Abstract—This paper proposes an end-to-end fine-grained visual categorization system, termed Part-based Convolutional Neural Network (P-CNN), which consists of three modules. The first module is a Squeeze-and-Excitation (SE) block, which learns to recalibrate channel-wise feature responses by emphasizing informative channels and suppressing less useful ones. The second module is a Part Localization Network (PLN) used to locate distinctive object parts, through which a bank of convolutional filters are learned as discriminative part detectors. Thus, a group of informative parts can be discovered by convolving the feature maps with each part detector. The third module is a Part Classification Network (PCN) that has two streams. The first stream classifies each individual object part into image-level categories. The second stream concatenates part features and global feature into a joint feature for the final classification. In order to learn powerful part features and boost the joint feature capability, we propose a Duplex Focal Loss used for metric learning and part classification, which focuses on training hard examples. We further merge PLN and PCN into a unified network for an end-to-end training process via a simple training technique. Comprehensive experiments and comparisons with state-of-the-art methods on three benchmark datasets demonstrate the effectiveness of our proposed method.

Index Terms—Part Localization Network, Part Classification Network, Duplex Focal Loss, Fine-grained Visual Categorization

1 Introduction

Fine-grained visual categorization (FGVC), which aims to classify subordinate categories, such as bird species, car models or aircraft variants, has been attracting increasing research interest [1-35]. The FGVC task is highly challenging even for powerful deep learning algorithms [6, 7, 10, 12, 16, 17, 19, 23-26, 28, 33, 36-42] because those categories can only be distinguished by subtle visual appearance differences located at some critical parts of foreground objects along with the presence of scene cluttering. In addition, large appearance variances in scale, pose, lighting, and viewpoint of objects within the same subordinate category further add to the complexity of the problem. Therefore, localizing and extracting discriminative features not only from global images but also, more importantly, from distinctive parts play a crucial role in improving FGVC performance.

Generally, most existing methods [5-7, 12, 15, 17, 18, 24, 28, 31, 43-46] follow the common pipeline in which they first find the objects or their distinctive parts and then they discriminate which subcategory each image belongs to according to the extracted discriminative features. For example, a fully convolutional part localization network is designed in [17] to locate multiple object parts. Despite of its effectiveness, the method heavily relies on manually-labeled strong part annotations, which greatly limits the usability and scalability of the fine-grained recognition algorithms. In contrast, some weakly supervised FGVC frameworks [7, 24, 28, 31, 43, 44] attempt to learn part localization networks under weak supervision that only require image category labels in the training phase and have been receiving increasing attention in recent years. A representative work is the Channel Grouping Network method proposed in [43], which clusters spatially-correlated feature channels for generating multiple discriminative parts. Another interesting work is Spatial Transformer Networks [7], which builds upon a differentiable attention mechanism to exploit a localization network to fit the transformation parameters for identifying a series of discriminative regions. Although promising part localization results are reported, some part localization networks are computationally expensive as heavy operations are involved for learning localization parameters. Other networks exploit the alternating optimization approach, so the networks cannot be trained in an end-to-end fashion together with part classification network, which will cause information loss and inefficiency.

Apart from locating distinctive parts, it is critical to learn discriminative features for fine grained image recognition. Although great processes of FGVC have been achieved due to the impressive feature learning power of Convolutional Neural Networks (CNNs), the large intra-class variances (pose, scales, color, etc.) and subtle inter-class visual differences, as shown in Fig.1, still remain as the major challenges. These challenges are supposed to degrade the performance of fine-grained image recognition. In this situation, it is highly desirable to learn an efficient and accurate part localization network together with more powerful CNN feature representations that have small within-class scatter while maintaining big between-class separation.

- *Junwei Han is with the School of Automation, Northwestern Polytechnical University, Xi'an, China. E-mail: junweihan2010@gmail.com.*
- *Xiwen Yao is with the School of Automation, Northwestern Polytechnical University, Xi'an, China. E-mail: yaoxiwen@nwpu.edu.cn.*
- *Gong Cheng (Corresponding author) is with the School of Automation, Northwestern Polytechnical University, Xi'an, China. E-mail: gcheng@nwpu.edu.cn.*
- *Xiaoxu Feng is with the School of Automation, Northwestern Polytechnical University, Xi'an, China. E-mail: fengxiaox@mail.nwpu.edu.cn.*
- *Dong Xu is with School of Electrical and Information Engineering, University of Sydney, NSW, 2006, Australia. Email: dong.xu@sydney.edu.au.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2933510, IEEE Transactions on Pattern Analysis and Machine Intelligence

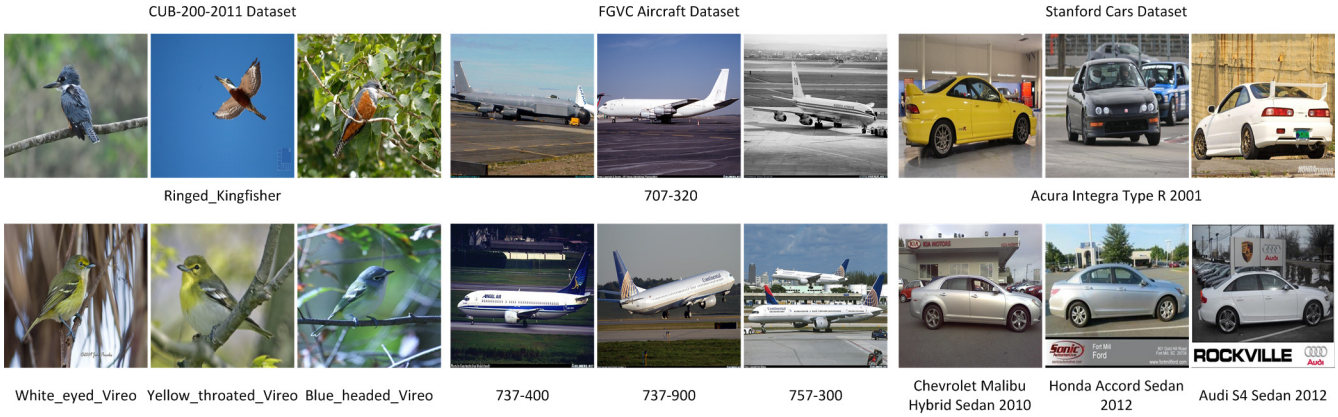


Figure 1: Illustration of challenges in fine-grained visual classification: the first row shows the large variances in the same subcategory and the second row shows the small variances among different subcategories.

To address the aforementioned limitations and also boost the state-of-the-arts of FGVC, *in this paper we propose a novel, unified, fine-grained visual categorization framework, termed Part-based Convolutional Neural Network (P-CNN). P-CNN consists of three modules, namely Squeeze-and-Excitation (SE) block, Part Localization Network (PLN) and Part Classification Network (PCN), used for feature recalibration, distinctive part localization, and image classification, respectively. Our main contributions are summarized as follows. 1) The PLN is trained in an unsupervised learning fashion, through which a set of convolutional filters are learned as discriminative part detectors. Thus, a group of informative object parts can be discovered by convolving the recalibrated feature maps with each learned part detector. 2) The PCN classifies each individual part into image-level categories and meanwhile combines part-level local features and image-level global feature for a final classification with a novel Duplex Focal Loss (DFL), which reshapes the metric learning loss and part classification loss functions to focus training on hard examples and thus down-weighting easy examples for discriminative feature learning and robust classification. 3) The PLN and PCN can be merged into a unified network for an end-to-end training by sharing the full-image convolutional features that alternates between the training for part localization task and the training for part classification task. Comprehensive experiments and comparisons with state-of-the-arts on three widely-used FGVC datasets demonstrate the superiority of our P-CNN model.*

This paper is organized as follows. Section 2 gives a brief review of related work. Section 3 firstly gives an overview of the proposed method and then describes the details of the framework. Section 4 presents comprehensive experimental results and analysis. Section 5 concludes this paper.

2 Related Work

Fine-grained visual categorization has been extensively studied for the last few decades. In this section, we will review the related works from two aspects including discriminative feature learning and part localization.

2.1 Discriminative Feature Learning

Compared with traditional low-level handcrafted features based methods [1, 3, 47], deep CNN models have significantly boosted the performance of FGVC due to the impressive feature learning power capability.

To better characterize subtle differences from fine-grained categories, the bilinear CNN structure [10, 30] and a set of its extensions are proposed. The basic idea is to exploit two independent CNNs to compute the pairwise feature interactions for capturing the image local differences and has achieved the state-of-the-art results on several public datasets for fine-grained classification [10, 27, 30, 39, 48-50]. Besides, in [45], the spatial relationship of discriminative regions is encoded in a recursive way to generate spatially expressive representations while the work in [37] proposes to encode deep CNN filter responses via spatially weighted combination of Fisher Vectors. The work in [21, 41] proposes to use object/part masks for selecting useful and meaningful convolutional descriptors to perform fine-grained image classification and retrieval.

Several recent works attempt to take advantage of different information to enhance feature representations. In [42], two complementary part-level and object-level visual descriptions are combined while the work in [51] builds up on a dynamic internal representation by incrementally combining the information of objects from coarse-to-fine granularity. Similarly, [22] proposes a multi-level coarse-to-fine object description by feeding the information from original image, object bounding box, foreground object segmentation, and part segmentation into the CNN models and finally concatenating their outputs. In addition to using only the vision information, recent methods [19, 29, 52] propose to further exploit rich prior information from either structured knowledge bases or unstructured texts to jointly learn deep embedded representations, which is demonstrated to be useful for distinguishing the subtle differences among subordinate categories.

An alternative stream of research is to explore metric learning to learn discriminative representations for fine-grained recognition [11, 16, 38, 44, 53]. Metric learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2933510, IEEE Transactions on Pattern Analysis and Machine Intelligence

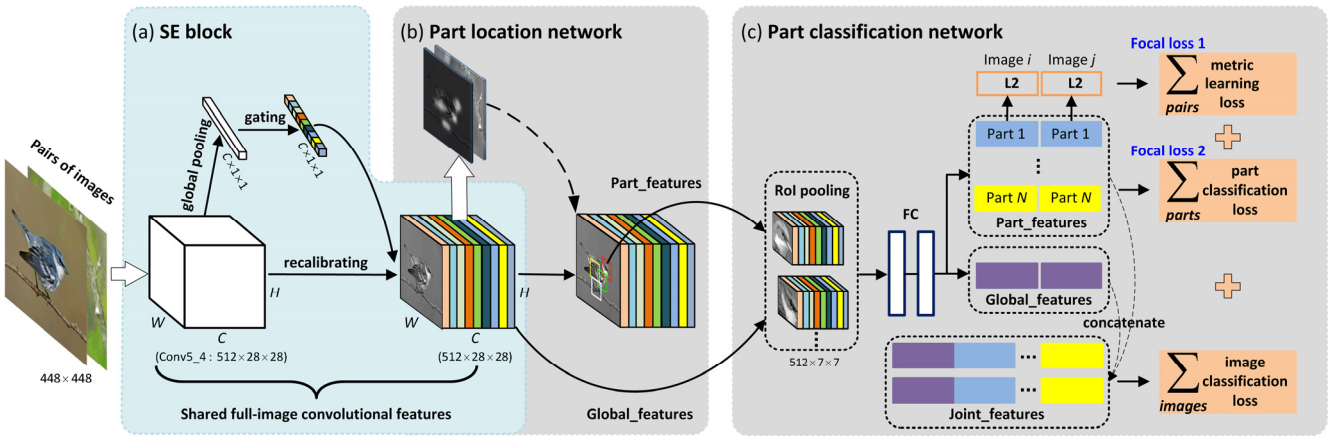


Figure 2: Illustration of our proposed P-CNN based FGVC framework. It is composed of three modules. The first module is a SE block used to recalibrate channel-wise feature responses. The second module is a Part Localization Network used to discover distinctive object parts. The third module is a Part Classification Network that classifies each individual part into image-level categories and also combines part-level local features and object-level global feature into a joint feature for the final classification. These three modules are merged into a unified network for an end-to-end training by sharing their full-image convolutional features between the tasks of part localization and part classification.

aims to learn a valid distance metric, measured by which the samples from the same class are as close as possible, while the samples from different classes are as far as possible. Two kinds of constraints including pairwise constraints and triplet constraints are widely used to capture the semantic similarity among images. To mine informative triplets for constrained feature learning, a deep metric learning approach is proposed in [16] to perform online hard triplet sampling with humans in the loop. Similarly, a smart mining process is processed to train the proposed triplet-based deep metric learning model in [38] by using the global structure loss with a triplet loss. Besides, the work in [53] attempts to model the relation between the fine-grained labels and attributes in a hierarchical manner and further seamlessly embed such hierarchical structures in a generalized triplet loss to constrain the CNNs for learning more discriminative features. Pairwise constraints, namely multi-attention multi-class constraints are introduced in [44] to pull same-attention same-class features closer, while push different-attention or different-class features away.

In contrast to the previous methods, this paper proposes a novel Duplex Focal Loss (DFL) for metric learning loss and part classification loss to automatically focus on training on hard examples and thus down-weight easy parts, which allows learning more discriminative features for capturing subtle differences between dissimilar classes.

2.2 Sophisticated Part Localization

Some early works [1, 6, 17, 41] mainly focus on leveraging the manually labeled object and part annotations to localize distinctive regions for fine-grained recognition. In [1], a strongly-supervised deformable part model is trained in a structured learning framework while the popular R-CNN [54] framework is employed to detect parts in addition to the whole object in [6]. However, the heavy manual annotation costs largely limit the large-scale practical application.

Recently, a more general scenario that does not rely on object/part annotations at the training phase has received increasing interests. Among these works, one important research direction is to design visual attention models for attending discriminative regions in order to imitate the way how humans perform visual sequence recognition. The work in [12] is among the first attempts for applying attention to FGVC, which uses a deep recurrent neural architecture with the iterative attention selection mechanism. The information of multi-resolution glimpses is recursively processed to output the next location and the final prediction is computed through the sequence of glimpses. However, the computational burden is high because calculating features at each glimpse in [12] requires forwarding GoogLeNet three times, which leads to very slow training and testing speed. In contrast, the method proposed in [18] is much more computationally efficient because of its fully-convolutional architecture, and it is capable of simultaneously focusing its glimpse on multiple visual attention regions. The aforementioned attention models are non-differentiable and are trained with the reinforcement learning technique to learn task-specific policies. Additionally, a two-level attention model including object-level and part-level attention is proposed in [15] to progressively attend the objects and their parts. The main drawback is that two types of attention models are independent and cannot be trained in an end-to-end fashion.

Alternatively, significant progresses have been made by learning the part localization network for localizing discriminative parts. The work in [17] designed a fully convolutional part localization network under strong part-level supervision to locate multiple object parts. A localization network that fits the transformation parameters for localizing a series of discriminative regions is designed in Spatial Transformer Networks [7]. Despite that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2933510, IEEE Transactions on Pattern Analysis and Machine Intelligence

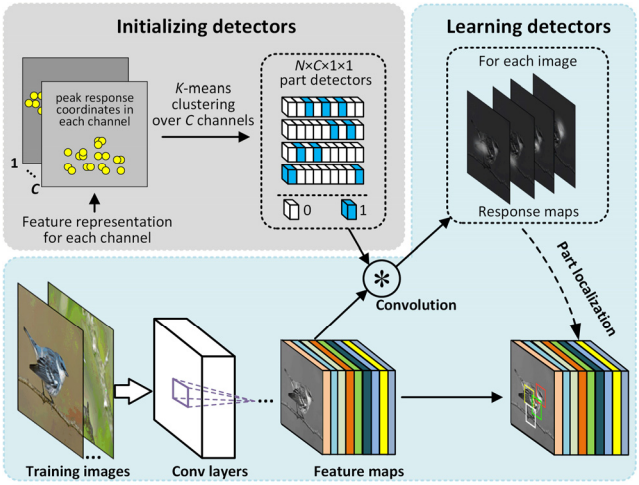


Figure 3: Illustration of our proposed Part Localization Network used to discover distinctive object parts. It aims to learn a set of discriminative part detectors in an unsupervised learning fashion to discover informative object parts.

it can be inserted into existing convolutional architectures and can be trained in an end-to-end fashion using back-propagation, the localization parameters involved in a number of hidden fully- connected or convolutional layers are heavy, which limit its efficiency in real applications. Moreover, the recurrent attention proposal network is designed in [28] to recurrently refine the location of one attention area to obtain incremental performance gains and achieves state-of-the-art performance.

The most relevant work to ours comes from [43], which proposed a Channel Grouping Network to group spatially-correlated feature channels for generating multiple discriminative parts. *Compared with [43], the advantages of our work are three folds. First, in order to localize distinctive object parts more accurately, we exploit channel attention learning to adaptively recalibrate channel-wise feature responses by selectively emphasizing informative class-specific features and suppressing less useful ones. Second, the work in [43] involves a group of fully-connected layers for part localization, in contrast to [43], our proposed part localization network is a fully convolutional network, which treats a bank of 1×1 convolutional filters as discriminative part detectors, thus enabling end-to-end training efficiently. Third, in [43], the part localization network and the final image classifier are trained independently, different from [43], we use the same full-image convolutional layers for both part localization network and image classification network in a unified network.*

3 The Proposed Method

3.1 Overview of the Proposed Method

The architecture of the proposed method is illustrated in Figure 2. The central goal of our method is to discover distinctive object parts and learn discriminative features not only from the global images but also from the discriminative object parts.

To this end, we propose a novel fine-grained visual categorization system, termed Part-based Convolutional Neural Network (P-CNN), which is composed of three modules. The first module is a Squeeze-and-Excitation (SE) block [55], which uses global information to adaptively recalibrate channel-wise feature responses by selectively emphasizing informative channels and suppressing less useful ones. The second module is a deep fully convolutional network to locate distinctive object parts and we name it as Part Localization Network (PLN). The PLN is trained in an unsupervised learning fashion, through which a bank of convolutional filters are learned as discriminative part detectors. Thus, a group of informative object parts can be discovered by convolving the excited feature maps (through the SE block) with each learned part detector. The third module is a Part Classification Network (PCN) that has two streams. The first stream focuses on classifying each individual object part into image-level categories. The second stream concatenates part-level local features and global image feature from an image together into a joint feature with softmax function for the final classification. In order to learn more powerful part-level features and hence boost the joint feature capability, we propose a new Duplex Focal Loss (DFL) used for metric learning and part classification, in which we reshape the loss functions in order to focus on training on hard examples and thus down-weighting easy examples.

In order to further merge PLN and PCN into a unified network for end-to-end training by sharing the full-image convolutional features, we propose a simple training scheme that alternates between the training for part localization task and the training for part classification task.

3.2 Squeeze-and-Excitation (SE) Block

The Squeeze-and-Excitation block is a high capacity computational unit that can be inserted into any layer of an off-the-shelf CNN to improve its representational capacity. It explicitly models the interdependencies between channels by adaptively recalibrating channel-wise feature responses. As shown in [55], in the early layers, the features are typically class agnostic, whereas in later layers, the features become increasingly specialized and respond to different inputs in a highly class-specific manner. Our goal is to ensure that the network is able to selectively emphasize informative channels and suppress less useful ones so that they can help to better discover distinctive object parts. Consequently, in this paper, we insert the SE block into a higher layer Conv5_4 (VGG-19 model is used in this work) to excite informative features in a class-specific manner.

The basic structure of the SE building block is illustrated in Figure 2 (a). For the Conv5_4 features obtained with the VGG-19 model, we construct a corresponding SE block to perform feature recalibration as follows [55]. The Conv5_4 features are first passed through a squeeze operation (i.e., global average pooling), which aggregates each global feature map across spatial dimensions $H \times W$ to produce a 1×1 channel-wise descriptor. This descriptor embeds the global distribution of channel-wise feature responses by shrinking the information from the global receptive field of the network. In this way, we obtain a

C×1×1 channel descriptor vector for the input features of C×H×W. Then, we use an excitation operation to fully capture channel-wise dependencies. This is achieved by employing a simple gating mechanism with a sigmoid activation, which takes the channel descriptor vector aggregated in the squeeze operation as the input and outputs the activation vector acting as channel excitations (or channel weights). In order to limit model complexity and aid generalization, the gating mechanism is parameterized by forming a bottleneck with two fully connected layers, i.e., a dimensionality-reduction layer and a dimensionality increasing layer. Finally, the excited features of the SE block are obtained by recalibrating Conv5_4 features with the learned activations. In this way, the SE block intrinsically introduces dynamics conditioned on the input to boost feature discriminability. The benefit of using SE block for FGVC will be demonstrated in our experiments.

3.3 Part Localization Network (PLN)

Figure 3 illustrates the architecture of our proposed Part Localization Network (PLN). It aims to learn a set of discriminative part detectors in an unsupervised learning fashion to discover informative object parts used for subsequent FGVC. The previous works have found that a convolutional feature channel corresponds to a certain type of visual pattern, but it is difficult to extract rich part information with a single channel. Inspired by the work [43], we propose a more intuitive method to learn our PLN as follows.

Given a set of training images $\mathbb{X}=\left\{X_1, X_2, \cdots, X_{|\mathbb{X}|}\right\}$, we pass them through a series of convolutional and pooling layers including the aforementioned SE block to obtain their feature maps. As the work [43], for each feature channel, we represent it as a position vector (see Figure 3) whose elements are the peak response coordinates over all training images, which is given by

$$\left[\left(t_x^1, t_y^1\right), \left(t_x^2, t_y^2\right), \cdots, \left(t_x^{|\mathbb{X}|}, t_y^{|\mathbb{X}|}\right)\right] \tag{1}$$

where $\left(t_x^i, t_y^i\right)$ is the peak response coordinate of the i-th image in the training set, and $|\mathbb{X}|$ is the number of training images. Then, *we adopt K-means method to cluster all C feature channels into N groups by using Eqn. (1) as their feature representations, where each group aggregates spatially-correlated patterns from a group of channels whose peak responses appear in neighboring locations*. The resultant n-th group is represented by an indicator function $\mathbb{I}\{\cdot\}$ over all feature channels, which is given by

$$\boldsymbol{D}_n=\left[\mathbb{I}\{1\}, \cdots, \mathbb{I}\{c\}, \cdots, \mathbb{I}\{C\}\right] \tag{2}$$

where $\mathbb{I}\{c\}$ equals to one if the c-th feature channel belongs to the n-th cluster and zero otherwise. Thus, each C×1×1 clustering indicator vector is regarded as an initialized part detector.

For each input image X, we can obtain its part-based response map for the n-th C×1×1 part detector (also can be regarded as a convolutional filter) by convolving its feature maps with this part detector, which is given by

$$\mathbf{M}_n(X)=\operatorname{sigmoid}\left(\boldsymbol{D}_n^{\mathrm{T}}\mathbf{U}\right)=\operatorname{sigmoid}\left(\sum_{c=1}^{C} d_c \mathbf{U}_c\right) \tag{3}$$

where $\mathbf{U}$ is the feature maps of the input image X, $\mathbf{U}_c$ is the c-th feature channel of convolutional features $\mathbf{U}$, and d_c is the c-th element of $\boldsymbol{D}_n$. $\mathbf{M}_n(X)$ is normalized with a sigmoid activation to obtain the final part response map, which indicates the occurrence probability of a specific object part.

Since there is no available bounding box information for object parts, in order to train compact and discriminative part detectors in an end-to-end fashion, we introduce a loss function as [43] by focusing on compactness and diversity restrictions on the part response maps, which is computed by

$$L_{PLN}=\min \sum_{n=1}^{N}\sum_{i=1}^{|\mathbb{X}|}\left(Dis\left(\mathbf{M}_n(X_i)\right)+\lambda_1 Div\left(\mathbf{M}_n(X_i)\right)\right) \tag{4}$$

where λ_1 is a trade-off parameter in order to control the relative importance of these two terms.

The first term in Eqn. (4) is a distance measure function to encourage a compact distribution, which is defined by

$$Dis\left(\mathbf{M}_n(X_i)\right)=\sum_{(x,y)\in\mathbf{M}_n} m_n(x,y)\left[\left\|x-t_x^i\right\|^2+\left\|y-t_y^i\right\|^2\right] \tag{5}$$

where $m_n(x,y)$ denotes the response amplitude of $\mathbf{M}_n$ at the coordinate (x,y).

The second term in Eqn. (4) is a diversity function used to support a diverse response distribution from different part response maps, which is defined by

$$Div\left(\mathbf{M}_n(X_i)\right)=\sum_{(x,y)\in\mathbf{M}_n} m_n(x,y)\left[\max_{k\neq n} m_k(x,y)-mrg_1\right] \tag{6}$$

where n, k are the indexes of different response maps. mrg_1 is a margin to enable the loss less sensitive to noises.

Once we have learned a set of C×1×1 part detectors with have high response to certain discriminative regions, by convolving the feature maps with these detectors we can obtain a set of part response maps. Therefore, a discriminative object part can be discovered simply by picking the location with the maximum value in the entire response map.

3.4 Part Classification Network (PCN)

As shown in Figure 2 (c), our Part Classification Network (PCN) has two streams. The first stream aims to learn more powerful part features by designing a Duplex Focal Loss (DPL), which reshapes the loss functions to focus on hard examples and thus down-weight easy examples during the training process. The second stream concatenates part-level local features and global image feature from an image together into a joint feature with softmax function for the final classification.

Specifically, given an input image X, for each discovered object part and the original image, we uniformly pass their convolutional features through a region of interest (RoI) pooling layer and two fully-connected (FC) layers to obtain the part-level local features and the global

image feature. For the VGG-19 model, we set the size of the RoI pooling layer to 512×7×7 and the FC parameters are directly transferred from VGG-19. In order to train PCN in an end-to-end fashion, we propose a novel loss function as follows.

$$L_{PCN} = \min\left(\underbrace{FL_1 + \lambda_2 FL_2}_{\substack{\text{duplex focal loss used for} \\ \text{part classification and metric learning}}} + L_{cls} \right) \tag{7}$$

The first two terms in Eqn. (7) are our newly proposed duplex focal loss used to classify each individual object part into image-level categories, and to explicitly model the similarity and dissimilarity constraints of a paired object parts, respectively. λ_2 is a trade-off parameter in order to control the relative importance of these two focal loss terms. Their detailed formulations are designed as follows.

$$FL_1 = \sum_{i=1}^{|\mathbb{X}|} \sum_{n=1}^{N} -\left(1 - y_{i,n}\right)^{\gamma} \log y_{i,n} \tag{8}$$

where $y_{i,n}$ is the predicted softmax probability for the n-th part from the i-th image for the class with label $\hat{y} = 1$ ($\hat{y}$ is the ground-truth label). γ is a tunable focusing parameter used to generate a modulating factor $\left(1 - y_{i,n}\right)^{\gamma}$ for the softmax cross entropy loss. In this way, we can reshape the loss function to down-weight easy examples and thus focus on hard examples during the training process.

$$FL_2 = \sum_{q,p} \begin{matrix} \left(1 - \max\left(y_q, y_p\right)\right)^{\gamma} \ell_{qp} D^2\left(x_q, x_p\right) + \\ \left(1 - \max\left(y_q, y_p\right)\right)^{\gamma} \left(1 - \ell_{qp}\right) \max\left(0, mrg_2 - D^2\left(x_q, x_p\right)\right) \end{matrix} \tag{9}$$

where $\ell_{qp} \in \{1, 0\}$ is a label indicator to indicate if two corresponding parts $\left(x_q, x_p\right)$ (e.g., a bird head versus a bird head from two images) are from the same class or not. Specifically, if the two corresponding parts $\left(x_q, x_p\right)$ are from the same class, they will be considered as a similar pair and ℓ_{ij} equals to one; otherwise, they will be considered as a dissimilar pair and ℓ_{ij} equals to zero. $D\left(x_q, x_p\right)$ is the squared Euclidean distance between a paired L2-normalized vectors x_q and x_p, which is computed by

$$D\left(x_q, x_p\right) = \left\| x_q - x_p \right\|_2 \tag{10}$$

The first term in Eqn. (9) penalizes a similar pair that is too far apart. For dissimilar pairs, we hope that their distances are bigger than a margin mrg_2, so the second term in Eqn. (9) is used to penalize the dissimilar pair distances for being smaller than a margin by using the hinge loss function. Besides, in order to further focus on hard examples and down-weight easy examples during the training process, we also add a modulating factor $\left(1 - \max\left(y_q, y_p\right)\right)^{\gamma}$ to the metric learning loss, where y_q is the predicted softmax probability of the q-th part for the class with label $\hat{y} = 1$.

The third term in Eqn. (7) is a softmax cross entropy loss used for the final classification, which takes the joint features as the input by combining part-level local features and global image feature together.

3.5 Joint Training of PLN and PCN

So far, both PLN and PCN are still trained independently. This will modify their convolutional layers. It therefore needs to develop a scheme that allows for sharing the convolutional layers between the two networks, rather than learning two networks separately.

To this end, we propose a simple 4-step training technique, motivated by the work [43]. Specifically, the first step trains the PLN as described in Section 3.3. This network is initialized with the pre-trained VGG-19 model and then trained for part localization task in an end-to-end fashion. The second step trains the PCN by using the object parts discovered with the PLN at step-1. This classification network is also initialized by the pre-trained VGG-19 model. At this time, the two networks do not share full-image convolutional layers. The third step uses the classification network from step-2 to initialize PLN training, but we fix the shared full-image convolutional layers and only fine-tune N part detectors. Now the two networks share the convolutional layers. Finally, we fine-tune the unique layers of PCN by keeping the shared convolutional layers and the discovered parts fixed. As such, both networks share the same full-image convolutional layers and form a unified network.

4 Experiments

To demonstrate the effectiveness of the proposed method, we construct comprehensive experiments, which are organized as follows. We first describe the benchmark datasets and implementation details. Next, the classification results and comparisons with state-of-the-art methods are presented. We finally present a series of ablation studies to demonstrate the impact of each component in our proposed FGVC system.

4.1 Datasets

Experiments are conducted on three challenging datasets, including Caltech-UCSD Birds (CUB-200-2011) [56], FGVC-Aircraft [57] and Stanford Cars [58], which are widely used to evaluate fine-grained visual categorization methods.

CUB-200-2011 [56] dataset is the most widely used dataset for FGVC. It includes 11,788 images from 200 different bird species. The images are split into the training and testing sets with 5994 images for training and 5794 images for testing. Each image in this dataset is associated with detailed annotations, including image-level labels, object bounding boxes, part locations, and binary attributes, which can be used for fully supervised learning. Note that in all our experiments, only image-level labels are employed during training without using any part or bounding box annotation. Classifying bird subcategories is challenging because of the different poses, viewpoints and cluttered background.

FGVC-Aircraft [57] dataset contains 10,000 images from 100 aircraft variants, which are split into equal train-

ing, validation and testing sets. And each image is provided with a bounding box annotation and an image-level class label. Some aircraft variants have extremely subtle differences that only can be distinguished by the number of windows in the model. Compared to CUB-200-2011, airplanes appear in relatively clear background and tend to occupy a significantly larger portion of the image.

Stanford Cars [58] dataset consists of 16,185 images from 196 classes of cars, which is divided with 8144 images for training and 8041 images for testing. Categories are typically annotated at the level of Year, Maker, Model, e.g., "2012 Tesla Model S" and "2012 BMW M3 coupe." Unlike the FGVC-Aircraft dataset, most of the cars occupy a small portion of image in more cluttered background. Thus, accurate part localization would play a more significant role here.

We adopt top-1 accuracy as the evaluation metric to comprehensively evaluate the classification performances of our P-CNN based FGVC method and baseline methods, which is defined as the number of images that are correctly classified divided by the number of testing images.

4.2 Implementation Details

To make fair comparisons with the state-of-the-art methods, we implement the proposed P-CNN method based on the widely used CNN model VGG-19 [36], which is pre-trained on ImageNet. For all three datasets, the P-CNN takes as input images with the size of 448×448 pixels. The size of the part bounding box is set to 96×96 pixels and the number of parts is set to 4 as the setting in [43]. And the final feature dimensionality for each object part and the global image is 4096D.

For a better training of P-CNN, the additional introduced parameters of the corresponding three modules, i.e., the SE block, the part localization network and the part classification network, are initialized as follows. Specifically, as for the SE block, we firstly perform max pooling on the output feature maps of the SE block to obtain new feature maps with the size of 512×7×7, and then add two fully connected layers upon the feature maps to train a fine-grained image classification system with only global image features. The learned CNN layers including the excited Conv5_4 and its previous layers are used to initialize the backbone network of our P-CNN. For the part localization network, we use a similar technique as [43], i.e., we adopt K-means clustering to cluster all C feature channels based on their coordinates of peak responses over all training images into N groups and use the clustering indicator vectors to initialize the parameters of the introduced N convolutional filters (see Figure 3). As for the two fully connected layers in the part classification network, the parameters are directly transferred from the corresponding layers of VGG-19 pre-trained on ImageNet.

After performing initialization as mentioned above, optimization is performed by using stochastic gradient descent (SGD) with momentum 0.9 *and weight decay 0.0005*. *The epoch number is set to be about 60-100 depending on the datasets and the learning rate is set to be 0.001 and to be multiplied by 0.1 every 20 epochs*. For PLN training, the batch size is 16 images. For PCN training and the joint training of PLN and PCN, 4 similar image pairs and 4 dissimilar image pairs are selected to form a mini-batch size of 8 image pairs. During training, if we use randomly sampled image pairs, many of them satisfy the metric learning constraint well and give nearly zero loss in Eqn. (9). That is, those easy image pairs have less effect for updating model parameters. This makes the training process inefficient and even unstable. To address this issue, we use a hard example mining scheme: only the hard examples that violate the metric learning constraint are included into the training process. In addition, the trade-off parameters λ_1 in Eqn. (4) and λ_2 in Eqn. (7) are set to 2 and 0.1, respectively. The tunable focusing parameter in Eqns. (8) and (9) is set to 2.

4.3 Comparisons with the State-of-the-art Methods

The results of our proposed P-CNN method and the comparisons with state-of-the-art methods are reported in Table 1. Rows 1-3 show annotation-based methods that employ object bounding boxes or part annotations for model training. Rows 4-13 give annotation-free methods including our P-CNN method that use only image-level labels for training. According to Table 1, our method outperforms most of the state-of-the-art methods by a noticeable margin on all three datasets, which clearly demonstrates the effectiveness of the proposed method. The detailed comparison analysis on these three datasets is as follows.

CUB-200-2011 dataset. As shown in Table 1, the two earlier methods, i.e., PS-CNN [17] and Part-RCNN [6], achieve comparable results based on the AlexNet model, which require both bounding box and parts annotations or bounding box annotations alone to learn the part localizations directly from the images or pre-computed object proposals. The Mask-CNN method [41] obtains state-of-the-art performance of 85.7%, which also relies on additional annotations of bounding box and parts to learn discriminative features. In contrast, our proposed P-CNN method works without any additional annotations but outperforms them with a large margin.

Furthermore, our method outperforms those annotation-free, part localization-based methods, such as STN [7], RA-CNN [28], and MA-CNN [43], with accuracy improvements of 3.2%, 2.0% and 0.8%, respectively. STN [7] is a spatial transformer network used to learn invariance to scale, warping by feature transforming. RA-CNN [28] presents a recurrent attention CNN framework to locate discriminative areas recurrently for better classification performance. MA-CNN [43] proposes a part localization sub-network by clustering and grouping feature channels. Our method shares a similar part localization manner to MA-CNN, but ours can be efficiently trained in an end-to-end fashion. More importantly, our proposed P-CNN method further focuses on learning discriminative features for the parts by employing metric learning. Compared with the MAMC method [44] that uses a stronger baseline CNN model (ResNet-50) network for learning discriminative features, our method using the weaker VGG-19 network still achieves a modest accuracy improvement of 1.1%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2933510, IEEE Transactions on Pattern Analysis and Machine Intelligence

Table 1: Accuracies (%) of our proposed method and 12 state-of-the-art methods on the CUB-200-2011 dataset, the FGVC Aircraft dataset, and the Stanford Cars dataset. "BBox" and "Parts" refer to object bounding box annotations and part annotations, respectively. "n/a" denotes "not available".

Method	Base Model	Annotations	Accuracy		
			CUB-200-2011	FGVC Aircraft	Stanford Cars
PS-CNN [17]	AlexNet	BBox+Parts	76.6	-	-
Part-RCNN [6]	AlexNet	BBox	76.4	-	-
Mask-CNN[41]	VGG-16	BBox+Parts	85.7	-	-
STN [7]	Inception	n/a	84.1	-	-
MAMC [44]	ResNet-50	n/a	86.2	86.5	92.8
OPAM [59]	VGG-16	n/a	85.8	-	92.2
Bilinear CNN [30]	VGG-16	n/a	84.1	86.6	91.3
PDFR [37]	VGG-16	n/a	84.5	-	-
AutoBD [42]	VGG-16	n/a	81.6	-	88.9
DVAN [51]	VGG-16	n/a	79.0	-	87.1
RA-CNN [28]	VGG-19	n/a	85.3	-	92.5
MA-CNN [43]	VGG-19	n/a	86.5	89.9	92.8
P-CNN (Ours)	VGG-19	n/a	**87.3**	**90.6**	**93.3**

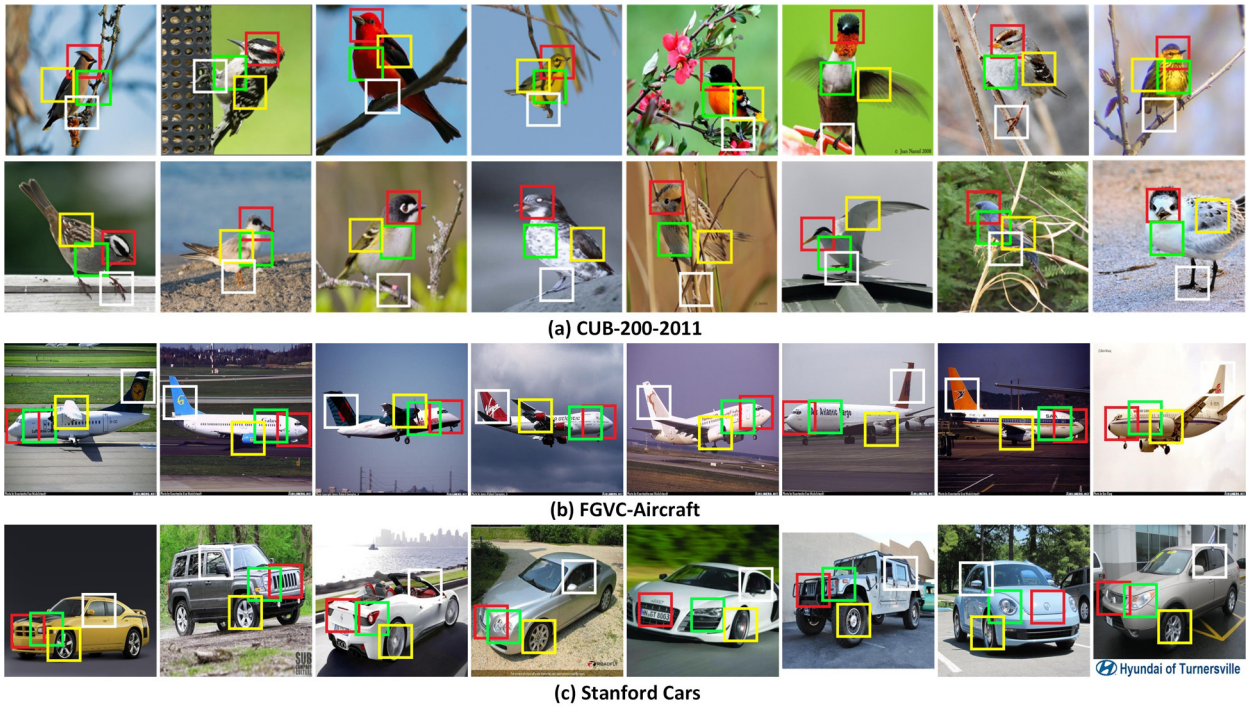


Figure 4: Part localization results on the CUB-200-2011 bird dataset, the FGVC-Aircraft dataset and the Stanford Cars dataset. The part detectors can localize consistently discriminative parts on all these three datasets.

FGVC-Aircraft and **Stanford Cars** datasets. The proposed P-CNN method achieves excellent performances on the FGVC Aircraft dataset and Stanford Cars dataset. On the FGVC Aircraft dataset, our method obtains significant accuracy improvements of 4.1% and 4.0% when compared with the MAMC method [44] and the Bilinear CNN [30] method. Compared with the strong baseline MA-CNN [43], our method also achieves the improvement of 0.7%. More importantly, our method only needs one feed-forward stage for localizing parts and extracting discriminative features from these parts for classification. These results further validate the effectiveness and efficiency of our method. On the Stanford Cars dataset, we have the same observation as on the CUB-200-2011 data set and the FGVC-Aircraft dataset. Specifically, our method outperforms most of the baseline methods with

Table 2: Comparisons of different methods for part localization in terms of classification accuracy (%) on the CUB-200-2011 dataset, the FGVC-Aircraft dataset and the Stanford Cars dataset.

Method	Accuracy		
	CUB-200-2011	FGVC-Aircraft	Stanford Cars
P-CNN (initial clustering without SE block)	82.2	82.1	86.5
P-CNN (initial clustering with SE block)	83.7	84.4	89.1
P-CNN (independently train PLN and PCN)	86.8	89.5	92.5
P-CNN (jointly train PLN and PCN)	87.3	90.6	93.3
MA-CNN (initial) [43]	82.0	-	-
MA-CNN (L_{cng}) [43]	85.3	-	-
MA-CNN (L_{cls}+L_{cng}) [43]	86.5	89.9	92.8

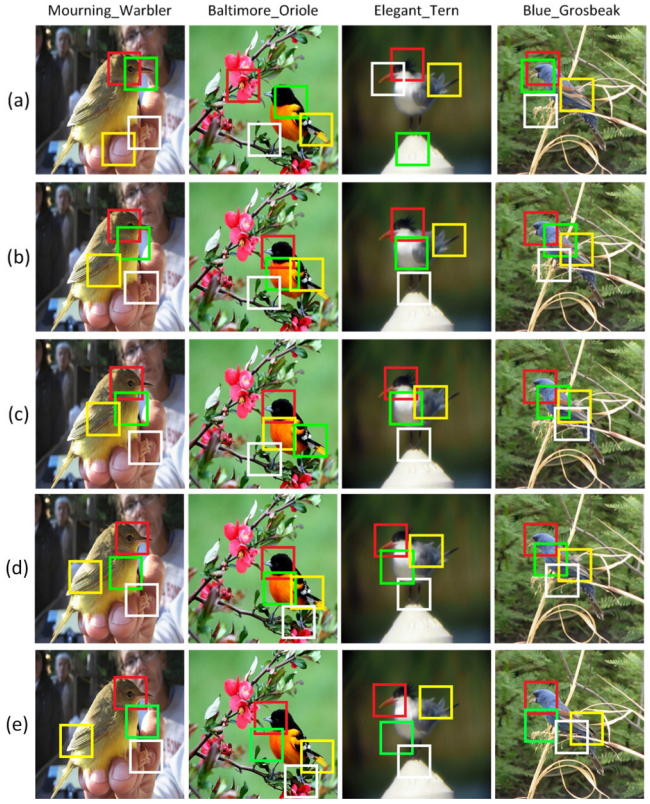


Figure 5: Part localization results: (a) the initial parts by channel clustering without SE block, (b) the initial parts by channel clustering with SE block, (c) the intermediate parts by training PLN, (d) the final parts by jointly training PLN and PCN, and (e) the final parts by using MA-CNN method [43].

the large margin. Compared with the second best method (the MAMC method [44] and the MA-CNN method [43]), we still achieve an accuracy improvement of 0.5%. This demonstrates the powerful capability of discovering distinctive parts and learning discriminative features of our proposed method.

4.4 Ablation Analysis

In this section, we perform a series of detailed ablation studies to evaluate the contributions of some key components involved in our proposed method on the CUB-200-2011 dataset, *FGVC-Aircraft dataset and Stanford-Cars dataset*.

4.4.1 Part Localization Analysis

The parts of some *individual bird*, *aircraft and car* examples located by the proposed part localization network (with joint training of PLN and PCN) are shown in Figure 4. As can be seen, despite of that the birds appear in different poses and viewpoints with cluttered background, our PLN method can learn discriminative part detectors to consistently localize the parts of head, breast, wing and feet. *For aircraft and car categories, consistently discriminative part areas are also successfully detected. This is mainly because the convolutional filters-based part detectors and their fine-grained features can be jointly learned in an end-to-end fashion by using back-propagation without losing discriminative information constrained in the similar and dissimilar image pairs.*

To further demonstrate the benefits of the newly proposed SE block and our proposed joint training technique for PLN and PCN, four bird examples with their visualized part localization results are presented in Figure 5. They are: (a) the initial parts by channel clustering without the SE block (i.e., feature recalibration), (b) the initial parts by channel clustering with SE block, (c) the intermediate parts by training PLN, and (d) the final parts by jointly training PLN and PCN. As can be seen, some part detectors initialized by channel clustering without the SE block attend on the regions of background. For example, the red and green part detectors both localize the regions of background in the images of "*Baltimore_Oriole*" and "*Elegant_Tern*" due to the strong color characteristics of the background regions. Besides, some part detectors (e.g., the red and green detectors for "*Blue_Grosbeak*" image) localize parts with large overlap, which could achieve less classification accuracy gains when compared with that from different part locations. After performing feature recalibration through the SE block by emphasizing informative channels and suppressing less useful ones, the initial part detectors are more robust to deal with the background disturbance and can localize most parts from the foreground objects. However, the part detectors still cannot locate consistent parts for all these four birds. This problem can be solved (see Figure 5(c)) by training a part localization network through optimizing Eqn. (4). Further on, we can jointly train PLN and PCN in order to obtain more discriminative part detectors that can consistently

work on different foreground birds, as shown in Figure 5(d).

The quantitative comparison on part localization is conducted in terms of classification accuracy. All compared methods employ the proposed P-CNN method to perform fine-grained classification, but with different part localization settings. The accuracies are listed in Table 2. As can be seen, performing feature recalibration through the SE block before channel clustering learns a better initial part detector and improves the classification accuracy by *1.5% on the CUB-200-2011 dataset. Large improvements of 2.3% and 2.6% are achieved on the FGVC-Aircraft dataset and the Stanford-Cars dataset, respectively. Significant improvements of 3.1% and 3.4% for independently training the proposed PLN are achieved when compared with initial clustering with SE block on the CUB-200-2011 dataset and the Stanford-Cars dataset, respectively. Remarkable improvement of 5.1% is gained on the FGVC-Aircraft dataset. Additionally, by jointly training PLN and PCN can further improve the classification accuracy. Slight improvements of 0.5% and 0.8% are obtained on the CUB-200-2011 dataset and the Stanford Cars dataset, respectively. About 1.1% gain is achieved on the FGVC-Aircraft dataset.*

To better illustrate the effectiveness of the proposed PLN, we also present the part localization results of MA-CNN [43] in Figure 5 (e) and the corresponding classification accuracy in Table 2 for ease of comparison. As can be seen, on the CUB-200-2011 dataset we obtain close classification results with MA-CNN for initial clustering parts. However, the classification accuracy of our P-CNN that jointly trains PLN and PCN can achieve better performance than MA-CNN on all three datasets, which clearly demonstrates the superiority of the proposed PLN. As shown in Figure 5, our PLN can localize more discriminative parts than MA-CNN, especially from the images with cluttered backgrounds. This is mainly because that our designed PLN exploits a bank of 1×1 convolutional filters instead of a group of fully-connected layers in MA-CNN to perform part localization, which is more efficient to enable PLN to be trained in an end-to-end fashion along with the final PCN, which can promote to learn more discriminative parts. Besides, SE block is employed to recalibrate feature maps, which adaptively emphasizes informative channels and is beneficial for a better input of PLN.

4.4.2 Image Classification Analysis

Our P-CNN based fine-grained visual categorization system is composed of SE block, PLN and PCN. As described above, we have analyzed the performances of part localization and the impacts of the SE block and joint training for part localization. In this subsection, we further analyze the impact of some key components (e.g., SE block, part features, metric learning, and duplex focal loss) for the final image classification task.

Base Network: Firstly, a base network is built upon the VGG-19 model with slight modifications. That is, we add a RoI pooling layer upon the Conv5_4 layer, followed by two fully connected layers. The base network is trained with global image classification loss, and the corresponding classification accuracy is provided in Table 3.

Effectiveness of the SE block: The SE block adaptively recalibrates channel-wise feature responses by modeling interdependencies between channels and has shown to be beneficial for improving the part localization performance. Here we insert a SE block to the Conv5_4 layer of the base network to evaluate its impact for classification. *As shown in Table 3, adding a SE block solely can offer 2.5% and 2.2% accuracy improvements compared to the base network on the CUB-200-2011 dataset and the FGVC-Aircraft dataset, respectively. Significant improvement of 4.1% is achieved on the Stanford Cars dataset.*

Importance of part-features: Global image features have the limited ability of characterizing the subtle differences among fine-grained categories. Here we evaluate the importance of localizing parts and fusing part-based feature representations for fine-grained image classification. Specifically, the SE block and the proposed PLN are both added to the base network. And the new network is trained with the part classification loss and image classification loss. The part classification task attempts to classify each individual part into image-level categories. The image classification task aims to combine the part-level local features and global image-level features together into a joint feature for final classification. *As can be seen from Table 3, involving localizing parts and fusing part-based representations leads to significant accuracy improvements of 7.6%, 8.1% and 6.8% compared to the results obtained by using only image-level information on the CUB-200-2011 dataset, FGVC-Aircraft dataset and Stanford Cars dataset, respectively. The results clearly demonstrate it is necessary to fuse part-based representations for fine-grained image recognition.*

Superiority of metric learning: Here we evaluate the role of metric learning for learning discriminative features from the localized distinctive parts in improving fine-grained classification performance. Specifically, the image classification loss and part classification loss are added with a metric learning loss. The introduced metric learning loss enforces the features of the corresponding parts from similar image pairs to be as close as possible, while those from dissimilar image pairs to be as far as possible. We can observe from Table 3 that the network with metric leaning outperforms the network without metric leaning with a clear margin (0.9% accuracy improvement) *on the CUB-200-2011 dataset. Besides, 1.3% and 1.4% accuracy improvements are achieved on the FGVC-Aircraft dataset and the Stanford Cars dataset, which again demonstrates the superiority of learning discriminative part-level features by using metric learning.*

Duplex focal loss: Not all the localized parts contribute equally to the part-based feature learning and classification. Here we design a duplex focal loss to reshape the metric learning loss and part classification loss to focus training on hard parts and thus down-weighting easy parts. In the feature learning process, hard parts from similar image pairs usually appears with large intra-class variances and are enforced to be closer, while hard parts from dissimilar image pairs usually appears with subtle inter-class differences and are enforced to be further. Through this, more discriminative features are learned. As shown in Table 3, the duplex focal loss can lead to 0.6% accuracy gains when compared with the network

Table 3: Ablation studies of the P-CNN on the CUB-200-2011 dataset, the FGVC-Aircraft dataset and the Stanford Cars dataset.

		Global feature		Joint feature		
Base Network		√	√	√	√	√
SE block			√	√	√	√
Part features with traditional classification loss				√	√	
Part features with traditional metric learning loss					√	
Part features with focal classification loss						√
Part features with focal metric learning loss						√
Accuracy	CUB-200-2011	75.7	78.2	85.8	86.7	87.3
	FGVC-Aircraft	77.9	80.1	88.2	89.5	90.6
	Stanford Cars	80.2	84.3	91.1	92.5	93.3

without using the duplex focal loss *on the CUB-200-2011 dataset. For FGVC-Aircraft dataset and Stanford Cars dataset, 0.9% and 0.8% performance improvements are achieved, respectively. All the gains clearly illustrates that it is beneficial to pay more attention on hard parts in the training to learn more discriminative features, which further improves fine-grained image recognition*.

5 Conclusion

In this paper, we have proposed an elegant and effective end-to-end fine-grained visual categorization system, termed Part-based Convolutional Neural Network (P-CNN), which builds upon Squeeze-and-Excitation (SE) block to recalibrate feature maps for obtaining a better input for the subsequent Part Localization Network (PLN) and Part Classification Network (PCN). The PLN is performed by learning a set of convolutional filters as discriminative part detectors. Then, the obtained parts are individually classified by the PCN and are further combined with object-level global features for final classification with a novel Duplex Focal Loss (DFL), which reshapes the metric learning loss and part classification loss by focusing on hard examples and thus down-weighting easy examples for discriminative feature learning and robust classification. In the experiments, we have comprehensively evaluated the proposed method for the FGVC tasks on three widely-used datasets. On all three datasets, we have achieved promising accuracies when compared with the state-of-the-art methods.

Acknowledgment

This work was supported in part by the National Science Foundation of China (NSFC) under Grants 61701415, 61772425, 61522207, 61773315, and 61790552, in part by the Young Star of Science and Technology in Shaanxi Province under grant 2018KJXX-029, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102018zy023.

J. Han and X. Yao contributed equally to this work. G. Cheng is the corresponding author.

References

[1] S. Branson, O. Beijbom, and S. Belongie. Efficient Large-Scale Structured Learning. In *CVPR*, 2013, 1806-1813.

[2] A. Angelova, and S. Zhu. Efficient Object Detection and Segmentation for Fine-Grained Recognition. In *CVPR*, 2013, 811-818.

[3] S. Gao, W. H. Tsang, and Y. Ma. Learning Category-Specific Dictionary and Shared Dictionary for Fine-Grained Image Categorization. *IEEE Trans. Image Process.*, 23(2): 623-634, 2013.

[4] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. In *CVPR*, 2014,

[5] J. Krause, T. Gebru, J. Deng, L. J. Li, and F. F. Li. Learning Features and Parts for Fine-Grained Recognition. In *ICPR*, 2014, 26-33.

[6] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *ECCV*, 2014, 834-849.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. In *NIPS*, 2015, 2017-2025.

[8] J. Krause, H. Jin, J. Yang, and F. F. Li. Fine-grained recognition without part annotations. In *CVPR*, 2015, 5546-5555.

[9] D. Lin, X. Shen, C. Lu, and J. Jia. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 2015, 1666-1674.

[10] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015, 1449-1457.

[11] Q. Qian, R. Jin, S. Zhu, and Y. Lin. Fine-grained visual categorization via multi-stage metric learning. In *CVPR*, 2015, 3716-3724.

[12] P. Sermanet, A. Frome, and E. Real. Attention for Fine-Grained Categorization. In *ICLR*, 2015, 224-30.

[13] M. Simon, and E. Rodner. Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks. In *ICCV*, 2015, 1143-1151.

[14] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang. Multiple Granularity Descriptors for Fine-Grained Categorization. In *ICCV*, 2015, 2399-2406.

[15] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015, 842-850.

[16] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. Fine-grained

Categorization and Dataset Bootstrapping using Deep Metric Learning with Humans in the Loop. In *CVPR*, 2016, 1153-1162.

[17] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-Stacked CNN for Fine-Grained Visual Categorization. In *CVPR*, 2016, 1173-1182.

[18] X. Liu, T. Xia, J. Wang, and Y. Lin. Fully Convolutional Attention Networks for Fine-Grained Recognition. *arXiv 1603.06765V4*, 2016.

[19] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning Deep Representations of Fine-Grained Visual Descriptions. In *CVPR*, 2016, 49-58.

[20] Y. Wang, J. Choi, V. I. Morariu, and L. S. Davis. Mining Discriminative Triplets of Patches for Fine-Grained Classification. In *CVPR*, 2016, 1163-1172.

[21] X. S. Wei, J. H. Luo, J. Wu, and Z. H. Zhou. Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. *IEEE Trans. Image Process.*, PP(99): 1-1, 2016.

[22] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Coarse-to-Fine Description for Fine-Grained Visual Categorization. *IEEE Trans. Image Process.*, 25(10): 4858-4872, 2016.

[23] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas. SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Recognition. In *CVPR*, 2016, 1143-1152.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016, 2921-2929.

[25] F. Zhou, and Y. Lin. Fine-Grained Image Classification by Exploring Bipartite-Graph Labels. In *CVPR*, 2016, 1124-1133.

[26] S. Cai, W. Zuo, and L. Zhang. Higher-Order Integration of Hierarchical Convolutional Activations for Fine-Grained Visual Categorization. In *ICCV*, 2017, 511-520.

[27] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie. Kernel Pooling for Convolutional Neural Networks. In *CVPR*, 2017, 3049-3058.

[28] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017, 3.

[29] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo. Knowledge-Embedded Representation Learning for Fine-Grained Image Recognitio. In *IJCAI*, 2018,

[30] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1309-1322, 2018.

[31] Y. Wang, V. I. Morariu, and L. S. Davis. Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition. In *CVPR*, 2018,

[32] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. In *CVPR*, 2018,

[33] X. Zhe, S. Huang, Y. Zhang, and D. Tao. Webly-supervised Fine-grained Visual Categorization via Deep Domain Adaptation. *IEEE Trans Pattern Anal Mach Intell*, PP(99): 1100-1113, 2018.

[34] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In *ICCV*, 2017, 5552-5561.

[35] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. One-Shot Fine-Grained Instance Retrieval. In *ACM MM*, 2017, 342-350.

[36] K. Simonyan, and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015,

[37] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking Deep Filter Responses for Fine-Grained Image Recognition. In *CVPR*, 2016, 1134-1142.

[38] B. Harwood, K. B. G. Vijay, G. Carneiro, I. Reid, and T. Drummond. Smart Mining for Deep Metric Learning. In *ICCV*, 2017, 2840-2848.

[39] S. Kong, and C. Fowlkes. Low-Rank Bilinear Pooling for Fine-Grained Classification. In *CVPR*, 2017, 7025-7034.

[40] Z. Xu, D. Tao, S. Huang, and Y. Zhang. Friend or foe: Fine-grained categorization with weak supervision. *IEEE Trans. Image Process.*, 26(1): 135-146, 2017.

[41] X. S. Wei, C. W. Xie, and J. Wu. Mask-CNN: Localizing Parts and Selecting Descriptors for Fine-Grained Image Recognition. *Pattern Recog.*, 76(1): 704-714, 2018.

[42] H. Yao, S. Zhang, C. Yan, Y. Zhang, J. Li, and Q. Tian. AutoBD: Automated Bi-Level Description for Scalable Fine-Grained Visual Categorization. *IEEE Trans. Image Process.*, 27(1): 10-23, 2018.

[43] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017,

[44] M. Sun, Y. Yuan, F. Zhou, and E. Ding. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. In *ECCV*, 2018,

[45] L. Wu, Y. Wang, X. Li, and J. Gao. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Trans. Cyber.*, 10.1109/TCYB.2018.2813971, 2018.

[46] Y. Zhang, X. S. Wei, J. Wu, J. Cai, J. Lu, V. A. Nguyen, and M. N. Do. Weakly Supervised Fine-Grained Categorization With Part-Based Image Representation. *IEEE Trans. Image Process.*, 25(4): 1713-1725, 2016.

[47] L. Xie, Q. Tian, M. Wang, and B. Zhang. Spatial Pooling of Heterogeneous Features for Image Classification. *IEEE Trans. Image Process.*, 23(5): 1994-2008, 2014.

[48] Q. Sun, Q. Wang, J. Zhang, and P. Li. Hyperlayer Bilinear Pooling with application to fine-grained categorization and image retrieval. *Neurocomput.*, 282(1): 174-183, 2018.

[49] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact Bilinear Pooling. In *CVPR*, 2015, 317-326.

[50] X. Wei, Y. Zhang, Y. Gong, J. Zhang, and N. Zheng. Grassmann Pooling as Compact Homogeneous Bilinear Pooling for Fine-Grained Visual Classification. In *ECCV*, 2018,

[51] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimedia*, 19(6): 1245-1256, 2017.

[52] X. He, and Y. Peng. Fine-Grained Image Classification via Combining Vision and Language. In *CVPR*, 2017, 7332-7340.

[53] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding Label Structures for Fine-Grained Feature Representation. In *CVPR*, 2016, 1114-1123.

[54] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*, 2014, 580-587.

[55] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018,

[56] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds200-2011 Dataset. *California Institute of Technology*, 2011.

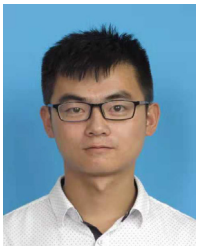
[57] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-

Grained Visual Classification of Aircraft. *HAL - INRIA*, 2013.
[58] J. Krause, M. Stark, D. Jia, and F. F. Li. 3D Object Representations for Fine-Grained Categorization. In *ICCVW*, 2013, 554-561.
[59] Y. Peng, X. He, and J. Zhao. Object-Part Attention Model for Fine-Grained Image Classification. *IEEE Trans. Image Process.*, 27(3): 1487-1500, 2018.

Junwei Han received his B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems in 1999, 2001, and 2003, respectively, all from Northwestern Polytechnical University, Xi'an, China, where he is currently a professor. He was a research fellow at Nanyang Technological University, The Chinese University of Hong Kong, Dublin City University, and the University of Dundee from 2003 to 2010. His research interests include computer vision and brain-imaging analysis. He is an associate editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Human-Machine Systems, Neurocomputing, and Machine Vision and Applications.

Xiwen Yao received his B.S. and Ph.D. degree from the Northwestern Polytechnical University, China, in 2010 and 2016, respectively. He is currently a research assistant of Northwestern Polytechnical University. His research interests include computer vision and remote sensing image processing, especially on fine-grained image classification and object detection.

Gong Cheng received the B.S. degree from Xidian University, Xi'an, China, in 2007, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2013, respectively. He is currently a Professor with Northwestern Polytechnical University, Xi'an, China. His main research interests are computer vision and pattern recognition.



Xiaoxu Feng received the B.E. degree from the Inner Mongolia University, Hohhot, China, in 2017. He is currently working toward the PhD degree at Northwestern Polytechnical University. His research interests include computer vision and image processing, especially on object detection and scene classification.

Dong Xu received his B.E. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, in 2001 and 2005, respectively. He is currently a professor with the School of Electrical and Information Engineering, the University of Sydney, Australia. He is a Fellow of the IEEE and a Fellow of the International Association of Pattern Recognition.