CA-PMG: Channel attention and progressive multi-granularity training network for fine-grained visual classification

Peipei Zhao¹ Qiguang Miao¹ Hang Yao¹ Xiangzeng Liu¹ Ruyi Liu¹ Maoguo Gong²

¹ School of Computer Science and Technology, Xidian University, Xi’an, Shaanxi, China
² School of Electronic Engineering, Xidian University, Xi’an, Shaanxi, China

Correspondence
Qiguang Miao, School of Computer Science and Technology, Xidian University, Xi’an 710071, Shaanxi, China. Email: qgmiao@xidian.edu.cn; xzliu@xidian.edu.cn

Abstract
Fine-grained visual classification is challenging due to the inherently subtle intra-class object variations. To solve this issue, a novel framework named channel attention and progressive multi-granularity training network, is proposed. It first exploits meaningful feature maps through the channel attention module and captures multi-granularity features by the progressive multi-granularity training module. For each feature map, the channel attention module is proposed to explore channel-wise correlation. This allows the model to re-weight the channels of the feature map according to the impact of their semantic information on performance. Furthermore, the progressive multi-granularity training module is introduced to fuse features cross multi-granularity. And the fused features pay more attention to the subtle differences between images. The model can be trained efficiently in an end-to-end manner without bounding box or part annotations. Finally, comprehensive experiments are conducted to show that the method achieves state-of-the-art performances on the CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets. Ablation studies demonstrate the effectiveness of each part in our module.

1 INTRODUCTION

Fine-grained image classification (FGVC) aims to distinguish multiple sub-categories within a super-category, such as different species of birds, and identifying the models or makes of vehicles. Different from conventional classification task (e.g. ImageNet classification [1]) that is to distinguish basic-level categories, fine-grained image classification has visually similar categories, and each category contains fewer training samples. Moreover, fine-grained image classification is very challenging as subcategories tend to own small variance in object appearance and thus can only be classified by some subtle or local differences. For example, we discriminate breeds of birds depending on the shape of their beak or the colour of their back. Therefore, fine-grained visual classification is a much more challenging computer vision problem than conventional classification owing to the subtle intra-class object variations among sub-categories.

In recent years, fine-grained visual classification performance has been significantly improved on species of birds, cars and aircrafts mainly due to the development of convolutional neural networks (CNNs). Early works mostly learnt object part localisation and feature representation with the assistance of manual annotations [2–6]. However, manual object part annotations are difficult to obtain. Weakly-supervised frameworks have received more and more attention in recent researches [7–10]. These methods can be able to locate more discriminative local regions. In a recent work, called the Progressive multi-granularity (PMG) [11], this approach encourages the model to learn the complementary information across different image granularities. In this paper, we focus on investigating a novel network that combines channel attention with the PMG training framework.
Multi-granularity training framework has been explored for fine-grained visual classification to address the large intra-class variations. The framework works in steps during training, where at each step the training pays attention to cultivating granularity-specific information with a corresponding stage of the network. It starts with more stable and finer granularities, and gradually moves onto coarser granularities to avoid the confusion caused by large intra-class variations. Although this framework can focus on more discriminative information, it ignores the channel relationship. To tackle this issue, we investigate the channel relationship which can focus on each channel according to different weights. Extensive experiments demonstrate the effectiveness of our approach.

Main contributions of this paper can be summarised as follows:

- We propose a channel attention module to obtain more discriminative features. The module can improve the representational ability of a network by explicitly modelling the interdependencies between the channels of feature maps.
- We investigate the progressive training strategy to fuse features from different granularities. Meanwhile, we study a random jigsaw patch generator that can encourage the network to learn features at specific granularities.
- We empirically demonstrate and confirm the effectiveness of our approach using deep neural networks, such as ResNet [12] trained for image classification tasks on various datasets including Caltech-UCSD birds (CUB-200-2011) [13], Stanford cars (CAR) [14] and FGVC-aircraft (AIR) [15].

The rest of this paper is organised as follows. Section 2 briefly reviews the related work. Section 3 describes our approach that contains the random jigsaw patch generator, the channel attention module and the progressive training strategy. After, the experiments and analysis are shown in Section 4. Finally, we conclude this work in Section 5.

2 RELATED WORK

2.1 Fine-grained visual classification

The conventional image classification has many applications, such as ImageNet classification, gender classification [16], point cloud classification [17], railway track surface defect classification [18], hyperspectral image classification [19, 20] and the other classification tasks [21–23]. These tasks belong to coarse classification. For example, there is a dog in an image. In the ImageNet classification task, the image can be classified as dog. But we cannot know the specific dog species. And fine-grained image classification can classify the specific dog species.

Benefiting from the development of neural networks, early FGVC research mainly focused on strongly-supervised methods with extra annotations such as bounding box [2, 24–26]. From a set of images in a special domain, labelled with part locations and class, [2] proposed a novel part-stacked CNN architecture that consisted of a fully convolutional network and a two-stream classification network. Based on manually-labelled strong part annotations, the fully convolutional network could locate multiple object parts and the two-stream classification network that encoded part-level and object-level cues simultaneously. [24] presented a method, called part-based one-versus-one features (POOFs), that could learn a large set of discriminative intermediate-level features. To make full use of part information, [25] proposed an effective flowchart named hierarchical part matching (HPM). The HPM introduced several novel modules to integrate into image representation, including geometric phrase pooling (GPP), hierarchical structure learning (HSL) and foreground inference and segmentation. [26] proposed a part localisation approach which leveraged deep convolutional features computed on bottom-up region proposals.

Recent studies about fine-grained visual classification have moved from strongly-supervised learning strategy with extra annotations to weakly-supervised learning strategy with only category labels [9, 10, 27–29]. The weakly supervised learning methods focused on locating more complementary parts, the most discriminative parts and parts of various granularities. [9] designed a new multi-agent cooperative learning scheme to effectively localise informative regions without fine-grained bounding-box or part annotations. [10] proposed a new part learning approach, named a multi-attention convolutional neural network (MA-CNN), where part generation and feature learning could reinforce each other. [27] proposed a novel recurrent attention convolutional neural network (RA-CNN) for FGVC without part annotations or bounding box. The method could learn discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. To avoid missing any important object parts, [28] proposed a novel representation method, called weakly supervised complementary parts model, that retrieved information suppressed by dominant object parts detected by convolutional neural networks. To overcome the difficulty of learning diverse experts from limited data, [29] promoted diversity among experts by combining an expert gradually-enhanced learning strategy and a Kullback–Leibler divergence based constraint.

There are few methods to fuse information from these discriminative parts together better. But how to fuse features from different parts is still a challenge in the field of fine-grained image classification. In this work, we propose a novel method that can fuse small-granularity and large-granularity features.

2.2 Visual attention

Visual Attention has been widely used in a variety of computer vision applications. It can capture subtle inter-class differences in FGVC. For example, [30] proposed a method that consisted of two parts: a differentiable one-squeeze multi-excitation (OSME) model and a multi-attention multi-class (MAMC) constraint. The OSME captured features from multiple attention regions to localise different parts. And the MAMC could enforce the correlations among different parts. To deal with intra-class variances and low inter-class variances, [31] proposed an attention convolutional binary neural tree which
incorporated convolutional operations along the edges of the three structures and utilised the routing functions in each node to determine the computational paths from root to leaf within the tree as deep neural networks. [32] proposed a novel model, called “filtration and distillation learning” (FDL), that enhanced the region attention of discriminative parts for FGVC. [33] proposed a novel recurrent attention network which paid attention to marginal differences to obtain more representative features. The multi-attention convolutional neural network (MA-CNN) [10] jointly learnt part proposals and the feature representations on each part.

In this paper, we integrate the channel attention into the progressive training strategy to capture more discriminative regions for FGVC.

### 2.3 Progressive training

Progressive training approach was originally proposed for generative adversarial networks [34]. It started with low-resolution images, and then progressively increased the resolution by adding layers to the networks. Recently, to deal with extreme super-resolution scenarios, [35] proposed a progressive cascading residual network that alleviated the instability of training caused by the training very deep convolutional neural network on extremely low-resolution inputs. [36] designed a progressive multi-scale approach which reconstructed a high-resolution image in intermediate steps by progressively performing a 2× upsampling of the input from the previous level. [37] introduced a novel unconditional generative model, called SinGAN, that could be trained on a single natural image. SinGAN could be utilised within a simple unified learning framework to deal with various image manipulation tasks, such as harmonisation, animation, super-resolution, editing and paint-to-image from a single image. It can be seen from recent works that progressive training strategy has been widely used for generation tasks.

For FGVC, [11] designed a new progressive training strategy to fuse information from previous levels of granularity and cultivate the inherent complementary properties across different granularities. In this paper, we integrate the channel attention into the progressive training to design a simple but effective network that can learn more discriminative information with a series of training stages. The main difference between our approach and the PMG is that our approach proposes a channel attention module to reweight the channels of the feature map according to the impact of their semantic information on performance.

### 3 APPROACH

In this section, we present our proposed training framework, called channel attention and progressive multi-granularity training network (CA-PMG). It can efficiently and accurately focus on discriminative regions without any part and object annotations. As shown in Figure 2, the framework of our approach contains two parts: 1) A differentiable channel attention module that captures features from attention regions. 2) A progressive multi-granularity (PMG) training module that learns stable fine-grained features in the shallower layers and progressively moves attention to the learning of abstract features of large granularity level in the deeper layers.

#### 3.1 Jigsaw puzzle generator

Jigsaw puzzle solving [38] was utilised for self-supervised task in representation learning. [11] introduced a jigsaw puzzle generator to produce input images for different steps of progressive training. In this paper, our goal is to design different granularity regions and force the network to exploit information specific to the corresponding granularity level at each training step. Then, we also utilise the jigsaw puzzle generator to form different input images. Assuming \( P \in R^{3 \times W \times H} \), we divide it into \( n \times n \) patches. And, the size of the patch is \( 3 \times \frac{W}{n} \times \frac{H}{n} \). As shown in Figure 1, these patches are randomly shuffled and merged into a new image \( P' \in R^{3 \times W' \times H'} \).

The hyper-parameter \( n \) controls the granularities of patches. The choice of hyper-parameter \( n \) needs to meet two conditions. The first one is that the patch size should be smaller than the receptive field of the corresponding stage, otherwise, the performance of the network will be reduced. The second one is that the size of the patch should be proportional to the receptive fields of the stages. Normally, the receptive field of each stage is about double than that of the previous stage. Therefore, we set \( n = 2^{l-1}+1 \).

As shown in Figure 1, we utilise the jigsaw puzzle generator to process the initial image and obtain a new image that is the input of our network. The new image uses the same label \( y \) as the ground truth. We input the new image and obtain the output \( y_j \) of the \( l \)th stage. Then, we use the cross entropy loss to optimise the parameters of our module.
It should be clarified that the size of the patch is smaller with the increase of \( n \). The smaller patch cannot guarantee the completeness of all the parts. Therefore, we should appropriately choose the size of \( n \). We will discuss the procedure in the Section 4.2.

### 3.2 Channel attention

The visual attention models have been used to explore weakly supervised part localisation. The previous works can be roughly divided into two categories. Firstly, attention is known as part detection. In the other words, each attention captures a certain area. For example, [39] proposed a novel recurrent neural network model that was able to capture information from an image or video by adaptively selecting a sequence of regions and locations. The model only processed the selected regions. Ref. [40] introduced a novel learnable module, called the spatial transformer, which used a spatial transformation to a feature map. [27] presented a novel recurrent attention convolutional neural network that could capture discriminative region attention. Secondly, attention is used for imposing a soft mask on the feature map, which starts from activation visualisation [41, 42]. Later, it is extended for improving the recognition performance [43, 44] and localising parts [45, 46]. Our method also falls into this category. We adopt the channel attention to extract and describe discriminative regions in the input image.

As shown in Figure 2, our framework can be implemented on any backbone feature extractor, such as ResNet50 and ResNet101 [12]. Let us \( \mathbf{X} \in \mathbb{R}^{H \times W \times C} \), where \( H \), \( W \), \( C \) are the height, width and number of channels of an input image. \( \mathbf{F}_j : \mathbf{X} \rightarrow \mathbf{U}_j, \mathbf{U}_j \in \mathbb{R}^{(H/W) \times (W/W) \times C_j} \). \( \mathbf{F}_j \) denotes a series of convolutional operators and \( \mathbf{U}_j \) denotes the feature map at \( j \)th stage, where \( j = 1, 2, \ldots, L \). We can write the feature map

\[
\mathbf{U}_j = [u_1^j, u_2^j, \ldots, u_C^j]. \tag{1}
\]

In order to exploit channel dependencies, we first utilise global average pooling to generate channel-wise statistics. The attention module aggregates the feature map \( \mathbf{U}_j \) across spatial dimensions \( H_j \times W_j \) to generate a channel descriptor \( \mathbf{z}_j = [z_1^j, z_2^j, \ldots, z_C^j] \in \mathbb{R}^C \). The \( r \)th channel of \( \mathbf{z}_j \) is calculated by:

\[
z_r^j = \frac{1}{H_j \times W_j} \sum_{i=1}^{H_j} \sum_{w=1}^{W_j} \mathbf{U}_j(i, w). \tag{2}
\]

Then, to capture channel dependencies, we use a gating mechanism to learn a nonlinear interaction between channels.

\[
\mathbf{m}_j = \sigma(W_2 \delta(W_1 \mathbf{z}_j)), \tag{3}
\]

where \( \sigma \) and \( \delta \) denotes the Sigmoid and ReLU [47] functions respectively, \( \mathbf{W}_1 \in \mathbb{R}^{C \times C_j} \) and \( \mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{2}} \). \( \mathbf{m}_j \) denotes a non-mutually-exclusive relationship among channels. Therefore, our attention module utilises it to re-weight the channels of the feature map \( \mathbf{U}_j \),

\[
\mathbf{S}_j = [m_1^j u_1^j, m_2^j u_2^j, \ldots, m_C^j u_C^j] \in \mathbb{R}^{H \times W \times C_j}, \tag{4}
\]

where \( \mathbf{S}_j \) refers to an attention map.

By adopting the idea of attention, our approach implements a simple yet effective attention mechanism that can re-weight the channels of the feature map. And the channel attention module
can enable an end-to-end training on fine-grained visual classification and help to boost feature discriminability.

### 3.3 Progressive training

We propose to build our module based on the channel attention and progressive learning strategy to effectively capture discriminative information. The progressive training starts from the low stage and then progressively adds new stages for training. Due to the limitation of the receptive field and representation ability at the low stage, the model will be forced to capture discriminative information from local details. With the increase of stages, the network gradually locates discriminative information from local details to global structure.

Our goal of introducing the progressive training strategy is to impose classification loss on different intermediate stages. Hence, we introduce convolution block $B_i^{conv}$ that takes $i$th stage output $S_i$ as input and reduces it to a vector representation $V_i = B_i^{conv}(S_i)$. To predict the probability distribution, we define a classification module $B_i^{class}$ that consists of the Batchnorm [48], the Elu [49] and two fully-connected stages. We input the $V_i$ into the classification module. Then, we obtain the probability distribution $y_i = B_i^{class}(V_i)$.

In general, our module adopts cross entropy loss $L_{CE}$ to minimise the distance between ground truth label $y$ and prediction probability distribution $y_i$. The $L_{CE}$ can be calculated by:

$$L_{CE}(y_i, y) = - \sum_{i=1}^{m} y \times \log(y_i), \quad (5)$$

where $m$ is the number of categories. $y_i$ represents the probability that the input $X$ belongs to the $i$th category at the $l$ stage.

To improve the performance of fine-grained image classification, we concatenate the outputs of the last multiple stages as:

$$V_{concat} = concat(V_{L-\Delta+1}, ..., V_{L-1}, V_L), \quad (6)$$

where $\Delta$ is the number of last stages. This is followed by a classification model $y_{concat} = H_{concat}^{class}(V_{concat})$. Then, we use the cross entropy loss to optimise the model. The formula is as follows:

$$L_{CE}(y_{concat}, y) = - \sum_{i=1}^{m} y \times \log(y_{concat}). \quad (7)$$

In order to help each stage in the model work together, all parameters whose are utilised in the current prediction will be optimised, although they have been updated in the prior step.

We obtain the prediction probability distributions $y_i$ and $y_{concat}$. If we only utilise $y_{concat}$ as the prediction, the final result of our module can be written as:

$$C_1 = \arg\max(y_{concat}). \quad (8)$$

### 4 EXPERIMENT RESULTS AND DISCUSSION

In this section, we will describe the datasets utilised in this paper, the implementation details and experiment results. We evaluate the performance of our proposed approach on three challenging fine-grained image classification datasets: Caltech-UCSD birds (CUB-200-2011) [13], Stanford cars (CAR) [14] and FGVC-aircraft (AIR) [15].

CUB-200-2011 dataset includes 200 categories for fine-grained visual classification. And each category has about 30 training images. The dataset also contains 5994 training samples and 5794 test samples.
TABLE 1 Comparison results with state-of-the-art methods. our method$^*$ is the result of $y_{concat}$ and combined accuracy$^*$ is the combination of multiple results as shown in Figure 2. The best results are indicated in bold, and the second-best results are indicated in underline.

| Method              | Reference | Backbone | CUB-200-2011 | stanford cars | Aircraft |
|---------------------|-----------|----------|--------------|---------------|----------|
| DFL-CNN [8]         | CVPR18    | ResNet50 | 87.4         | 93.1          | 91.7     |
| NTS-Net [9]         | ECCV18    |          | 87.5         | 93.9          | 91.4     |
| TASN [50]           | CVPR19    |          | 87.9         | 93.8          | -        |
| DCL [7]             | CVPR19    |          | 87.8         | 94.5          | 93.0     |
| MGE-CNN [29]        | ICCV19    |          | 88.5         | 93.9          | -        |
| S3N [51]            | ICCV19    |          | 88.5         | 94.7          | 92.8     |
| DBTNet [52]         | NIPS19    |          | 87.5         | 94.1          | 91.2     |
| MC-Loss [53]        | TIP20     |          | 87.3         | 93.7          | 92.6     |
| DF-GMM [54]         | CVPR20    |          | 88.8         | 94.8          | 93.8     |
| ACNet [31]          | CVPR20    |          | 88.1         | 94.6          | 92.4     |
| CIN [55]            | AAAI20    |          | 87.5         | 94.1          | 92.6     |
| FDL [32]            | AAAI20    |          | 88.6         | 94.3          | 93.4     |
| API-Net [56]        | AAAI20    |          | 87.7         | 94.8          | 93.0     |
| Our method          | ResNet50  |          | 88.4         | 95.1          | 93.1     |
| Our method (combined accuracy) & ResNet50 & 89.2 & 95.2 & 93.3 |
| Our method          | ResNet101 |          | 88.1         | 95.1          | 93.4     |
| Our method (combined accuracy) & ResNet101 & 89.2 & 95.2 & 93.8 |

Stanford cars dataset contains 196 categories. There are 8144 training samples and 8042 test samples.

Aircraft dataset contains 3334 training samples, 3333 validation samples and 3333 test samples with 100 categories. There are about 90 images available for each category. Aircraft samples from the dataset are taken from various angles, and the categories are assigned based on aircraft model and production year.

4.1 Implementation details

We implement all experiments using PyTorch [57] with version higher than 1.6.0 over three Tesla P40 GPUs. Our proposed approach is evaluated on the widely used backbone network ResNet50 and ResNet101. The total number of stages in ResNet is $L = 5$. In order to get the best performance, we set $S = 3$. In the training phase, the input images are resized to $550 \times 550$ and randomly cropped to $448 \times 448$ with random horizontal flipping. During testing, we augment inputs by resizing images to $550 \times 550$ and cropping from centre into $448 \times 448$.

In this paper, we utilise stochastic gradient descent (SGD) to optimise our network. The initial learning rates are 0.0002 for pre-trained convolutional layers, and a $10\times$ multiplier is used for newly added layers. The learning rate is reduced by the cosine annealing schedule [58] during training. SGD optimiser is utilised with momentum 0.9 and weight decay 0.0005. We train our module for 200 epochs with batch size 16 and measure the top-1 classification accuracy from the last epoch.

4.2 Comparisons with state-of-the-art methods

To fully verify the effectiveness of our method, Table 1 provides quantitative experimental results on CUB-200-2011, Stanford Cars and FGVC-aircraft.

We first analyse the results on the CUB-200-2011 dataset in Table 1. The multi-stage outputs combined result of our approach achieves state-of-the-art performance. Compared with DFL-CNN [8] which enhances the mid-level learning capability of CNNs by learning a bank of convolutional filters to extract class-specific discriminative patches, our method obtains a better result with the improvement of 1.8%. Our method exceeds the MGE-CNN [29] by 0.7%, even though it exploits information of various granularities by building several different networks. Compared with DF-GMM [54], our method boosts the accuracy by 0.4%.

As shown in Table 1, our approach exhibits similar performance on Stanford cars. On Stanford Cars, our approach obtains state-of-the-art performance with ResNet50 and ResNet101 as the base model. The cars in the Stanford cars are much more rigid and the performance of $y_{concat}$ is good enough, so that the performance of combining multi-stage outputs has not been significantly improved. Meanwhile, our performance of combining multi-stage outputs exceeds the second best methods DF-GMM [54] and API-Net [56] by 0.4%, and exceeds the third best method S3N [51] by 0.5%.

It is observed that our method achieves competitive result on the FGVC-aircraft dataset. On Aircraft, our approach obtains
The importance of channel attention

The progressive training strategy

Ablation study

The jigsaw puzzle generator

Accuracy "accuracy" is the result of $y_{concat}$ and "combined accuracy" is the combination of multiple results as shown in Figure 2.

Table 2: The accuracy and combined accuracy of our proposed approach by using different hyper-parameters $s$ without the assistance of jigsaw puzzle generator. "accuracy" is the result of $y_{concat}$ and "combined accuracy" is the combination of multiple results as shown in Figure 2.

| $s, n$ | Accuracy (%) | Combined accuracy (%) |
|-------|--------------|-----------------------|
| 1, {1,1} | 85.6 | 85.8 |
| 2, {1,1,1} | 87.6 | 88.2 |
| 3, {1,1,1,1} | 87.9 | 88.4 |
| 4, {1,1,1,1,1} | 87.5 | 88.0 |
| 5, {1,1,1,1,1,1} | 86.7 | 87.2 |

State-of-the-art performance with ResNet101 as the base model. In ResNet101, compared with the result of $y_{concat}$, the combined result still obtains better result (0.4%). In ResNet50, our method exceeds all of the comparison methods except DF-GMM and FDL [32]. In ResNet101, our approach outperforms MC-Loss [53] by 1.2%, even though it builds a mutual-channel loss to delve into individual feature channels. For CIN [55] which first learns the complementary features from the correlated channels and then distinguishes the subtle visual differences between images for final classification, we outperform it by 1.2%.

4.3 Ablation study

We conduct ablation studies to verify the effectiveness of the jigsaw puzzle generator, the progressive training strategy and the channel attention. We conduct experiments on the CUB-200-2011 dataset and use ResNet50 as the backbone network. The total number of stages $L$ in ResNet50 is 5.

4.3.1 The progressive training strategy

To validate the effectiveness of the progressive training, we design experiments without jigsaw puzzle generator. As shown in Table 2, $S$ increases from 1 to 5. The $y_{concat}$ is kept for all runs and number of steps is $S + 1$. Compared with results in

Table 2, when $S < 4$ the increasing of $S$ improves the performance of model. When $S = 4$, both the accuracy and the combined accuracy start to decrease. This may be because low stage layers mainly pay attention to features that are irrelevant to the class. But the additional supervision force low stage layers to focus on class-relevant features. Then it affects the decrease in the overall performance.

4.3.2 The jigsaw puzzle generator

In Table 3, we analyse the results of our approach with the help of the jigsaw puzzle generator on CUB-200-2011 dataset. The hyper-parameter $s = 2^l - l + 1$, where $l$ denotes /th stage. In Table 3, when $S < 4$, both the accuracy and the combined accuracy of our model on the basis of the progressive training are proportional to the value of $S$. These indicate that the jigsaw puzzle generator can boost the performance of our model when $S < 4$. When $S \geq 4$, both the accuracy and the combined accuracy of our model decrease with the increase in the value of $S$. These denotes that the jigsaw puzzle generator does not show any advantages when $S \geq 4$. The possible reason is that when $S \geq 4$ the split patches are too small to keep useful information, which confuses our module training.

4.3.3 Importance of channel attention

Channel attention (CA) is important in capturing discriminative regions. In Table 4, we discuss the results of our approach with the help of the channel attention. From Table 3, it can be concluded that when $S = 3$ our approach obtains the best results.
Therefore, we set $\delta = 3$ in this part. For $y_{\text{concat}}$, using channel attention boosts the accuracy by 0.2% and 0.1% than without CA on cars and air datasets respectively. We also notice that for combining multi-stage outputs, using CA can offer 0.2%, 0.2% and 0.1% performance improvement compared to the method without the CA (89.0% vs. 89.2%, 95.0% vs. 95.2% and 93.2% vs. 93.3%) on CUB, cars and air datasets. These indicate the effectiveness of the channel attention module.

4.3.4 With and without pretraining

We know that the above experiments are based on ImageNet pre-training. To validate the effectiveness of our method, we make the experiments with and without pre-training in Table 5. As shown in Table 5, the accuracy of the models without pre-training is lower than that of the mode with pre-training. However, the model without pre-training and the model with pre-training have the same trend that the combined accuracy is higher than the accuracy. It indicates the effectiveness of our proposed method.

4.4 Visualisation

In order to illustrate the advantages of our proposed approach, we apply the Grad-CAM to implement visualisation for last three stages’ convolution layer of both baseline model and our approach. As shown in Figure 3, columns (d)–(f) are the visualisation of the convolution layers from the third to the fifth stage of our model’s backbone, which combines the channel attention and the progressive training strategy. From (a) and (b), we know that the model of ResNet50 is disturbed by the background. Compared (a) with (d), we can know that our model pays more attention to discriminative parts and reduces background the attention to background. Meanwhile, compared (b) with (e), it is concluded that our model reduces the focus on the background. This indicates that our module can help the model locate useful information at earlier stages. Therefore, our approach can improve the representational ability of a network.

5 CONCLUSION

In this paper, we have proposed a novel CNN that combines the progressive training strategy with the channel attention for fine-grained visual classification. Firstly, we have introduced a simple jigsaw puzzle generator to produce input images that contain information of different granularity levels. Then, we have utilised the channel attention module for re-weighting the channels of feature map to extract discriminative features. Finally, we have introduced a novel training strategy that can fuse multi-granularity features in a progressive manner. Our method does not require part annotation or bounding box, and can be trained in an end-to-end manner. Extensive experiments on three challenging fine-grained datasets demonstrate that our approach obtains state-of-the-art performances.
ACKNOWLEDGMENT
The work was jointly supported by the National Key R&D Program of China under Grant No. 2018YFC0807500, the National Natural Science Foundations of China under grant No. 61772396, 61772392, 61902296 and Xi’an Key Laboratory of Big Data and Intelligent Vision under grant No.201805053ZD4CG37, the National Natural Science Foundation of Shaanxi Province under grant No. 2020JQ-330, 2020JM-195, the China Postdoctoral Science Foundation under grant No. 2019M663640, Guangxi Key Laboratory of Trusted Software (No. KXX20261), National Key Research and Development Project under Grant No.B019030051.

ORCID
Qiguang Miao https://orcid.org/0000-0002-2872-388X

REFERENCES
1. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115(3), 211–252 (2015)
2. Huang, S. et al.: Part-stacked cnn for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1173–1182. IEEE, Piscataway, NJ (2016)
3. Lei, J., et al.: Fast mode decision based on grayscale similarity and inter-view correlation for depth map coding in 3D-HEVC. IEEE Trans. Circuits Syst. Video Technol. 28(3), 706–718 (2016)
4. Lin, D. et al.: Deep lac: Deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1666–1674. IEEE, Piscataway, NJ (2015)
5. Krause, J. et al.: Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5546–5555. IEEE, Piscataway, NJ (2015)
6. Branson, S. et al.: Bird species categorization using pose normalized deep convolutional nets. arXiv:1406.2952, (2014)
7. Chen, Y. et al.: Distraction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5157–5166. (2019)
8. Wang, Y., Morariu, V.I., Davis, L.S.: Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4148–4157. IEEE, Piscataway, NJ (2018)
9. Yang, Z. et al.: Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision, pp. 420–435. IEEE, Piscataway, NJ (2018)
10. Zheng, H. et al.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217. IEEE, Piscataway, NJ (2017)
11. Du, R. et al.: Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: European Conference on Computer Vision, pp. 153–168. Springer, Berlin, Heidelberg (2020)
12. He, K. et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. IEEE, Piscataway, NJ (2016)
13. Wah, C. et al.: The caltech-ucsd birds-200-2011 dataset. (2011)
14. Krause, J. et al.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops, pp. 554–561. IEEE, Piscataway, NJ (2013)
15. Maji, S. et al.: Fine-grained visual classification of aircraft. arXiv:1306.5151, (2013)
16. Alghahili, M., Li, Z., Ali, H.A.: Deep feature learning for gender classification with covered/camouflaged faces. IET Image Proc. 14(15), 3957–3964 (2020)
17. Zhang, X., et al.: Hybrid feature CNN model for point cloud classification and segmentation. IET Image Proc. 14(16), 4086–4091 (2020)
18. Ye, J., et al.: Method for automatic railway track surface defect classification and evaluation using a laser-based 3d model. IET Image Proc. 14(12), 2701–2710 (2020)
19. Wang, Q., He, X., Li, X.: Locality and structure regularized low rank representation for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57(2), 911–923 (2019)
20. Wang, Q., Lin, J., Yuan, Y.: Salient band selection for hyperspectral image classification via manifold ranking. IEEE Trans. Neural Networks Learn. Syst. 27(6), 1279–1289 (2016)
21. Kurmi, Y., Chaurasia, V.: Classification of magnetic resonance images for brain tumour detection. IET Image Proc. 14(12), 2808–2818 (2020)
22. Chandran, M.C., Jose, M.V.: Optimised hybrid classifiers for automatic hep-2 cell classification. IET Image Proc. 14(16), 4316–4328 (2020)
23. Xu, Y., et al.: Underwater sonar image classification using generative adversarial network and convolutional neural network. IET Image Proc. 14(12), 2819–2825 (2020)
24. Berg, T., Belbeaufre, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 955–962. IEEE, Piscataway, NJ (2013)
25. Xie, L. et al.: Hierarchical part matching for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1641–1648. IEEE, Piscataway, NJ (2013)
26. Zhang, N. et al.: Part-based r-cnn for fine-grained category detection. In: European conference on computer vision. pp. 834–849. Springer, Berlin, Heidelberg (2014)
27. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4438–4446. IEEE, Piscataway, NJ (2017)
28. Ge, W., Lin, X., Yu, Y.: Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3034–3043. IEEE, Piscataway, NJ (2019)
29. Zhang, L. et al.: Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8331–8340. IEEE, Piscataway, NJ (2019)
30. Sun, M. et al.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 805–821. Springer, Berlin, Heidelberg (2018)
31. Ji, R. et al.: Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10 468–10 477. IEEE, Piscataway, NJ (2020)
32. Liu, C. et al.: Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11 555–11 562, ACM Press, New York (2020)
33. Shroff, P. et al.: Focus longer to see better: Recursively refined attention for fine-grained image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 868–869. IEEE, Piscataway, NJ (2020)
34. Karras, T. et al.: Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196, (2017)
35. Ahn, N., Kang, B., Sohn, K.-A.: Image super-resolution via progressive cascading residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 791–799. IEEE, Piscataway, NJ (2018)
36. Wang, Y. et al.: A fully progressive approach to single-image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 864–873. IEEE, Piscataway, NJ (2018)
37. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4570–4580. IEEE, Piscataway, NJ (2019)
38. Wei, C. et al.: Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1910–1919. IEEE, Piscataway, NJ (2019)

39. Mnih, V. et al.: Recurrent models of visual attention. arXiv:1406.6247, (2014)

40. Jaderberg, M. et al.: Spatial transformer networks. arXiv:1506.02025, (2015)

41. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833. Springer, Berlin, Heidelberg (2014)

42. Zhou, B. et al.: Object detectors emerge in deep scene CNNs. arXiv:1412.6856, (2014)

43. Wang, F. et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164. IEEE, Piscataway, NJ (2017)

44. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. IEEE, Piscataway, NJ (2018)

45. Zhou, B. et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929. IEEE, Piscataway, NJ (2016)

46. Rosenfeld, A., Ullman, S.: Visual concept recognition and localization via iterative introspection. In: Asian Conference on Computer Vision, pp. 264–279. Springer, Berlin, Heidelberg (2016)

47. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 807–814. ACM Press, New York (2010)

48. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp. 448–456. PMLR Press, Cambridge, MA (2015)

49. Clevert, D.-A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (ELUs). arXiv:1511.07289, (2015)