A weakly supervised spatial group attention network for fine-grained visual recognition

Jiangjian Xie[1,2,3] · Yujie Zhong[1] · Junguo Zhang[1,2] · Changchun Zhang[1,2] · Björn W Schuller[3,4,5]

Abstract

The fine-grained visual recognition is to classify several sub-categories affiliated to the same basic-level category, which is highly challenging because the same sub-category with large variance and different sub-categories with small variance. Previously approaches generally localize the targets or parts first, then determine which sub-category the image is attached to. They depend on target or part annotations, which are labor-intensive and a barrier to moving towards practical use. Other methods indirectly extract recognizable areas from the high-level feature maps, ignoring the spatial relationships between the target and its parts, which may cause inaccurate recognition. In this paper, we propose a weakly supervised spatial group attention network (WSSGA-Net) for fine-grained bird recognition. According to the spatial relationships between the target and its parts, we embed the spatial group attention (SGA) module into the WSSGA-Net to highlight the correct semantic feature regions by establishing a semantic feature space enhancement mechanism. In addition, we apply moment exchange (MoEx) to generate new feature maps by exchanging two input image feature moments for data augmentation. Comprehensive experiments indicate that our approach significantly has a better performance than the state-of-the-art approaches on the standard bird image datasets Bird-65, CUB200-2011 and fine-grained dataset Stanford Cars.

Keywords Classification · Fine-grained image · Bird recognition · Weakly supervised network · Moment exchange · Spatial group attention

1 Introduction

Fine-grained visual recognition, in contrast to traditional visual recognition, is extremely challenging, aiming to recognize multiple sub-categories under the same basic-level category. Birds are one of those basic-level categories with hundreds of sub-categories, and precise recognition of birds is crucial for their conservation and scientific study [1]. However, there are two main challenges in bird image recognition:

1. *High intra-subcategory variance.* As illustrated in the first row of Fig. 1, the left group of four images belongs to the same sub-category of the Black Footed Albatross, but they are quite different in poses, views, feathers, and further more. It is possible for humans to incorrectly recognize them into different sub-categories. The same situation exists in the case of the right group of Black Stork examples.
2. *Low inter-subcategory variance.* The second row of Fig. 1 illustrated that four images of the left group belong to four different sub-categories, but their black appearance look similar. These sub-categories have similar global appearance, and they only have some discriminable areas of the bodies, such as the mouth. It is difficult for humans to distinguish them apart. The right group is in the same boat. It is also difficult to access precise

✉ Jiangjian Xie
shyneforce@bjfu.edu.cn

Changchun Zhang
zhangchangchun@bjfu.edu.cn

[1] School of Technology, Beijing Forestry University, Beijing 100083, P. R. China

[2] Research Center for Biodiversity Intelligent Monitoring, Beijing Forestry University, Beijing 100083, P. R. China

[3] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany

[4] GLAM – Group on Language Audio and Music, Imperial College London, London SW7 2AZ, Germany

[5] Centre for Interdisciplinary Health Research, University of Augsburg, Augsburg 86159, Germany

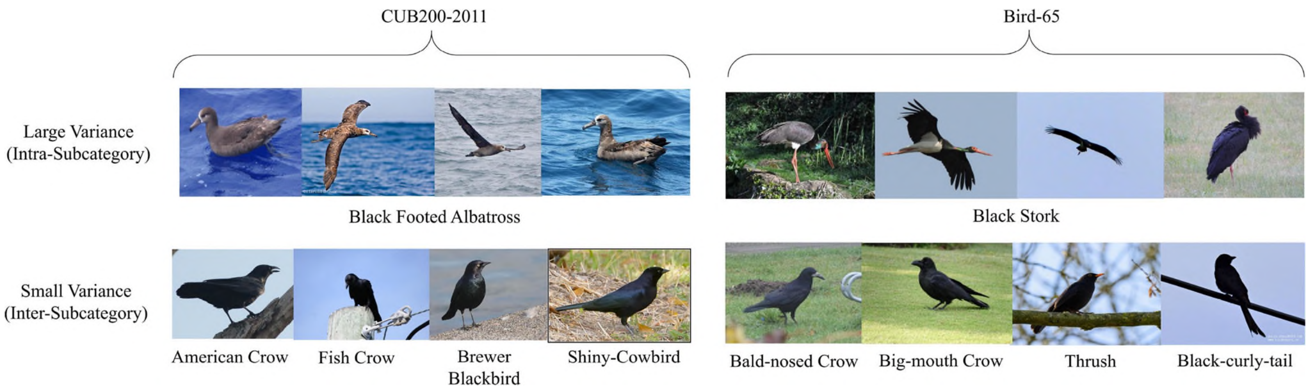


Fig. 1 Instances of the CUB-200-2011 and Bird-65 datasets. The first line illustrated the same sub-category with large variance, and the second line illustrated different sub-categories with small variance

classification accuracy by using only the current state-of-the-art coarse-grained convolutional neural networks (CNNs), such as ResNet [2], VGG [3] and Inception [4]. Therefore, instructing model to learn discriminable representations in detail is important for fine-grained visual recognition of birds. Fine-grained visual recognition bears the potential to detect subtle differences (such as subtle local differences) between sub-categories, then provide better classification performance than coarse-grained visual classification.

Currently, existing fine-grained recognition approaches have achieved great advancement. However, these approaches face an critical issue which needs to be addressed urgently: high labor consumption on image annotation. The image annotations (e. g., the image-level sub-category annotation, bounding boxes of the object, and part localization) are demanded in the training stage of numerous existing approaches, and even in the testing stage. The manual labelling are sometimes uncontractual and labor-consuming in the practical applications. To transform the classification of fine-grained image into applications, it is the optimal choice to use as few annotations as possible. This is the *first problem.*

To address the first problem, researchers start to concentrate on how to reach potential performance with the setting of weakly supervised where only image-level annotations are applied in both the training and testing stages. Weakly supervised fine-grained recognition indicates that the network utilizes image-level annotations to find subtle features of the targets in the images, which are then employed to recognize the targets. Zheng et al.[5] proposed the progressive attention convolutional neural network (PA-CNN), which contained two sub-networks to localize different parts of the target on multiple scales. Kim et al. [6] boosted the performance of fine-grained visual recognition in three stages: learning part-wise features, generating hard negative sample features and fine-grained visual recognition. However, when the discriminable regions were selected, they neglect the spatial relationships between parts and the target, but the spatial relationships were very useful for finding the discriminable regions in an intuitive understanding. This caused large regions of background noise and inaccurate recognition. This is the *second problem.*

To solve the second problem, the spatial group attention (SGA) module is included in this paper. It is useful to obtain the equivalent semantic features at the corresponding spatial location of the raw image in the specific semantic group. However, since the absence of supervision of region-specific details and the potential presence of noise in the image, the spatial allocation of semantic features can be somewhat confusing, which greatly weakens the representations of learning. To make each group of features spatially robust and well allocated, we apply the SGA module that scales the feature vectors at all locations using an attention mask within every feature group. We utilize this attention mask to eliminate possible noise and highlight the corresponding semantic feature areas. This simple and effective mechanism described above is the named SGA module that requires few extra parameters and computations.

Due to the rarity of some bird species and the secrecy of bird activity, there is a lack of their image data. Since limited data may lead to non-convergence, overfitting or local optima during model training, limited bird image data is the *third problem*. Data augmentation is a common solution that increases the amount of data for training through creating larger data variance. Many researches have proven that it is effective in computer vision field, such as object detection, image classification and image segmentation. Previous works usually use random data augmentations such as croping, droping, cutout, or cutmix to preprocess the original image. However, random data augmentations generate more noise, which may reduce the efficiency of training and affect the quality of feature extraction. Besides, to reduce training time and increase stability, the moments (i. e., standard deviation and mean) of potential features are usually eliminated as background noise. But the moments play a more important role in the image generation field. Researches have proven that the moments extracted from the normalization of instance and location can catch style and shape the image information roughly. These moments are essential for the generation process of the data augmentation.

To deal with the third challenge, we swap the feature moments of the two input images in the feature space to generate feature maps containing information about both the two images as new training data, and we name this method Moment Exchange (MoEx) for short. This method does not introduce background noise nor lose the semantic information of the original images. Therefore, it can effectively achieve the purpose of increasing the number of data for training to optimise the model performance.

In general, this paper has three main contributions:

1. We design a weakly supervised spatial group attention network (WSSGA-Net) for fine-grained bird image recognition.
2. The spatial group attention (SGA) module is applied in our WSSGA-Net approach to eliminate possible noise and highlight the key semantic feature regions to boost the model performance.
3. In the feature extraction stage of the WSSGA-Net training, we present the MoEx data augmentation approach to extract a new feature map fusing two input picture features, increasing the amount of data for training and improving network performance.

The remaining of the paper is composed as follows. Firstly, the related works are reviewed, including fine-grained image classification and data augmentation in Section 2,next the proposed weakly supervised spatial group attention network (WSSGA-Net) is described in Section 3. In Section 4, we conduct extensive experiments to demonstrate the effectiveness of WSSGA-Net. At last, we draw a conclusion in Section 5.

2 Related work

In this section, the related work of fine-grained image recognition and data augmentation is reviewed.

2.1 Fine-grained Image Recognition

Convolutional neural networks (CNN) were initially proposed for image classification. However, these basic models have low performance for fine-grained visual recognition, it is challenging to concern the subtle differences between the parts of objects without a unique design. Now, various approaches have been proposed to distinguish such distinct fine-grained categories.

Many methods use annotations of part locations and attributes to focus on local features. Zhang et al. [7] designed Part R-CNN to extend R-CNN [8] to classify targets and locate partial regions with a geometric prior, and then predict fine-grained categories from pose-normalized representations. Branson et al. [9] introduced a graph-based clustering algorithm for computing local features of the object poses, which is helpful for classification. Lin et al. [10] introduced a feedback structure called Deep Localization, Alignment and Classification (Deep LAC) incorporating localization, alignment and classification as three sub-networks. Additionally, Valve Linking Function (VLF) was designed to reduce alignment and classification errors, increasing part locating and assisting classification.

To reduce the cost of a traditional location annotation, weakly supervised approaches that require only image-level annotation have gradually emerged. There are three major techniques among existing weakly supervised approaches for fine-grained visual recognition: end-to-end feature coding, localization-classification subnetworks, and visual attention. These three classes of approaches are introduced sequentially in the following.

A. End-to-end feature coding based methods End-to-end feature encoding-based methods incline to extract discriminable features directly by establishing robust networks for fine-grained visual recognition. Lin et al. [11] proposed bilinear CNNs to represent an image as an ensemble outer product of features exported from two bilinear models, thus encoding higher-order statistics of convolutional activation and enhancing mid-level learning. Yu et al. [12] utilized the hierarchical bilinear pooling method (HBP) to capture the feature, which had better classification accuracy compared with other bilinear pooling methods. Min et al. [13] introduced an advanced method which simultaneously normalized the bilinear representation with square root, low rank, and sparsity called multi-objective matrix normalization method (MOMN), which were three regularizers that not only compressed the bilinear features and facilitated the generalization of the model, but also stabilized the second-order information. However, the above approaches with powerful models imply that high capacities and computational effort make them not conducive to applications.

B. Localization-Classification Subnetworks Based Methods The subnetwork for localization-classification is utilized to localize the distinctive parts of a region through a localization subnetwork. Next, the localization subnetwork located the features of the targets and then feed back to the classification subnetwork. Zhang et al. [14] designed an advanced CNN architecture which combined semantic part detection and abstraction (SPDA-CNN) for fine-grained image recognition, which consisted of two subnetworks: the subnetwork of detection was utilized for part localization and the subnetwork of classification was utilized to classify fine-grained image. Yang et al. [15] designed a multi-intelligence collaborative Navigator-Teacher-Scrutinizer Network (NTS-Net) to pinpoint the discriminative part. The training stage required no additional annotations, rather, the teacher agent assisted in locating the discriminative critical information by the navigation agent, and finally, these features are utilized for fine-grained recognition. Lin et al. [16] proposed an Increasing Specialized Generative Adversarial Network (IS-GAN), which was a three-scale framework consisting of a generative adversarial network for feature extraction and a patch proposal network for localization on each scale. Guo et al. [17] suggested a novel framework for progressive sampling to distinguish parts from coarse to fine scale detail learning, with three subnetworks for feature extraction at the whole, object and detail levels respectively. But these approaches have a problem: High time consumption, since each region proposal requires to pass two subnetworks respectively, and thousands of region proposals generally are generated from each image.

C. Visual Attention Based Methods Since the visual attention model-based methods can recognize discriminable targets in images with no extra annotation information, they have been extensively-used in the field of fine-grained visual recognition in current years. Fu et al. [18] utilized a recurrent attention CNN (RA-CNN) to recognize the position of an attention region and learn features of this region recursively, while the method focuses on only one part of the local area. So, they incorporated three scale features, i. e., three parts to achieve the final class. To recognize multi-attention regions simultaneously, Zheng et al. [19] utilized a Multi-Attention CNN (MA-CNN), which could localize several parts concurrently. To address heavy computational cost and a limited amount of parts in existing attention-based approaches, Zheng et al. [20] suggested a trilinear attention sampling network (TASN) to extract more fine-grained features, which was implemented by knowledge extraction in a student-teacher approach. Hu et al. [21] introduced a weakly supervised data augmentation network (WSDAN). This network can improve the recognition performance by generating an attention map to represent the discriminable part of the object through weakly supervised learning and performing data augmentation guided by the attention map. Zhang et al. [22] designed the Multi-branch and Multi-scale Attention Learning Network (MMAL-Net) containing two attention modules (AOLM and APPM) for localizing the objects and proposing differentiated components, respectively. Liu et al. [23] proposed a Subtler Mixed Attention Network (SMA-Net), which used a discriminative region localization module with a channel attention mechanism for region localization and a mixed attention module with feature extraction to focus on finer and differentiated regions. Ding et al. [24] designed an attention pyramidal convolutional neural network (AP-CNN). This model learnt high-level semantic and low-level detail features through a pyramidal hierarchy consisting of top feature paths and bottom attention paths. Wang et al. [25] presented an end-to-end Distinguished Feature Gaussian Mixture Model (DF-GMM). This model can alleviate the discriminable area spreading problem in higher-order feature mappings by adding a low-rank representation mechanism (LRM) to the model, enabling the discriminative region to be more accurately located on the new low-rank feature mapping. Unlike the above methods, the weakly supervised spatial group attention network (WSSGA-Net) proposed in our paper uses the SGA module to eliminate possible noise and highlight the corresponding semantic feature regions, which may obtain better fine-grained classification accuracy than the above methods.

2.2 Data augmentation

The existing data augmentation methods are mainly image-specific augmentation methods. Random space image augmentation approaches have been proposed and shown to be useful in enhancing the performance of deep learning networks, such as cropping and dropping.Gong et al. [26] proposed the KeepAugment method to preserve salient features and augment non-salient regions of an image to improve fidelity and increase diversity. Yun et al. [27] introduced the CutMix data augmentation method to fill a random region of the current image with a patch of another image. Yoo et al. [28] presented the CutBlur data augmentation technique, which improved diversity by cutting blocks of high-resolution images and pasting them to corresponding low-resolution images. Cubuk et al. [29] utilized Auto Augmentation to provide a search space for data augmentation strategies. It can utilize concrete strategies automatically to access the best accuracy for validation of the target dataset. Compared with data augmentation in the input space, feature space-based data augmentation methods can improve model performance more efficiently. Li et al. [30] designed an implied data augmentation approach – Moment Exchange(MoEx), which replaces the feature moments (mean and variance) of original image by another new image and inserts target labels of two images, forcing the model to focus on the feature moments (from new image) – the normalized features (from the original image). We incorporate this method into our fine-grained classification model to demonstrate its effectiveness on multiple datasets.

3 Approach

We introduce the proposed WSSGA-Net (illustrated in Fig. 2) in detail in this section. Firstly, the MoEx data augmentation

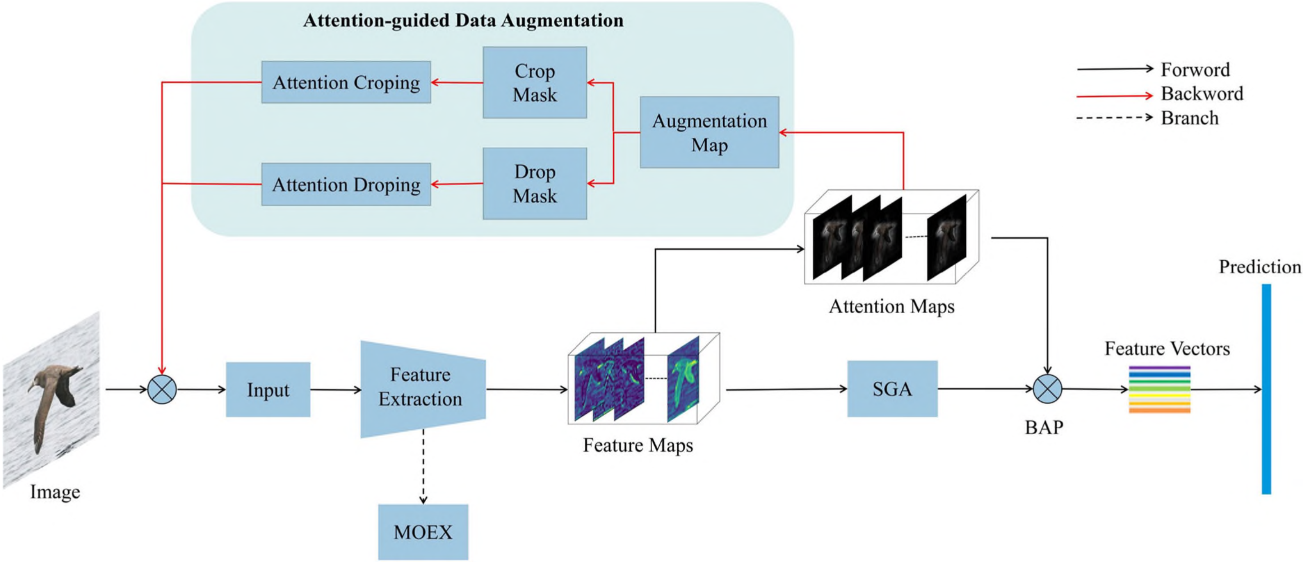


Fig. 2 Overview on the structure of the Weakly Supervised Spatial Group Attention Network

method is applied in the feature extraction stage to generate new feature maps fusing two input image features, resulting in the increased training data. Then, we deliver the obtained feature maps to the SGA module to eliminate possible noise and highlight the corresponding semantic feature regions. Meanwhile, they are transformed into part attention maps, which are used for attention-guided data augmentation to augment the input data. Further, enhanced feature groups are achieved by the SGA module based on the original feature maps, which are combined to generate new feature maps. Bilinear attention pooling (BAP) element-wise multiply feature maps after the SGA module and part attention maps to generate the feature vectors. Finally, we obtain predictions based on the feature vectors. The SGA module, BAP, MoEx, and attention-guided data augmentation are presented in the following subsections, respectively.

3.1 Attention learning

3.1.1 Spatial group attention module

The SGA module inspired by Spatial Group-wise Enhance [31] is depicted in Fig. 3. It is shown that for C channels, a $H \times W$ convolutional feature map is divided into G groups along the channel dimension. We first examine a certain group separately. In feature space, every group at every location has a representation vector, namely $\mathcal{X} = \{\mathbf{x}_{1\ldots m}\}$, $\mathbf{x}_i \in \mathbb{R}^{\frac{C}{G}}$, $m = H \times W$. This is based on assuming that this group catches concrete feature responses (such as bird eyes) gradually during the network learning. In the features, we can access high-level responses at the location of bird eye, while other locations almost have little activation and become zero vectors. But since the existing of similar patterns and the inevitable noise, it is often challenging for CNNs to access well-distributed feature responses. To solve this problem, the entire group space information is used to further improve the extracting ability of semantic features in important areas. The spatial averaging function $\mathcal{F}_{gp}(\cdot)$ averages the semantic vector of the group representation to get the global feature $\boldsymbol{g}$.

$$\boldsymbol{g} = \mathcal{F}_{gp}(\mathcal{X}) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \tag{1}$$

Then, utilizing this global feature, the equivalent importance coefficient is generated for every feature. Thereby for each position, we have:

$$c_i = \boldsymbol{g} \cdot \boldsymbol{x}_i \tag{2}$$

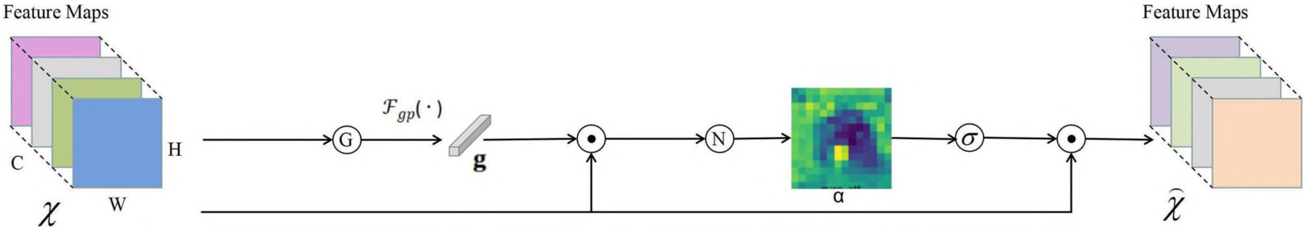


Fig. 3 The process of the SGA module

where $\boldsymbol{c}_i$ can also be expanded as $\|\boldsymbol{g}\| \|\boldsymbol{x}_i\| \cos(\theta_i)$, θ_i is the angle between $\boldsymbol{g}$ and $\boldsymbol{x}_i$. This represents that a direction (i.e., θ_i) closer to $\boldsymbol{g}$ are more possible to access a larger initial coefficient and features have a larger vector length. We normalize $\boldsymbol{c}$ over the feature space in order to eliminate the biased magnitude of coefficients between different samples, as is widely represented in:

$$\hat{c}_i = \frac{c_i - \mu_c}{\sigma_c + \epsilon}, \mu_c = \frac{1}{m}\sum_j^m c_j, \sigma_c^2 = \frac{1}{m}\sum_j^m (c_j - \mu_c)^2 \quad (3)$$

To assure that the insertion of normalization can represent the identity transformation, parameters γ, β are utilized for each coefficient $\hat{c}_i$, which scales and shifts the normalized value:

$$a_i = \gamma \hat{c}_i + \beta \quad (4)$$

In an individual SGA module, the amount of groups G is the same as the amount of γ, β, and the order of their scale is about tens (typically, 32 or 64), which is basically ignorable in comparison to the parameters of the whole network. At last, to access the augmented feature vector $\hat{x}_i$, the original $\boldsymbol{x}_i$ is enlarged by generating important coefficients $\boldsymbol{a}_i$ through a sigmoid function gate $\sigma(\cdot)$ over the feature space:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i \cdot \sigma(a_i) \quad (5)$$

and all augmented features from resulting feature group:

$$\hat{\mathcal{X}} = \{\hat{\boldsymbol{x}}_{1 \ldots m}\} \quad (6)$$

where $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{\frac{C}{G}}, m = H \times W$.

3.1.2 Bilinear attention pooling

Inspired by leveraging bilinear pooling, BAP (See Fig. 4) extracts features from the two-stream network layer. We element-wise reproduce the feature map F with every attention map A_k to access the M part feature maps F_k as presented in Eq.7:

$$F_k = A_k \odot F(k = 1, 2, \ldots, M) \quad (7)$$

where $\odot$ represents element-wise multiplication.

Next, in order to access k_{th} features f_k, a distinguishable part feature is extracted by an extra feature extraction function $g(\cdot)$, such as Global Maximum Pooling (GMP), Global Average Pooling (GAP), or convolutions,

$$f_k = g(F_k) \quad (8)$$

The feature of the object is indicated by part feature vectors $P \in R^{M \times N}$ which are overlapped by these part features f_k. Let $\Gamma(A, F)$ indicate bilinear attention pooling between feature maps F and attention maps A. It can be indicated in Eq.9,

$$P = \Gamma(A, F) = \begin{pmatrix} g(a_1 \odot F) \\ g(a_2 \odot F) \\ \ldots \\ g(a_M \odot F) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \ldots \\ f_M. \end{pmatrix} \quad (9)$$

3.2 Data augmentation

3.2.1 Moment exchange

Similar to Mixup and Cutmix, MoEx [30](see Fig. 5) fuses the normalized features of two training samples. Since the MoEx can be applied in each layer, we currently use two

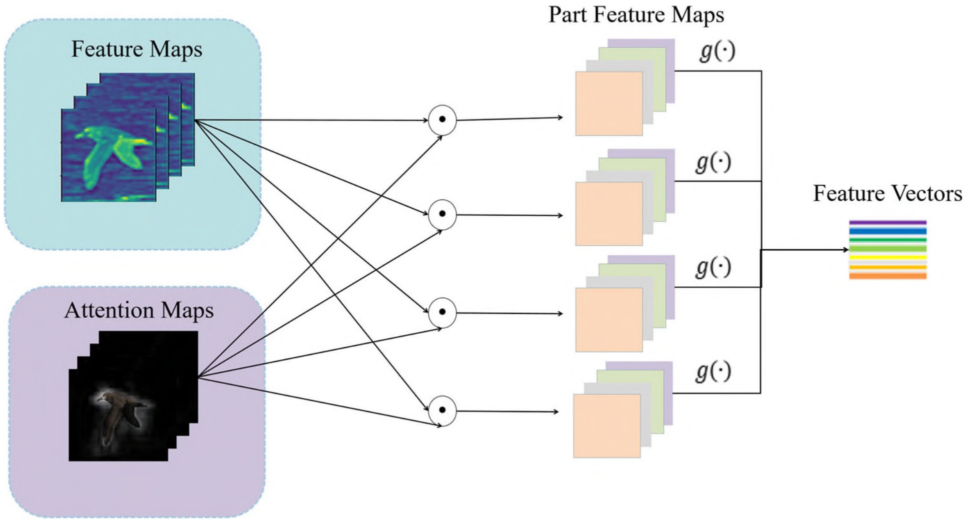


Fig. 4 The illustration of Bilinear Attention Pooling

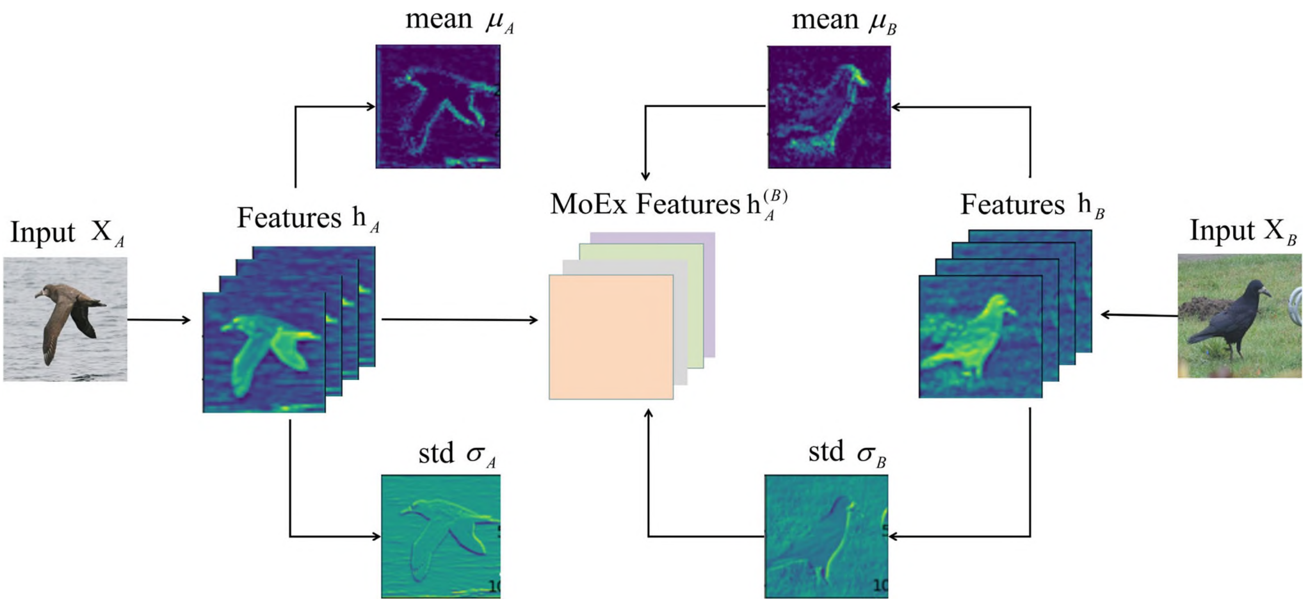


Fig. 5 The process of Moment Exchange (MoEx)

randomly selected features $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ as examples. $\hat{h}$, μ, and $\boldsymbol{\sigma}$ components are decomposed from the features of the layer inputs $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ through the within-instance normalization, respectively.

To motivate the network to take full use of the moments, we combine the feature of image $\boldsymbol{X}_A$ with the moments of image $\boldsymbol{X}_B$:

$$\boldsymbol{h}_A^{(B)} = F^{-1}\left(\hat{\boldsymbol{h}}_A, \boldsymbol{\mu}_B, \boldsymbol{\sigma}_B\right) = \sigma_B \frac{\boldsymbol{h}_A - \mu_A}{\sigma_A} + \boldsymbol{\mu}_B \tag{10}$$

Next, we continue the feature extraction process for these features $\mathbf{h}_A^{(B)}$, which contains the moments of image B, concealed in the features of image A.

3.2.2 Attention-guided Data Augmentation

We can make use of attention maps to conduct data augmentation efficiently. In training stage for every image, one of its attention maps A_k is chosen at random to conduct the process of data augmentation, and normalize it as k_{th} Augmentation Map $A_k^* \in R^{H\times W}$.

$$A_k^* = \frac{A_k - \min(A_k)}{\max(A_k) - \min(A_k)} \tag{11}$$

More detailed local features are extracted with augmentation map A_k^*. Specifically, we first obtain the Crop Mask C_k from A_k^* by setting element $A_k^*(i, j)$ which is greater than threshold $\theta_c \in [0, 1]$, as practised in Eq.12:

$$C_k(i, j) = \begin{cases} 1, & \text{if } A_k^*(i, j) > \theta_c \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Next, we search one bounding box B_k that can cover the entire chosen positive area of C_k and expand this area from the original image as the enhanced data for input. To encourage attention maps to represent multiple discriminative parts, we introduce attention dropping. Concretely, we acquire an attention Drop Mask D_k by setting element $A_k^*(i, j)$ which is greater than the threshold $\theta_d \in [0, 1]$, as presented in Eq.13:

$$D_k(i, j) = \begin{cases} 0, & \text{if } A_k^*(i, j) > \theta_d \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

The model will be prompted to extract other discriminable areas because the k_{th} part is removed from the image, which indicates the target can also be extracted better. Furthermore, this process will improve the accuracy of classification and localization.

4 Experiments

We proceed the experiments on two fine-grained bird image datasets for classification: CUB-200-2011 [32], and Bird-65. In addition, experiments on Stanford Cars dataset were conducted to demonstrate the generalizability of our approach. Our proposed WSSGA-Net approach is in comparison to

Table 1 Implementation Details

Parameter	Value
Image size	448×448
Feature extractor	ResNet-50
Batch size	32
Initial learning rate	0.001
Epochs	30
Number of attention maps	32
Optimizer	SGD

more than ten state-of-the-art approaches to validate its effectiveness and advantages.

4.1 Dataset and evaluation metric

Two datasets are applied in our experiments as follows. **CUB200-2011** is extensively-used dataset in fine-grained image recognition tasks, which includes 11 788 images of 200 bird sub-categories, among them, training set has 5994 images and testing set has 5794 images. There are an image-level sub-category label, a bounding box of the bird, and 15 part locations as annotations for every image. We just utilize an image-level sub-category label in the training stage in our experiments. **Bird-65** is a custom dataset from the bird database of Poyang Lake area, Jiangxi, China. It contains 6543 images of 65 bird sub-categories. There are 4580 images and 1963 images in the training set and testing set respectively. We only annotated every image with the image-level sub-category label. The same with CUB200-2011 dataset, we just use image-level sub-category label in the training stage. **Stanford Cars** is another widely-used dataset which contains 16185 images of 196 car subcategories. There are 8144images in the training set, and 8041images in the testing set. For each subcategory, 24 84 images are selected for training and 24 83 images for testing. There are an image-level sub-category label, a bounding box of the car for every image. We also only utilize image-level sub-category label in the training stage in following experiments.

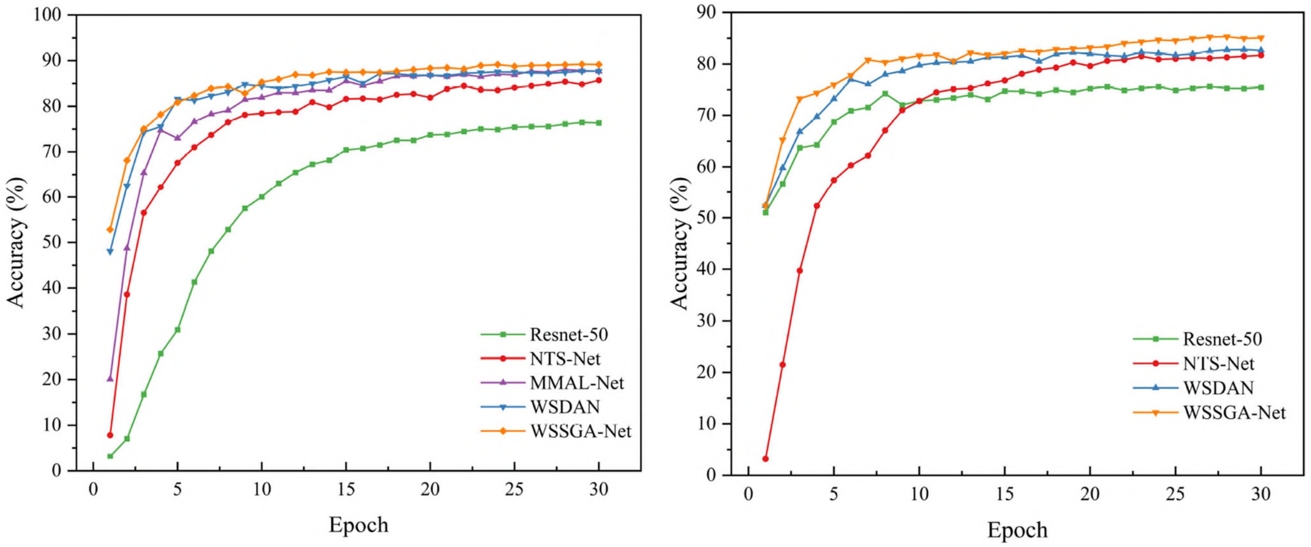
We apply accuracy as the evaluation metric to comprehensively measure the classification performances of our WSSGA-Net method and other approaches, which is extensively-used for measuring the performance of classification for fine-grained image. It represents the average classification result of our model over all classes. It is indicated as follows:

$$Accuracy = \frac{R_a}{R} \tag{14}$$

where R_{a} counts the amount of images which are classified properly and R implies the number of testing images.

Table 2 Comparison with other models on the CUB200-2011 dataset. * represents experimental results from existing method papers on the CUB200-2011 dataset

Methods	Train Annotation		Test Annotation		Accuracy(%)	CNN Features
	Object	Parts	Object	Parts		
Part-based R-CNN*[7]	✓	✓	✓	✓	76.4	AlexNet
Deep LAC*[10]	✓	✓	✓		84.1	AlexNet
SPDA-CNN*[14]	✓	✓	✓		85.1	VGGNet
Triplet-A*[33]	✓		✓		80.7	GoogleNet
Coarse-to-Fine*[34]	✓		✓		82.9	VGGNet
B-CNN*[11]					84.1	VGGNet
ST-CNN*[35]					84.1	GoogleNet
RA-CNN*[18]					85.4	VGGNet
NTS-Net[15]					85.8	ResNet-50
MA-CNN*[19]					86.5	VGGNet
DFL-CNN*[36]					87.4	VGGNet
Guided Zoom*[37]					87.7	ResNet-18
SMA-Net*[23]					87.7	ResNet-50
MMAL-Net[22]					88.0	ResNet-50
AP-CNN*[24]					88.4	ResNet-50
DF-GMM*[25]					88.8	ResNet-50
ResNet-50[2]					76.5	ResNet-50
WSDAN(baseline)[21]					87.8	ResNet-50
WSSGA-Net					**89.2**	ResNet-50



(a) CUB200-2011

(b) Bird-65

Fig. 6 Comparison with other models on the different datasets

4.2 Implementation Details

Our experimental platform is a computer with Intel(R) Xeon(R) Platinum 8255C CPU, Nvidia GeForce RTX 3090 GPU, and Ubuntu 18.04. The experiment setup is shown in Table 1, GAP is chosen as the feature pooling function $g(\cdot)$. θ_c and θ_d are the attention cropping and dropping threshold, which are both set to 0.5.

At the training stage of WSSGA-Net, the initial weights of the RestNet-50 are obtained by pre-training on the ImageNet 1K dataset. We apply a weight decay of 0.00001 with a momentum of 0.9 and set the initial learning rate with exponential decay of 0.9 after every 2 epochs.

4.3 Comparisons with State-of-the-art Approaches

We present the experimental results and analyses of our WSSGA-Net method on the above mentioned datasets as well as the state-of-the-art approaches in this subsection. Table 2 and Fig. 6(a) show the comparison performance on the CUB-200-2011 dataset. To conduct fair comparison, we list the target, part annotations, and feature extraction networks which all approaches utilize. CNN Features represent which network this approach applies to extract CNN features, such as VGGNet and GoogleNet, ResNet.

WSSGA-Net achieves the best performance among other approaches under the same weakly supervised setting that no target and part annotations are utilized in both training and testing stages, and accesses 0.4% higher accuracy than the best approach accuracy of DF-GMM (89.2% vs 88.8%). Both the low-rank reorganization representation used in the DF-GMM and the SGA module used in our WSSGA-Net approach consider the spatial context of the discriminative region, but the MoEx module and the attention-guided data augmentation applied by our WSSGA-Net enable us to obtain more than enough target features to get results ahead of DF-GMM. Our WSSGA-Net approach improves by 1.4% over our baseline method (WSDAN), verifying the effectiveness of the further improvement in our WSSGA-Net method. The main reason for the accuracy improvement is that the SGA module focuses on the spatial group features of the targets. So our WSSGA-Net method can extract the target features more comprehensively to improve the classification accuracy. The secondary reason is that MoEx performs data augmentation in feature space, complementing the attention-guided data augmentation of the baseline method (WSDAN). Altogether, our WSSGA-Net method benefits from two simultaneous data augmentation methods that can provide enough target features for the network to improve classification accuracy.

Our approach has a better performance than the approaches which are based on the CNN structures, such as ST-CNN and Bilinear-CNN. ST-CNN adopt GoogleNet with batch normalization to obtain 82.3% through only conducting fine-tuning on the CUB-200-2011 dataset. VGGNet and VGG-M are two different CNNs in Bilinear-CNN. Both two approaches are proposed earlier using basic CNNs architecture resulting in weak feature extraction and easily disturbed by background information noise. The SGA module in our WSSGA-Net approach can guide the network to learn the target features, thus achieving an accuracy improvement of 5.1%.

Moreover, even we compare with state-of-the-art approaches with supervision in both the training and testing stages, such as Triplet-A, or Coarse-to-Fine, find that our WSSGA-Net

Table 3 Comparison with other models on the Bird-65 dataset

Methods	Accuracy(%)
ResNet-50[2]	75.7
NTS-Net[15]	81.7
WSDAN[21]	82.8
WSSGA-Net	**85.3**

Table 5 Comparison with other random data augmentation

Model	Accuracy(%)	
	CUB200-2011	BIRD-65
WSDAN	87.8	82.8
+CutMix[27]	87.9	83.1
+MoEx	**88.7**	**84.0**

approach obtains an improvement by at least 8.2%. Furthermore, our approach superiors to methods that use both object and part annotations, such as Deep LAC. The experimental results show that weakly supervised methods do not definitely lose accuracy owing to the lack of object-level labels, but that network structure optimization is required to maintain accuracy. We apply neither object nor part annotations in our WSSGA-Net method, which enables fine-grained image classification tailored to actual applications.

In addition, we compared ResNet-50, NTS-Net, WSDAN, and WSSGA-Net to analyse the performance on the Bird-65 dataset. The results are presented in Table 3 and Fig. 6(b). The trend of results on this dataset is similar as on the CUB-200-2011 dataset: Under the same weakly supervised setting, our WSSGA-Net method achieves the best performances among state-of-the-art approaches, which has a 2.5% enhancement over compared approach that has the best classification results.

4.4 Effectivenesses of Components in Our WSSGA-Net Method

In the following two aspects, specific experiments are conducted to show the effectiveness of components in our WSSGA-Net method: *1. Effectiveness of spatial group attention:* In this subsection, in order to carry out a more in-depth evaluation of the effectiveness of the SGA, we present the accuracies of the baseline model (WSDAN) combined with the SGA and three other commonly used attention modules in Table 4. We can discover that the SGA in our WSSGA-Net method on the CUB200-2011 dataset can improve the classification accuracy of the WSDAN method by 1.2%. In contrast to the convolutional block attention module (CBAM) that also focuses on image spatial relationship, SGA is improved by 0.8% over the WSDAN model. Moreover, SGA performs better than the efficient channel attention (ECA) which focuses on the channel relationship. Besides, there is a simple, parameter-free attention module for convolutional neural networks (SimAM) that optimizes an energy function to find the importance of each neuron. The classification accuracy of the WSDAN model with SimAM is 88.5%, which is lower in accuracy than the WSDAN model with SGA by 0.5%. The SGA focuses on generating the attention factor in each semantic group for each spatial location, thus enabling each group to enhance its autonomously learned representations and boost the classification accuracy. Furthermore, the trends of results on the Bird-65 dataset are similar to those on the CUB-200-2011 dataset: the WSDAN method with SGA achieves the best performances among the other attention modules.

2. Effectiveness of moment exchange: MoEx is an implied data augmentation approach that promotes the network to take full use of the moment information also for recognition networks. Since our method is fast, conducts wholly in feature space, and mixes different features, one can effectively integrate it with attention-guided data augmentation methods in the WSDAN model. As shown in Table 5, the accuracy improvements of MoEx in our WSSGA-Net approach over random data augmentation (CutMix) for the WSDAN method are greater on both datasets. Because both random data augmentation and attention-guided data augmentation in WSDAN are directly applied on the input images, the overlap between these two methods results in an insignificant improvement. However, MoEx is a feature-space data augmentation that can be used in conjunction with

Table 4 Comparison with other attention modules

Model	Accuracy(%)	
	CUB200-2011	BIRD-65
WSDAN	87.8	82.8
+CBAM[38]	88.2	83.3
+ECA[39]	88.4	83.6
+SimAM[40]	88.5	84.2
+SGA	**89.0**	**84.6**

Table 6 Effectivenesses of components in our WSSGA-Net method

SGA	MoEx	Accuracy(%)	
		CUB200-2011	Bird-65
×	×	87.8	82.8
×	✓	88.7	84.0
✓	×	89.0	84.6
✓	✓	**89.2**	**85.3**

Table 7 Effects of group number G

Group number G	WSSGA-Net Acc(%)
8	88.53
16	88.91
32	89.15
64	89.23
128	88.65

attention-guided data augmentation in WSDAN to promote the network to learn features.

Besides, as demonstrated in Table 6, MoEx is compatible with the SGA. Both of the modules we added contribute considerably to the final 89.2% accuracy, which is 1.4% higher as compared to the WSDAN model. As expected, the application of MoEx in our WSSGA-Net method can also increase the classification accuracy on the Bird-65 dataset, which is 1.2% higher compared with the WSDAN model. And, if MoEx is not applied in our WSSGA-Net approach, the classification accuracy declines by 0.7%.

3. Effects of group number G: In our WSSGA-Net method, the group number G controls the amount of different semantic sub-features in the SGA module. In case the total number of channels is fixed, too few groups will limit semantic diversity; In contrast, too many groups will lead to weaker feature representation for each semantic response. It is possible that there is a suitable hyperparameter G to balance semantic diversity and the ability of representing each semantic to improve network performance. We can observe that the classification accuracy of our WSSGA-Net method on the CUB200-2011 dataset shows a trend of increasing first and then decreasing in Table 7 and Fig. 7. Based on experimental results we usually recommend the group number G to be 64.

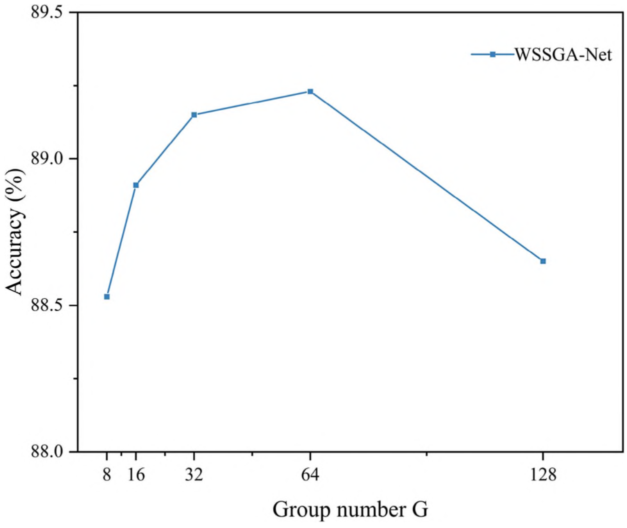


Fig. 7 Effects of group number G

Table 8 Effects of initialization parameter γ and β

γ	β	WSSGA-Net Acc(%)
0	0	89.23
0	1	89.03
1	0	88.91
1	1	89.11

4. Effects of initialization parameter γ and β: The parameter γ and β in the SGA moduel have a slight but non-negligible impact on experimental results. We assign values 0,1 to γ and β respectively to see the effect of the initialization parameters on the experimental results. During the initial stage of network training, since the ordinary patterns of semantic learning has not yet been completely formulated in convolutional feature maps, it may be suitable to abandon the attention mechanism for a moment, but let the model learn the basic semantic features first. The attention modules need to be gradually turned in effect after the initial training phase. In Table 8 we can discover that the model with both γ and β set to 0 has the highest classification accuracy. *5. Effects of normalization layer:* Since different samples in the same semantic group have inconsistent distribution of features, it is difficult to learn robust importance coefficients without normalization. As shown in Table 9, we carry out experiments by rermoving the normalization layer from SGA moduel and observe that the classification accuracy of our WSSGA-Net method significantly decreases.

4.5 Visualization analysis

We further explore whether our proposed WSSGA-Net method can concentrate on the bird in the image through a heat map visualization. We visualize the feature mapping in the final convolutional layer by Gradient-weighted Class Activation Mapping (Grad-CAM) [41]. The maximum connected area with a high response value in the heat map indicates our intended target object region, since it may show how various areas in the raw image contributed to the right categorization. For a direct contrast, we superimpose the heat maps accessed from the visualization of the ResNet-50, WSDAN, and WSSGA-Net directly on the original images. Then, we can identify the focal regions of the networks. As shown in Fig. 8, the regions of the image that the network is

Table 9 Effects of normalization layer

Normalization	WSSGA-Net Acc(%)
✓	89.23
×	88.15

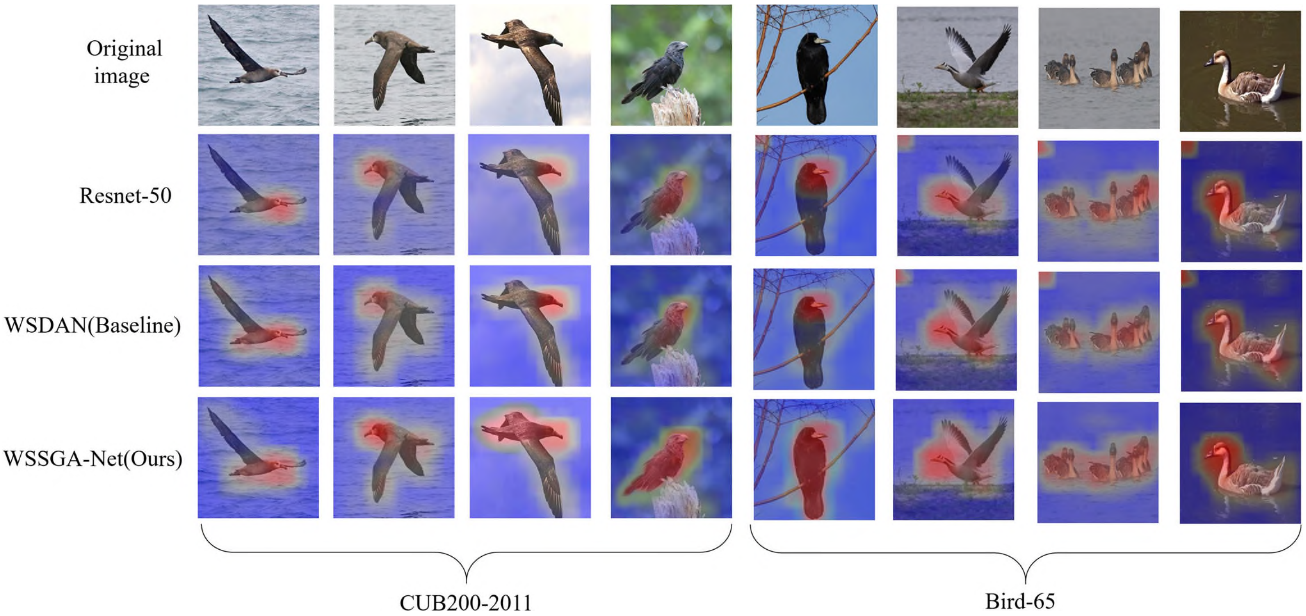


Fig. 8 Heat map visualization of different models

more concentrated on are those that are redder and brighter. The first line is the original images, the second through fourth lines are the results of heat map visualization of the ResNet-50, WSDAN, and WSSGA-Net, respectively. In the images of the ResNet-50 row and WSDAN row, the focused areas are mainly on the head of the birds, while in the image of our method row, the focused areas includes the entire bird objects. Through comparing the last three columns, it is evident that although ResNet-50 and WSDAN model focus inappropriately on some parts of the background, our WSSGA-Net model does not. In conclusion, our WSSGA-Net method allows the model to concentrate on the entire object, which is critical to locate the target and eliminate background information disturbance.

Table 10 Comparison with other models on the Stanford Cars dataset. * represents experimental results from existing method papers on the Stanford Cars dataset

Methods	Accuracy(%)	CNN Features
B-CNN*[11]	91.3	VGGNet
RA-CNN*[18]	92.5	VGGNet
NTS-Net[15]	93.8	ResNet-50
MA-CNN*[19]	92.8	VGGNet
DFL-CNN*[36]	93.8	VGGNet
Guided Zoom*[37]	93.0	ResNet-18
SMA-Net*[23]	94.4	ResNet-50
MMAL-Net[22]	95.0	ResNet-50
AP-CNN*[24]	95.4	ResNet-50
DF-GMM*[25]	94.8	ResNet-50
PMG*[42]	95.1	ResNet-50
WSDAN(baseline)[21]	94.3	ResNet-50
WSSGA-Net	**95.6**	ResNet-50

4.6 Generalization studies

To investigate the generalization capability of our WSSGA-Net method, we conduct experiments on the Stanford Cars dataset and compare with state-of-the-art approaches. Table 10 and Fig. 9 present the comparison performance on the Stanford Cars dataset. The classification accuracy of our

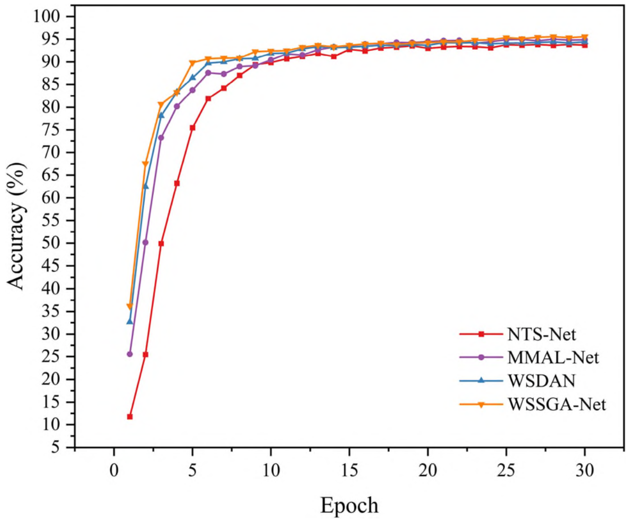


Fig. 9 Generalization experiments on the Stanford Cars dataset

WSSGA-Net method is still highest among state-of-the-art approaches and has a 1.3% enhancement over baseline WSDAN method.

In conclusion, our WSSGA-Net method not only outperforms state-of-the-art approaches for bird recognition but also has the best classification accuracy for car recognition. We demonstrate the superior performance of our method while taking into account generalization capabilities.

5 Conclusion

In this paper, we proposed a weakly supervised spatial group attention network (WSSGA-Net) method for fine-grained bird image recognition. First, we applied MoEx data augmentation in the feature space to provide more training data for the weakly supervised network. Then, the SGA facilitated the weakly supervised learning network to generate an attention factor for every spatial location to extract more discriminative image feature. Significantly, in order to fit to practical application, our WSSGA-Net method avoids the heavy labor-consumption of annotation. Extensive experimental results on the Bird datasets(i.e. Bird-65 and CUB200-2011) highlight the benefits of our WSSGA-Net method when compared to state-of-the-art methods. Furthermore, to confirm that our method is generalizable, we evaluated it on the Stanford Cars datasets, where it outperformed state-of-the-art methods. The efficiency of the SGA and MoEx modules in our WSSGA-Net method is validated by ablation studies and parameter-dependent experiments.

Our study focused on improving fine-grained visual recognition accuracy, however incorporating multiple modules increases the amount of network parameters and computing cost, making it challenging to deploy our WSSGA-Net model directly to edge devices. Therefore, in the future study, we will enable our WSSGA-Net to be more generic to satisfy the classification of multiple datasets. The other is that we will try to lightweight the WSSGA-Net in order to deploy it on edge devices for practical applications.

Acknowledgements This work was jointly supported by the Fundamental Research Funds for the Central Universities [No. 2021ZY70], the Beijing Municipal Natural Science Foundation [No. 6214040], the China Scholarship Council [No. 202106515010], and the Fundamental Research Funds for the Central Universities [No. BLX202129].

Data availability The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declarations

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

References

1. Zhao Z, Luo Z, Li J, Wang K, Shi B (2018) Applied Sciences 8(10):1906
2. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778
3. K. Simonyan, A. Zisserman, arXiv preprint http://arxiv.org/abs/1409.1556arXiv:1409.1556 (2014)
4. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2818–2826
5. Zheng H, Fu J, Zha ZJ, Luo J, Mei T (2019) IEEE Transactions on Image Processing 29:476
6. Kim T, Hong K, Byun H (2021) Neurocomputing 439:374
7. N. Zhang, J. Donahue, R. Girshick, T. Darrell, in *European Conference on Computer Vision* (Springer, 2014), pp. 834–849
8. R. Girshick, J. Donahue, T. Darrell, J. Malik, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 580–587
9. S. Branson, G. Van Horn, S. Belongie, P. Perona, arXiv preprint arXiv:1406.2952 (2014)
10. D. Lin, X. Shen, C. Lu, J. Jia, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1666–1674
11. T.Y. Lin, A. RoyChowdhury, S. Maji, in *Proceedings of the IEEE international conference on computer vision* (2015), pp. 1449–1457
12. C. Yu, X. Zhao, Q. Zheng, P. Zhang, X. You, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 574–589
13. Min S, Yao H, Xie H, Zha ZJ, Zhang Y (2020) IEEE Transactions on Image Processing 29:4996
14. H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, D. Metaxas, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1143–1152
15. Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, L. Wang, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 420–435
16. Lin Z, Gao W, Huang F, Jia J (2021) Knowledge-Based Systems 232:107480
17. Guo C, Lin Y, Chen S, Zeng Z, Shao M, Li S (2022) Knowledge-Based Systems 235:107651
18. J. Fu, H. Zheng, T. Mei, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4438–4446
19. H. Zheng, J. Fu, T. Mei, J. Luo, in *Proceedings of the IEEE international conference on computer vision* (2017), pp. 5209–5217
20. H. Zheng, J. Fu, Z.J. Zha, J. Luo, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2019), pp. 5012–5021
21. T. Hu, H. Qi, Q. Huang, Y. Lu, arXiv preprint arXiv:1901.09891 (2019)
22. F. Zhang, M. Li, G. Zhai, Y. Liu, in *International Conference on Multimedia Modeling* (Springer, 2021), pp. 136–147
23. Liu C, Huang L, Wei Z, Zhang W (2021) Applied Intelligence 51(11):7903
24. Ding Y, Ma Z, Wen S, Xie J, Chang D, Si Z, Wu M, Ling H (2021) IEEE Transactions on Image Processing 30:2826
25. Z. Wang, S. Wang, S. Yang, H. Li, J. Li, Z. Li, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2020), pp. 9749–9758
26. C. Gong, D. Wang, M. Li, V. Chandra, Q. Liu, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 1055–1064

27. S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, in *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 6023–6032
28. J. Yoo, N. Ahn, K.A. Sohn, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8375–8384
29. E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q.V. Le, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 113–123
30. B. Li, F. Wu, S.N. Lim, S. Belongie, K.Q. Weinberger, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2021), pp. 12,383–12,392
31. X. Li, X. Hu, J. Yang, arXiv preprint arXiv:1905.09646 (2019)
32. C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, california institute of technology (2011)
33. Y. Cui, F. Zhou, Y. Lin, S. Belongie, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1153–1162
34. Yao H, Zhang S, Zhang Y, Li J, Tian Q (2016) IEEE Transactions on Image Processing 25(10):4858
35. M. Jaderberg, K. Simonyan, A. Zisserman, et al., Advances in neural information processing systems **28** (2015)
36. Y. Wang, V.I. Morariu, L.S. Davis, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 4148–4157
37. Bargal SA, Zunino A, Petsiuk V, Zhang J, Saenko K, Murino V, Sclaroff S (2021) IEEE Transactions on Pattern Analysis and Machine Intelligence 43(11):4196
38. S. Woo, J. Park, J.Y. Lee, I.S. Kweon, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–19
39. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, in *Proceedings of the 2020 IEEE conference on computer vision and pattern recognition, IEEE, Seattle, WA, USA* (2020), pp. 13–19
40. L. Yang, R.Y. Zhang, L. Li, X. Xie, in *International conference on machine learning* (2021), pp. 11,863–11,874
41. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in *Proceedings of the IEEE international conference on computer vision* (2017), pp. 618–626
42. R. Du, D. Chang, A.K. Bhunia, J. Xie, Z. Ma, Y.Z. Song, J. Guo, in *European Conference on Computer Vision* (Springer, 2020), pp. 153–168

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jiangjian Xie received the B.S. degree from China Agricultural University, in 2007, and the Ph.D. degree from Beijing Jiaotong University, in 2013. He is currently an Associate Professor with Beijing Forestry University. His research interest includes intelligent progressing of forestry ecological environment information.

Yujie Zhong received the B.S. degree in School of Technology from Beijing Forestry University, China, in 2021. He is currently a master in School of Technology, Beijing Forestry University. His research interests include deep learning, image recognition.

Junguo Zhang received his B.S. and M.S. degrees in China University of Mining and Technology, in 2000 and 2003, respectively, and the D.E. degree in Beijing Forestry University. He visited the Forest Product Laboratory, USDA in 2012. He is the Dean of Research in School of Technology, Beijing Forestry University. He is committed in the research on the forestry information collection and intelligent processing. In addition, he has led nearly ten scientific projects supported by the National Natural Science Foundation of China, State Forestry Administration, etc.

Changchun Zhang received the M.S. degree in Control Theory and Control Engineering, from Beijing Technology and Business University in 2017. He received the Ph.D. degree in Computer Science and Technology from the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, in 2021. He is currently a lecturer in the School of Technology, Beijing Forestry University, China. His research interests include several topics in computer vision and machine learning.

Björn W. Schuller received his diploma, doctoral degree, habilitation, and Adjunct Teaching Professor in Machine Intelligence and Signal Processing all in EE/IT from TUM in Munich/Germany. He is Full Professor of Artificial Intelligence and the Head of GLAM at Imperial College London/UK, Full Professor and Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, co-founding

CEO and current CSO of audEERING, independent research leader within the Alan Turing Institute as part of the UK Health Security Agency. He is a Fellow of the IEEE and Golden Core Awardee of the IEEE Computer Society, Fellow of the BCS, Fellow of the ELLIS, Fellow of the ISCA, Fellow and President-Emeritus of the AAAC, Elected Full Member Sigma Xi, and Senior Member of the ACM.