PAPER

A Two-Stage Approach for Fine-Grained Visual Recognition via Confidence Ranking and Fusion

Kangbo SUN[†a)], ***Student Member and*** **Jie ZHU**[†b)], ***Nonmember***

SUMMARY Location and feature representation of object's parts play key roles in fine-grained visual recognition. To promote the final recognition accuracy without any bounding boxes/part annotations, many studies adopt object location networks to propose bounding boxes/part annotations with only category labels, and then crop the images into partial images to help the classification network make the final decision. In our work, to propose more informative partial images and effectively extract discriminative features from the original and partial images, we propose a two-stage approach that can fuse the original features and partial features by evaluating and ranking the information of partial images. Experimental results show that our proposed approach achieves excellent performance on two benchmark datasets, which demonstrates its effectiveness.

key words: *fine-grained, object location, attention, bilinear pooling, deep learning*

1. Introduction

Fine-grained visual recognition task is a great challenge for most deep learning networks. This task is aimed at distinguishing subtle differences among various classes of specific things, such as different birds and different aircraft. Size, shape, and texture of images play important roles in the image recognition task. However, the shape and size are similar among different categories in fine-grained image recognition, which leads to high difficulty in recognition. Therefore, it is important to distinguish partial textures in fine-grained image recognition. The bounding box as an auxiliary label can provide effective foreground information, which can greatly reduce the difficulty of fine-grained image classification. However, it is expensive to annotate bounding boxes artificially.

To benefit from bounding boxes without using them, many methods propose to utilize neural networks to generate bounding boxes in a weak supervision manner (with category label only). These methods obey the same manner: locating object with location networks under weak supervision and cropping out the regions inside bounding boxes as auxiliary data to train classification neural networks. Therefore, these methods can be defined as the two-stage methods with location process and classification process.

In the location process, Ge et al. [1] utilize the trained network from the object location task. The trained location network can not be optimized by fine-grained image recognition task, which results in the method not being an end-to-end method. Some methods [2], [3] propose to locate and classify the object within one end-to-end network with attention mechanism or Region Proposal Network (RPN) [4]. In the classification process, Hu et al. [2] utilize the auxiliary partial images to enhance the ability of the classification network by a data augmentation way. Zheng et al. [5] propose to utilize knowledge distillation to force the main branch to learn the information form the part branch as an alternative feature fusion option. Yang et al. [3] concatenate original feature and partial features together to get final classification. The two-stage methods similar to [1]–[3], [5] greatly improve the classification accuracy of fine-grained visual recognition tasks on related datasets.

In general, different partial images have different information for classification, and the more information, the more contribution to the final classification. Therefore, how to propose partial images with high information and how to fuse the original and partial features are essential for the two-stage fine-grained image recognition. In this paper, we propose the confidence ranking and fusion method to propose partial images and fuse original feature and partial features extracted by our two-stage network, which is shown in Fig. 1. Our proposed method adopts an attention-based location method to generate candidate partial images. To propose partial images with high information and extract dis-

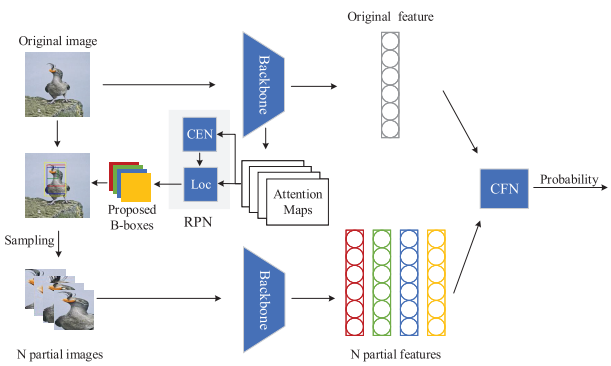

Fig. 1 The overview of our proposed approach. Backbone: the baseline based on the InceptionV3 and bilinear pooling, and the two backbones share same parameters. RPN: Region Proposal Network that proposes the top N informative partial regions, which includes Confidence Evaluation Network (CEN) and the location method. CFN: Confidence Fusion Network that fuses the original and the N partial features to make final decisions. N is determined as 4 in this figure.

Manuscript received February 4, 2020.
Manuscript revised July 8, 2020.
Manuscript publicized September 11, 2020.
[†]The authors are with Shanghai Jiao Tong University, China.
a) E-mail: kangbosun@sjtu.edu.cn
b) E-mail: zhujie@sjtu.edu.cn (Corresponding author)
DOI: 10.1587/transinf.2020EDP7024

criminative features from proposed partial images, we evaluate the information of the candidate partial images and select the most informative partial images, and fuse the original and partial features to make final decisions. The proposed approach could effectively extract discriminative information from these original and partial images.

To verify our approach, we conduct extensive experiments on related fine-grained image datasets. Experimental results show that our proposed model outperforms the baseline model with a large margin of 3.2% accuracy and gets a competitive performance on the CUB-200-2011 [6] dataset with 89.7% accuracy, and achieves state-of-the-art performance on the FGVC-Aircraft [7] dataset with 94.3% accuracy.

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 introduces our approach in detail. Section 4 provides experiments and results. Finally, conclusions are drawn in Sect. 5.

2. Related Works

Fine-grained image recognition. Deep learning and convolutional neural networks have achieved great success in image recognition and have replaced traditional manual feature extraction methods like SIFT [8] and HOG [9]. In fine-grained image recognition, classic CNN networks such as VGG [10], ResNet [11], and Inception [12] have also achieved competitive performance, but the recognition capabilities of these networks are still far from human beings. To address this problem, researches on fine-grained image recognition are mainly divided into two directions. One direction is to design a better network to extract more discriminative features, and the other direction is to provide the network with more discriminative images by using manually labeled bounding boxes to reduce the difficulty of recognition.

To design a better CNN extractor, Lin et al. [13] first adopt bilinear pooling in fine-grained visual recognition and achieve remarkable improvements. Many works about bilinear pooling focus on simplifying bilinear operations to reduce computational costs like [14], [15] or designing more complex bilinear networks to improve performance like [16]. In our work, we use the bilinear pooling to fuse the attention features and the CNN features to generate more discriminative features.

Bounding boxes can promote the performance of CNNs by a large margin, while it is expensive to get the manually labeled data. Many works [1]–[3] have been done to produce bounding boxes by neural networks in weak supervision manner. Ge et al. [1] utilize Mask-RCNN [17] network pre-trained by object location task to generate bounding boxes and partial images, which provide the classification network with more discriminative features. Hu et al. [2] propose to utilize attention maps to locate the object, and re-train the network in a data augmentation way. Zheng et al. [5] propose to zoom in the attention regions and utilize knowledge distillation to force the main branch to learn the information form the part branch as an alternative feature fusion option. Yang et al. [3] evaluate the region of preset anchors to locate object, and select the most possible partial images to train the network with original image together. In our work, we utilize the attention maps to generate bounding boxes as auxiliary data in one end-to-end network. Our proposed method does not need to preset anchors to locate the object parts and could extract features from the parts more effectively.

Object Location in weak supervision manner. Object location networks such as Mask-RCNN [17], SSD [18] and YOLO [19] have achieved impressive performance in object location. However, the large number of bounding boxes/part annotations burden networks with limited budgets. Weakly supervised methods aim to train networks with only image-level labels, which can alleviate this problem. Jie et al. [20] propose a self-taught learning network by selecting some high-response proposals and fine-tuning the network with these proposals. Diba et al. [21] propose to produce region proposals by using the attention maps. In our work, we apply the 1*1 convolution layer to generate attention maps and locate the high-response regions as the part annotations.

3. Approach

In this section, we introduce our method in detail. We first introduce the entire structure of our proposed model, and then explain in detail its three major components, including the baseline CNN model, our Region Proposal Network (RPN), and our Confidence Fusion Network (CFN).

3.1 Approach Overview

Our proposed model has three main components which are shown in Fig. 1. Our backbone network is built based on InceptionV3. Inspired by the work by [2], we extract the feature out of layer *Mix6e* and utilize the 1*1 convolution layer to generate attention maps. We adopt the bilinear pooling to fuse the attention features and the CNN features to generate more discriminative features. To generate partial images with high information, we design the RPN which includes Confidence Evaluation Network (CEN) and location method that work in pairs. The CEN evaluates the confidence score of each channel in the attention maps and sorts those channels by the score. Then, the location method utilizes Non-Maximum Suppression (NMS) method to select the top N most informative attention maps for cropping partial images. To utilize the relationship of confidence between features, we designed a special classification network named Confidence Fusion Network (CFN) to fuse the original feature and partial features. The details of our backbone model, RPN and CFN will be shown in Sect. 3.2, Sect. 3.3 and Sect. 3.4, respectively.

3.2 Bilinear Pooling in the Baseline Model

In our baseline model shown in the top half of Fig. 2, we uti-

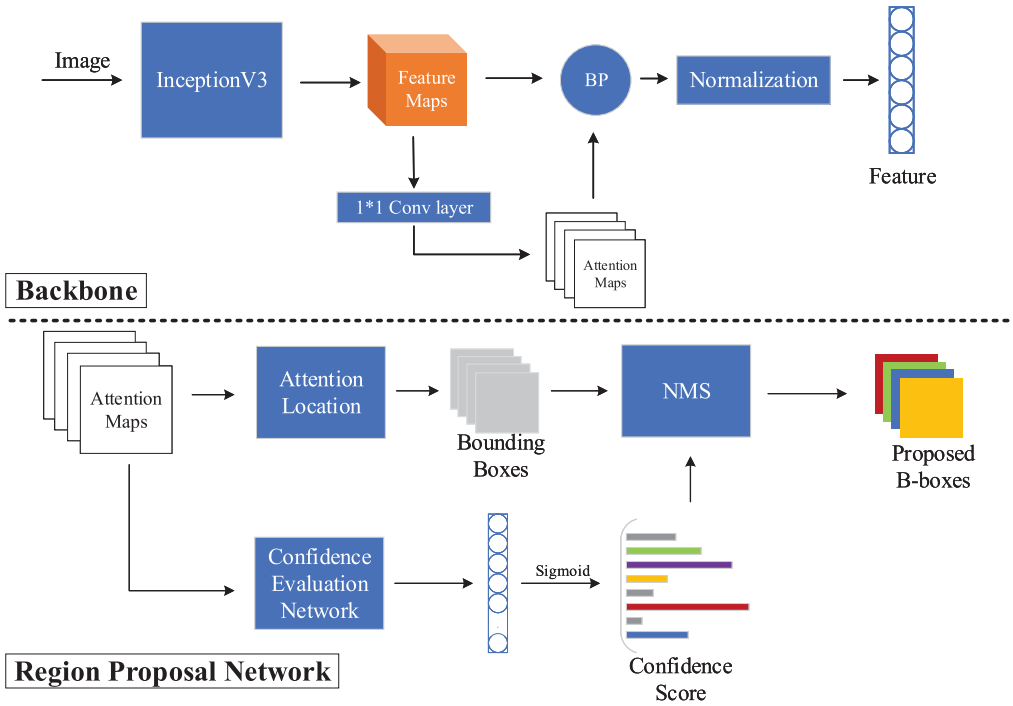

Fig. 2 Our backbone model and Region Proposal Network (RPN). InceptionV3: the CNN feature extractor, which generates feature maps out from *Mix6e* layer in InceptionV3. BP: bilinear pooling method detailed in Sect. 3.2. 1*1 Conv layer: to generate the attention maps from the feature produced by the InceptionV3 model. Normalization: to normalize the bilinear feature with sqrt and L2 normalization. Attention Location: our location method that can generate bounding boxes from the attention maps. Confidence Evaluation Network (CEN): to evaluate the confidence score for selecting the top N informative regions. NMS: Non-Maximum Suppression that can select the bounding boxes with less overlap.

lize bilinear pooling method to further enhance the feature by the InceptionV3 network. It is assumed that $F \in \mathcal{R}^{S\times K_1}$ is the feature maps obtained from the layer *Mix6e* of the InceptionV3 network, and the attention maps $A \in \mathcal{R}^{S\times K_2}$ is the attention maps, where K_1 and K_2 mean the number of channels, and $S = H * W$ is the spatial size of feature maps. Bilinear pooling is defined as:

$$B = A^T F \tag{1}$$

where $B \in R^{K_2\times K_1}$ is the output of bilinear pooling layer. For each element $B_{i,j}$ in B, there are:

$$B_{i,j} = {A_i}^T F_j = (a_i)^T x_j = \sum_{k=1}^{S} a_{k,i} x_{k,j} \tag{2}$$

where $B_{i,j}$ is the product of spatial location of the original feature in channel i of A and channel j of F, $a_i = (a_{1,i}, a_{2,i}, \ldots, a_{S,i})^T$ and $x_j = (x_{1,j}, x_{2,j}, \ldots, x_{S,j})^T$ mean the feature on channel i and j of A and F respectively, and $x_{j,i}$ means the feature on the position j of the channel i.

3.3 Region Proposal Network

To propose partial regions from the original images and the attention maps, we build Region Proposal Network (RPN) shown in the bottom of Fig. 2. The RPN first locates the partial regions with each channel of attention maps by the method introduced in Hu et al. [2]. Then, the appropriate partial regions will be chosen to crop partial images according to certain rules. This is because not all partial images can benefit the final classification and it is impossible to train all partial images generated by the attention maps. We assume that the more information a partial image has, the greater its contribution to the final classification. And, the contribution is called confidence score in this paper. Therefore, to achieve similar results as training all partial images by using only part of the partial images, we select the top N most informative partial images from all the images. To determine the top N informative partial images, we build the Confidence Evaluation Network (CEN) and the network takes the attention maps as input and estimates confidence score for each channel of the attention maps. To train CEN, the confidence labels for each channel are necessary for supervision. In this paper, we can produce N confidence labels by determining the classification cross-entropy of the top N confidence partial images. However, the number (M) of labels may not match the number (N, $N \le M$) of partial images used. To address this issue, we utilize the learning to rank method introduced by Yang et al. [3] to train the CEN.

Assuming that the confidence score evaluated by the CEN is given as $C = (C_1, C_2, \ldots, C_N, \ldots, C_M)$, $C_1 > C_2 >$

$\cdots > C_N > \cdots > C_M$, the top N informative partial images generated by the region proposal network are given as $P = (P_1, P_2, \ldots, P_N)$, and the corresponding probability evaluated by the classification network is given as $\mathcal{P}_p = (\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_N)$. However, it is not sufficient to supervise C (length M) with probability $\mathcal{P}_p$ (length N). To address this issue, we utilize the pairwise rank loss to train the CEN, which is shown in Eq. (3) and will be detailed in Sect. 3.5.

$$\mathcal{L}_{rank}(f, C, \mathcal{P}_p) = \sum_{(i,j),\mathcal{P}_i<\mathcal{P}_j} f(C_i, C_j) \tag{3}$$

where f is the hinge function that can penalize cases where the ordering of sequences C and $\mathcal{P}_p$ is inconsistent.

3.4 Confidence Fusion Network

In this section, we introduce the confidence fusion network in our model. It is given that the original feature produced by bilinear pooling is B_o and the corresponding N partial features are $B_{p1}, B_{p2}, \ldots, B_{pN}$. We first rank the partial features in confidence score order, and form these features as a sequence $(B_o, B_{p1}, B_{p2}, \ldots, B_{pN})$. We believe that these partial images sorted in confidence order are context-sensitive, so we use bidirectional LSTM for joint classification. To speed up the training process, we first use the Batch Normalization (BN) layer to normalize the features before the bidirectional LSTM. The proposed confidence fusion network is shown in Fig. 3.

3.5 Loss Function and Optimization

In this subsection, we introduce the loss function in detail. To train our model, we use the center loss to constrain the attention maps, the ranking loss to propose partial images and the classification loss to classify images.

Center loss. The center loss is proposed by Wen et al. [22] to solve face recognition. To ensure the RPN could propose different partial regions, different channels in the attention maps should have different high response parts. To guarantee that, we adopt center loss in our work to constrain the feature produced by bilinear pooling.

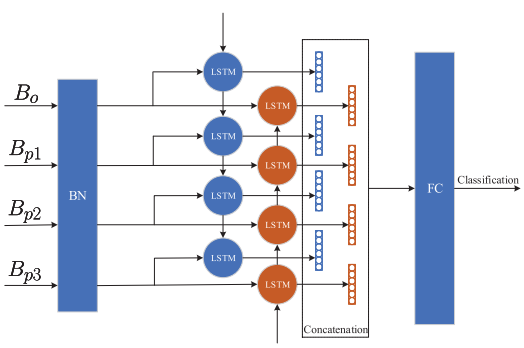

Fig. 3 Confidence Fusion Network (CFN) consists of a BN layer, a bidirectional LSTM and an FC layer, which makes final decisions based on the sequence features. N is determined as 3 in this figure.

It is assumed that $B \in \mathcal{R}^{M \times K}$ is the feature produced by bilinear pooling, and M is the number of channels in attention maps, K is the number of channels in CNN feature. For $B_i \in \mathcal{R}^{1 \times K}$, it is assumed that the center of B_i is c_i. For each category, the center loss is determined as Eq. (4).

$$\mathcal{L}_{center} = \sum_{i=1}^{M} \|B_i - c_i\|_2^2 \tag{4}$$

$$c_i \leftarrow c_i + \lambda(B_i - c_i) \tag{5}$$

where c_i is initialized as zeros, and updated by moving average determined in Eq. (5) for each category.

Ranking loss. The confidence score is given as $C = (C_1, C_2, \ldots, C_N, \ldots, C_M)$ in the sorted order, and the partial features are given as $P = (P_1, P_2, \ldots, P_N, \ldots, P_M)$. To further ensure that the top N informative partial images can be better proposed, we utilize the top R ($R \geq N$) informative partial features to train the ranking loss. Therefore, assuming that the selected partial features are $P = (P_1, P_2, \ldots, P_N, \ldots, P_R)$, and their confidence score and probability to the real label are as $C = (C_1, C_2, \ldots, C_N, \ldots, C_R)$ and $\mathcal{P}_p = (\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_N, \ldots, \mathcal{P}_R)$. To train the confidence evaluation network, we determine our ranking loss as Eq. (6).

$$\mathcal{L}_{rank}(C, \mathcal{P}_p) = \sum_{(i,j),\mathcal{P}_i<\mathcal{P}_j} \max(0, C_i - C_j) \tag{6}$$

The ranking loss function can encourage that C and $\mathcal{P}_p$ are in the same order.

Classification loss. We use the category cross-entropy function to train networks to make the final decisions. The decision loss is determined as follows:

$$\mathcal{L}_{cls1} = -\log(\mathcal{F}(X, P_1, P_2, \ldots, P_N)) \tag{7}$$

where $\mathcal{F}$ is our approach which outputs the final classification decision.

To teach the backbone model to extract features from the original and partial images, we define the loss function as follows:

$$\mathcal{L}_{cls2} = -\log(C_o(\mathcal{B}(X))) \tag{8}$$

$$\mathcal{L}_{cls3} = -\frac{1}{N}\sum_{i=1}^{N} \log(C_p(\mathcal{B}(P_i))) \tag{9}$$

where $\mathcal{B}$ means the backbone model which extract bilinear features from the original and partial images, and C_o, C_p mean two different single-layer FCs used to classify original and partial images, respectively.

The final classification loss is determined as follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{cls1} + \mathcal{L}_{cls2} + \mathcal{L}_{cls3} \tag{10}$$

Training loss and algorithm. The final loss in our experiments is determined as Eq. (11):

$$\mathcal{L}_{final} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{rank} + \beta\mathcal{L}_{center} \tag{11}$$

Algorithm 1 The training algorithm

Input: original images X, training epochs E, and $M, N, R, \lambda, \alpha, \beta$
Output: predict probability $\mathcal{P}$
while the trained epochs $< E$ **do**
1. Get the original image X and the category label Y
2. Generate feature maps F_o, attention maps A_o, bilinear features B_o, and predict probability $\mathcal{P}_o$ of original image
3. Locate the top R informative partial images $P = P_1, P_2, \ldots, P_N, \ldots, P_R$
4. Generate bilinear features B_p, and predict probability $\mathcal{P}_p$ of partial images P
5. Fuse B_o and $B_p[1:N]$ with CFN, and generate final predict probability $\mathcal{P}$
6. Calculate loss function $\mathcal{L}_{cls}$, $\mathcal{L}_{rank}$ and $\mathcal{L}_{center}$
7. Calculate gradient with BP, and update the network with SGD

end while

Table 1 Details of datasets.

Dataset	Categories	Training samples	Testing samples
CUB-200-2011	200	5994	5794
FGVC-Aircraft	100	6667	3333

where $\alpha = 1$ and $\beta = 3$ in our experiments. The final training algorithm is shown in Alg. 1.

4. Experiments

In this section, we introduce our experiments in detail, including datasets used, implement details, the performance of our approach. Finally, we visualize the proposed part locations of the object by RPN.

4.1 Datasets

We evaluate our proposed model on two benchmark datasets for fine-grained visual recognition: the CUB-200-2011 [6], FGVC-Aircraft [7]. Table 1 shows the details of these datasets, including the number of categories, the number of samples in the standard training/testing splits. In our experiments, we train our models in a weak supervision manner, which means that no bounding box or part annotation is used.

4.2 Implement Details

The input images are resized to 512×512 and randomly cropped into 448×448. Besides, we also use random rotation and random horizontal flip for data augmentation. The partial images proposed by RPN are resized to 224×224. We train all models using Stochastic Gradient Descent (SGD) optimizer with momentum of 0.9, weight decay of 1e-5, and the batch size is set to 4 on one GTX 1080Ti GPU. The initial learning rate is set to 0.001, with exponential decay of 0.9 after every 2 epochs.

Backbone network. We extract the out feature of layer *Mix6e* from InceptionV3 model and utilize the 1*1 convolution layer to generate attention maps, and finally use the bilinear pooling to generate bilinear features. The number of attention maps is determined as 32. The InceptionV3 network is pre-trained on ImageNet dataset. We use the signed square root and L2 normalization after bilinear pooling, which is widely applied in [13], [14], [16].

Region proposal network. We use non-maximum suppression (NMS) to select the top N most informative partial regions. To further ensure that the top N informative partial images can be better proposed, we utilize the top R ($R \geq N$) informative partial features to train the ranking loss. We first calculate all the bounding boxes and rank them with confidence score. We select those boxes one by one, and remove the boxes whose IoU with the selected boxes is greater than 0.5, and finally get bounding boxes with high confidence score and less overlapping regions.

Without special instructions, N is set to 4, R is set to 4, M is set to 32, and λ, α and β are set to 0.05, 1 and 3, respectively.

Table 2 Ablation experiments on NMS.

Method	Accuracy (%)
No NMS	87.6
NMS	89.4

Table 3 Ablation experiments on numbers of attention maps M.

Method	Acc (%)
$M = 16$	87.7
$M = 32$	89.4
$M = 64$	89.4

Table 4 Ablation experiments on numbers of partial images.

(N, R)	Acc (%)	(N, R)	Acc (%)
$(0, 4)$	84.7	$(1, 4)$	86.4
$(2, 4)$	88.1	$(3, 4)$	88.9
$(4, 4)$	89.3	$(4, 6)$	89.4
$(6, 6)$	89.4	$(6, 8)$	89.4

4.3 Evaluation and Analysis on the CUB-200-2011

To further understand our proposed model, we conduct extensive experiments on the CUB-200-2011 dataset, including ablation experiments on the sub-modules and the final performance of the proposed approach.

Ablation experiments on RPN. In our region proposal network, we train attention maps with center loss, and then we select the N most informative parts with NMS from M channels of attention maps. Table 2 shows that NMS method in the selecting process improves the accuracy with 1.8%, which means that NMS effectively enhances the richness of information. Table 3 shows that $M = 32$ is an appropriate size and continuing to increase M over 32 will not bring a significant performance gain, which is consistent with the work of Hu et al. [2].

Table 4 shows the effect of N and R on performance, and $N = 0$ means that the network does not utilize the partial images to make decisions, and the corresponding accuracy

of 84.7% is the classification accuracy of original images. Under the condition of $R = 4$, we study the influence of each partial image of the top N informative partial images respectively. The condition $R = 4$ is to ensure that CEN can better fit the true confidence order. The exponential results show that the top N ($N = 4$) informative partial images with confidence order improve the classification accuracy with 1.7%, 1.7%, 0.8% and 0.4%, respectively. In addition, the performance with $(N, R) = (4, 6)$ outperforms the performance with $(N, R) = (4, 4)$ with 0.1%, and continuing to increase N and R will not bring a significant performance gain.

Table 5 shows the impact of center loss on the final recognition accuracy. Due to the constraints of center loss, the attention maps can better focus on the parts of object, and the final classification accuracy is improved by 0.7%. However, excessive use of the center loss training network will cause abnormal location of the parts. Figure 4 shows the degradation phenomenon in learning the partial regions proposed by RPN on the CUB-200-2011 training dataset. As the epoch of training process increases, the average accuracy of partial images deteriorates, which means that attention maps focus on inappropriate regions. The degradation phenomenon is detrimental to the decision-making of CFN in the later epochs. To overcome the degradation phenomenon, we utilize an exponential decay of 0.1 after every 20 epochs for β in our work and set the initial β to 3. With the decay strategy, our proposed model achieves a competitive accuracy of 89.7% on the CUB-200-2011 testing dataset.

Ablation experiments on CFN. In our classification process, we first rank the partial features by confidence score, and then utilize the CFN to make final decisions. To demonstrate the effectiveness of our proposed CFN, we use the fully connected layers and LSTM to replace the bidirectional LSTM in CFN, respectively. The fully connected layers consist of two layers of FC with a hidden size of 2048, and a dropout layer [23] is included between the two layers of FC. The hidden state size of LSTM is 1024, which is twice that of Bi-LSTM. Table 6 shows that the bidirectional LSTM can better fuse the original feature and partial features, and the FCs even reduce accuracy compared with the baseline model. However, the FCs still outperforms the $(N, R) = (0, 4)$ conditions with a 1.0% accuracy. It is worth noting that if the partial features in confidence order are shuffled, the accuracy will be reduced by 0.9%.

Table 7 shows the comparison between our proposed method and other state-of-the-art methods on the CUB-200-2011 testing dataset. The InceptionV3 with BP is a strong baseline, which achieves 86.5% accuracy. While our proposed approach outperforms it with a large margin of 3.2% accuracy. Compared to NTS-net [3] which uses the preset anchors to produce partial images, we achieve a 2.2% improvement. Our proposed approach achieves a competitive accuracy compared with state-of-the-art methods on the CUB-200-2011 dataset.

Table 5 Ablation experiments on center loss.

Method	Accuracy (%)
$\beta = 0$	88.7
$\beta = 3$	89.4
β decay	89.7

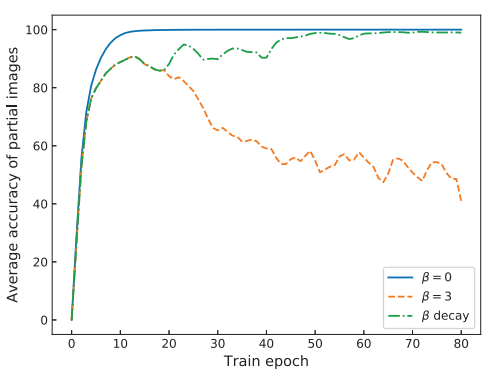

Fig. 4 The degradation phenomenon in learning the partial regions proposed by RPN on the CUB-200-2011 training dataset.

4.4 Evaluation on the FGVC-Aircraft Dataset

Table 8 shows the comparison between our proposed method and other state-of-the-art methods on FGVC-Aircraft testing dataset. Compared to NTS-net [3] which uses the preset anchors to produce partial images, we achieve a 2.9% improvement. Compared to WSDAN [2] which also uses InceptionV3 with BP as CNN extractor, we achieve a 1.3% improvement. Our proposed approach

Table 6 Ablation experiments on fusion methods.

Method	Accuracy (%)
Our baseline	86.5
BN + FCs	85.7
BN + LSTM	87.9
BN + Bi-LSTM	**89.4**
BN + Bi-LSTM (shuffle order)	88.5

Table 7 Comparison with state-of-the-art methods on the CUB-200-2011 testing dataset.

Network	Backbone	Accuracy (%)
BCNN [13]	VGGNets	84.1
STN [24]	InceptionV3	84.1
LRBP [15]	VGG-16	84.2
RA-CNN [25]	VGG-19	85.3
MA-CNN [26]	VGG-19	86.5
HBP [16]	VGG-16	87.1
DFL-CNN [27]	ResNet-50	87.4
NTS-Net [3]	ResNet-50	87.5
TASN [5]	ResNet-50	87.9
WSDAN [2]	InceptionV3	89.4
Ge et al. [1]	InceptionV3	**90.4**
Our baseline	InceptionV3	86.5
Ours ($\beta = 3$)	InceptionV3	89.4
Ours (β decay)	InceptionV3	**89.7**

Fig. 5 The visualization of the proposed partial regions by our model. The first column shows the original images from the CUB-200-2011 testing dataset. The second column shows the original images covered by the averaged attention map. The third column shows the heat maps. The 4-7th columns show the top four informative attention maps and the corresponding partial regions. The 8th column shows the all location of partial images in the original images.

Table 8 Comparison with state-of-the-art methods on FGVC-Aircraft testing dataset.

Network	Backbone	Accuracy (%)
BCNN [13]	VGGNets	86.6
RA-CNN [25]	VGG-19	88.4
MA-CNN [26]	VGG-19	89.9
HBP [16]	VGG-16	90.3
NTS-Net [3]	ResNet-50	91.4
DFL-CNN [27]	ResNet-50	92.0
WSDAN [2]	InceptionV3	93.0
Ours ($\beta = 3$)	InceptionV3	94.0
Ours (β decay)	InceptionV3	**94.3**

achieves state-of-the-art performance with 94.3% compared with previous methods on FGVC-Aircraft dataset.

4.5 Visualization

To further understand the partial regions proposed by our method, we visualize the average attention maps, the heat maps, and the proposed partial regions in Fig. 5. Attention maps and heat maps in the 2–3rd columns prove that attention maps could reveal the location of object. We use red, green, blue, and yellow to mark the location of the top four informative partial regions in the 4–7th columns. It is shown that the most informative region may be the head of birds in the CUB-200-2011 dataset in Fig. 5, and our approach could locate the most discriminative region and other high-information regions.

5. Conclusion

In this paper, we propose a two-stage approach for fine-grained visual recognition. Our proposed approach could locate and generate the most discriminative partial region and other high-information partial regions and fuse the original feature and partial features. We conduct our approach on extensive datasets, and the approach significantly improves the final accuracy with a large margin and achieves state-of-the-art performance accuracy on the FGVC-Aircraft dataset and a competitive accuracy on the CUB-200-2011 dataset.

Acknowledgments

This work is supported by the National Key Research Project of China under Grant No. 2017YFF0210903 and the National Natural Science Foundation of China under Grant Nos. 61371147 and 11433002.

References

[1] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3034–3043, 2019.
[2] T. Hu and H. Qi, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," arXiv preprint arXiv:1901.09891, 2019.
[3] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," Proc. European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol.11218, pp.438–454, 2018.
[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems 28, ed. C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, Curran Associates, pp.91–99, 2015.
[5] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.5007–5016, 2019.
[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Computation & Neural Systems Technical Report, CNS-TR-2011-001, California Institute of Technology, 2011.
[7] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," arXiv preprint arXiv:1306.5151, 2013.
[8] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol.60, no.2, pp.91–110, 2004.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05), pp.886–893, 2005.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.770–778, 2016.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1–9, 2015.
[13] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," Proc. IEEE International Conference on Computer Vision, pp.1449–1457, 2015.
[14] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.317–326, June 2016.
[15] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7025–7034, 2017.
[16] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," Proc. European Conference on Computer Vision (ECCV) 2018, pp.574–589, 2018.
[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Proc. IEEE International Conference on Computer Vision, pp.2980–2988, 2017.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, "SSD: Single shot multibox detector," European Conference on Computer Vision. Lecture Notes in Computer Science, vol.9905, pp.21–37, Springer, 2016.
[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.779–788, 2016.
[20] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.4294–4302, 2017.
[21] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.5131–5139, 2017.
[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," European Conference on Computer Vision, Lecture Notes in Computer Science, vol.9911, pp.499–515, Springer, 2016.
[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol.15, no.56, pp.1929–1958, 2014.
[24] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," Advances in Neural Information Processing Systems, pp.2017–2025, 2015.
[25] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.4476–4484, 2017.
[26] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," Proc. IEEE International Conference on Computer Vision, pp.5219–5227, 2017.
[27] Y. Wang, V.I. Morariu, and L.S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.4148–4157, 2018.

Kangbo Sun is a Ph.D. student at Shanghai Jiao Tong University. He received a bachelor's degree from Northwestern Polytechnical University in China in 2018. His research interests include computer vision, deep learning, image and video analysis, action recognition, and fine-grained visual analysis.

Jie Zhu received the Ph.D. degree in information system from Shanghai Jiao Tong University (SJTU), China. In 1999, he joined the Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU), China, where he is currently a Professor. His research interests include speech signal processing, multimedia system, deep learning etc. He has been awarded 8 patents and has published more than 150 technical publications.