Beyond the Attention: Distinguish the Discriminative and Confusable Features For Fine-grained Image Classification

Xiruo Shi
shixiruo@bupt.edu.cn
College of Computer Science, Beijing University of Posts and Telecommunications
Beijing, China

Liutong Xu
xliutong@tseg.org
College of Computer Science, Beijing University of Posts and Telecommunications
Beijing, China

Pengfei Wang*
wangpengfei@bupt.edu.cn
College of Computer Science, Beijing University of Posts and Telecommunications
Beijing, China

Yuanyuan Gao
gyy002005@163.com
Information Sciences Academy of China Electronics Technology Group Corporation
Beijing, China

Haifang Jian
jhf@semi.ac.cn
Institute of Semiconductors, Chinese Academy of Sciences
Beijing, China

Wu Liu*
liuwu1@jd.com
AI Research of JD.com
Beijing, China

ABSTRACT

Learning subtle discriminative features plays a significant role in fine-grained image classification. Existing methods usually extract the distinguishable parts through the attention module for classification. Although these learned distinguishable parts contain valuable features that are beneficial for classification, part of irrelevant features are also preserved, which may confuse the model to make a correct classification, especially for the fine-grained tasks due to their similarities. How to keep the discriminative features while removing confusable features from the distinguishable parts is an interesting yet changeling task. In this paper, we introduce a novel classification approach, named *L*ogical-b*a*sed *F*eature *E*xtraction Model (LAFE for short) to address this issue. The main advantage of LAFE lies in the fact that it can explicitly add the significance of discriminative features and subtract the confusable features. Specifically, LAFE utilizes the region attention modules and channel attention modules to extract discriminative features and confusable features respectively. Based on this, two novel loss functions are designed to automatically induce attention over these features for fine-grained image classification. Our approach demonstrates its robustness, efficiency, and state-of-the-art performance on three benchmark datasets.

KEYWORDS

fine-grained image classification; feature fusion; attention

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413883

ACM Reference Format:
Xiruo Shi, Liutong Xu, Pengfei Wang, Yuanyuan Gao, Haifang Jian, and Wu Liu. 2020. Beyond the Attention: Distinguish the Discriminative and Confusable Features For Fine-grained Image Classification. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413883

1 INTRODUCTION

Fine-grained image classification, also known as sub-category image classification, is a very popular research topic in the fields of computer vision and pattern recognition recently. The differences between fine-grained image classification and ordinary image classification are that large intra-class variance and small inter-class variance, as shown in Figure 1. Because the images of fine-grained are generally similar, and distinguished by the subtle and local differences which make fine-grained image classification tasks (e.g., bird species [31], dog species [12], car models [14] and FGVC-Aircraft [23]) still difficult.

Deep learning networks have made good progress on many tasks [6, 22, 27, 29, 36], due to the rapid development of deep learning networks [9, 15, 26, 30], the performance of fine-grained image classification has been continuously improved in recent years. Finding the distinguishable parts and learning the distinguishable features through the convolutional neural network plays a key role in fine-grained image classification. At present, there are two main methods to find distinguishing parts: (1) Obtain distinguishable parts or target bounding boxes directly by manually labeled information (annotation of bounding boxes on objects or parts), which has the advantage of not using extra computing resources. However, manually labeled information is expensive and difficult to obtain. (2) Through weakly supervised learning to automatically find distinguishable parts and extract distinguishable features [5**?**]. In particular, the use of the attention module [10, 20] has greatly helped weakly supervised learning. However, these methods ignore the distinguishable parts that may have confusable features. For example, a bird's eyes are the distinguishable parts, but when using the attention module to locate the bird's eye parts, it is inevitable

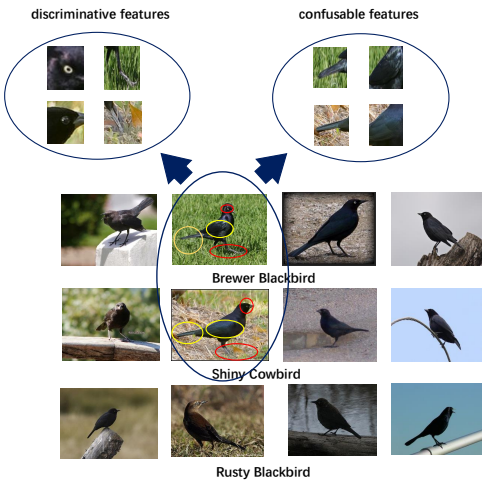

Figure 1: Example of large intra-class variance and small inter-class variance and the examples of discriminative features and confusable features from CUB-200-2011 dataset. We take the images in the circle as an example. The red circles are the distinguishable parts, which are what we want to obtain the discriminative features, and the yellow circles are the similar parts of the two images, which are the confusable features we want to obtain.

that the feather parts around the eye will be obtained at the same time. If other types of birds also have such feathers, the features obtained in this way will include the confusing feature of feathers.

To solve this problem, in this paper, we divide the features of the image into two categories: (1) Discriminative features, which can better distinguish fine-grained images. (2) Confusable features are features that confuse fine-grained image classification or unimportant features, as shown in Figure 1. We argue that the features obtained after adding the discriminative features and subtracting the confusable features are more beneficial to the classification task. This operation brings two benefits: (1)"The stronger the stronger, the weaker the weaker", the features that are helpful for classification will be further enhanced, and the features that are not helpful for classification will be further weakened (2) "Reduce the confusable features", we weaken easily confusable features to eliminate the influence of the confusable features that exist in distinguishable parts. We use squeeze and excitation operations in SENet [10] to obtain discriminative features and confusable features. SENet [10] adaptively recalibrates the characteristic response in terms of channels by explicitly modeling the interdependence between channels. It can learn to use global information to selectively emphasize information features and useless features. Inspired by SENet [10], this paper proposes to establish the relationship between different regions of the image so that it can selectively emphasize the useful and unwanted regions, the detailed method will be introduced in section 3.2. We add channel attention modules and regional attention modules at different stages of the standard backbone network to explore the relationships between different channels and different regions. We use different attention modules to extract different features. There are 4 different attention modules in each stage: (1) discriminative regional attention module, (2) confusable regional attention module, (3) discriminative channel attention module, and (4) confusable channel attention module. We use the discriminative loss to guide the discriminative regional attention modules and discriminative channel attention modules to extract discriminative region features and discriminative channel features. We use the confusable loss to induce confusable regional attention modules and confusable channel attention modules to extract confusable region features and confusable channel features. The discriminative features are obtained by adding the discriminative channel features and the discriminative region features. The confusable features are obtained by adding the confusable channel features and the confusable region features. In order to further enhance the discriminative features and confusable features, we fuse the same attribute features at different stages. The final features used for classification are obtained by adding the fused discriminative features and subtracting the fused confusable features. Our contributions can be summarized as follows:

1) To get better features for image classification, different from the previous methods, this paper not only considers the enhancement of distinguishable parts features, but divides the features into the discriminative features and the confusable features, and obtains the final features by adding the discriminative features and subtracting the confusable features. This operation can not only further enhance the discriminative features, but also we can reduce the negative impact of confusable features in the distinguishable parts on classification.
2) The discriminative features of the same category should be similar, and the confusable features do not have a good classification ability. Based on the above principles, we design two loss functions to induce attention modules to learn discriminative features and confusable features, respectively. The attention modules explore the importance between different channels and the importance between the image regions. The fusion of features of different layers of CNN has a positive effect on the classification task, to further enhance the classification features, we propose a feature fusion method for fusing features at different network levels.
3) We conduct comprehensive experiments on three challenging datasets (CUB Birds, Stanford Cars, Aircraft), and the proposed approach outperforms the state-of-the-art approaches on both datasets.

2 RELATED WORK

With the rapid development of deep networks, the emergence of various networks (e.g. AlexNet [15], VGGNet [26], ResNet [9]) has a positive impact on image classification. For fine-grained image classification, the key task is to obtain features of distinguishable parts for classification. For obtaining discriminative features, there are currently two methods.

One is to use manual labeling information to directly locate the distinguishable parts and then extract features [35, 37]. Zhang et al. [39] proposed a network with mid-level part abstraction layers.

This network is mainly composed of two sub-networks: a detection sub-network and a classification sub-network. Relying on manually labeled information is undoubtedly a simple and effective method [17, 19, 39, 40] for locating distinguishable parts. However, because it is difficult and expensive to obtain manual annotation information, weakly supervised fine-grained image classification methods are gaining more attention. At present, using weakly supervised learning to automatically induce deep neural networks to directly learn distinguishable features is called a one-stage learning approach. Using weakly supervised learning to locate distinguishable regions first, then extract relevant features for classification is called a two-stage learning method.

For one-stage learning methods [25, 38], Lin et al. [18] proposed a new way of feature fusion. They proposed a bilinear model, which uses two independent CNNs to calculate pairwise feature interactions to capture local differences in images. Transfer learning can also be used in various tasks [1, 3]. Cui et al. [3] explored the impact of image resolution on recognition and coping methods for long-tail data. They study transfer learning via fine-tuning from large scale datasets to small scale datasets. Chen et al. [2] proposed a "Destruction and Construction Learning" (DCL) method to enhance the difficulty of fine-grained classification and exercise the classification model to acquire expert knowledge. These methods use feature fusion or feature enhancement to induce neural networks to extract more effective features.

The main method for two-stage learning is to first locate the distinguishable regions and then extract the features of the distinguishable regions [5, 11, 33, 34]. Fu et al. [5] recurrently predicted the location of one attention area and extracted the corresponding features through a novel recursive attention convolutional neural network(RACNN). Wang et al. [33] proposed a correlation-guided discriminative learning model to fully mine and exploit the discriminative potentials of correlations for weakly supervised fine-grained image classification globally and locally. Hu and Qi proposed the WS-DAN model [11] introduced the data augmentation method into the task of fine-grained image classification.

At present, whether the distinguishable parts is obtained through manual annotation or is obtained through the attention module, it is may inevitable that the obtained distinguishable parts contains confusable features. To solve this problem while further enhancing distinguishable features, the method proposed in this paper not only extracts features that are easy to classify (discriminative features) but also the features that make classifications confusing (confusable features). By adding the discriminative features and subtracting the confusable features to get better classification features. Our model is end-to-end training and does not require any additional manual information except labels during the training process.

3 APPROACH

In this section, the model proposed in this paper is described in detail, including the design of the fusion of different levels of features (Sec. 3.1), the regional attention module (Sec. 3.2), and the design of the loss functions that induce attention module to extract discriminative features and confusable features (Sec. 3.3). The overall framework of the model is shown in Figure 2. Given input images, we obtain their features at the third, fourth and fifth stages of backbone network respectively, input these features into the attention modules to extract discriminative features and confusable features, and then fuse the same attribute features of different stages. The final features used for classification are obtained by adding discriminative features subtracting confusable features.

3.1 Feature Fusion at Different Layers

This paper proves its effectiveness on ResNet-50 and ResNet-101 respectively. As we all know, the feature extraction of ResNet is divided into 5 stages. Exploiting semantic features from different layers of CNNs has been shown to be beneficial to many vision tasks. We consider that too low-level features in convolutional neural networks are not abstract and not representative, top-level features lose some details. Therefore, we choose features in the 3, 4, and 5 stages for feature fusion. We send the features obtained in stages 3, 4 and 5 to the regional attention modules and channel attention modules suggested in Section 3.2 to obtain the relationships between regions and the relationships between channels. These attention modules obtain the features of discriminative region features, confusable region features, discriminative channel features and confusable channel features in the 3,4 and 5 stages, respectively, expressed as $\mathbf{r\_dis_i}$, $\mathbf{r\_con_i}$, $\mathbf{c\_dis_i}$, $\mathbf{c\_con_i}$, where $i \in (3, 4, 5)$ represents three stages of 3, 4, 5 respectively.

The discriminative features are obtained by adding the discriminative channel features and the discriminative region features. Similarly, the confusable features are obtained by adding the confusable channel features and the confusable region features. The final features used for classification are the original features plus the discriminative features minus the confusable features. The formula is as follows:

$$\mathbf{dis\_f_i} = \mathbf{r\_dis_i} + \mathbf{c\_dis_i} \quad (1)$$

$$\mathbf{con\_f_i} = \mathbf{r\_con_i} + \mathbf{c\_con_i} \quad (2)$$

$$\mathbf{F_i} = \mathbf{f_i} + \mathbf{dis\_f_i} - \mathbf{con\_f_i} \quad (3)$$

where $\mathbf{dis\_f_i}$ represents discriminative features at i stage, $\mathbf{con\_f_i}$ represents confusable features at i stage, $\mathbf{f_i}$ represents the original features that do not pass the attention modules at i stage, $\mathbf{F_i}$ represents the features obtained by enhancing the discriminative features and weakening the confusable features at i stage, $i \in (3, 4, 5)$.

We fuse features of the same attribute at different stages as shown in the above Figure 2. Formally, let $\mathbf{U_L}$ and $\mathbf{U_{L-1}}$ be the feature maps at stage L and $L-1$ ($L \in (3, 4, 5)$). First, the feature dimension of $\mathbf{U_{L-1}}$ is converted to be consistent with $\mathbf{U_L}$ through the convolution operation, and then the size of the feature map is consistent with that of $\mathbf{U_L}$ through the pooling layer. The procedure can be summarized as:

$$\mathbf{U} = \sigma(BN(\mathbf{W_1}(\mathbf{U_L} + pooling(\sigma(\mathbf{W_2}\mathbf{U_{L-1}}))))) \quad (4)$$

where $\mathbf{U}$ represents the features after fusion. σ represents the activation function, BN represents batch normalize, $pooling$ represents maximum pooling operation, $\mathbf{W_1}$ and $\mathbf{W_2}$ are the parameters that need to be learned.

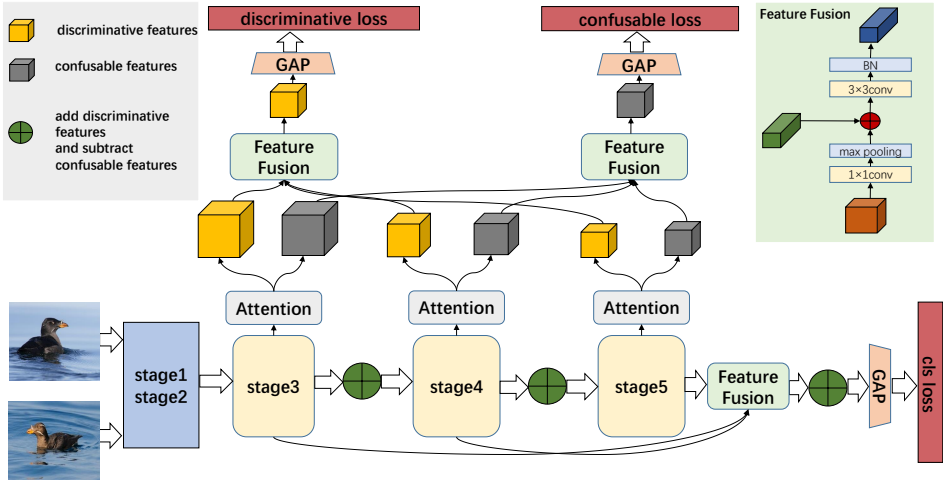

Figure 2: Overview of our approach. Given two input images under the same label, our model extracts discriminative features and confusable features through the attention modules at certain stages, and at the same time, we fuse the features of each stage to obtain better classification features. The final features used for classification are obtained by adding discriminative features subtracting confusable features. In the figure, discriminative loss is represented by L_{dis}, confusable loss is represented by L_{con}, and cls loss is represented by L_{cls}, which will be introduced in section 3.3. In the figure above, the yellow squares represent discriminative features, and the gray squares represent confusable features, and the green round symbols represent the operations of adding discriminative features subtracting confusable features which used to achieve the purpose of enhancing discriminative features and weakening confusable features. The upper right corner of the figure shows the module structure of different stages of feature fusion, the orange square represents the features of the L-1 stage, the green square represents the features of the L ($L \in (3, 4, 5)$) stage, and the blue square represents the fusion features.

3.2 Regional Attention

Inspired by SENet [10], we consider the relationships between image regions. Similar to SENet [10], we also use squeeze and excitation operations to establish the relationships between image regions. SENet [10] represents the entire spatial features on a channel as a global feature through a two-dimensional global pooling operation. In order to obtain all the channel features on each region, we need to first resize the features, and then extract the global channel features through one-dimensional global pooling. The specific details are described below.

Squeeze: Given an input $\mathbf{X} \in R^{C\times W\times H}$, we first need to perform a dimensional transformation on the input $\mathbf{X}$ and compress each feature map into a one-dimensional vector $\mathbf{X'} \in R^{C\times WH}$. Then exchange it with the channel dimension to make it $\mathbf{X'} \in R^{WH\times C}$. Here we use one-dimensional average pooling as the squeeze operation to make it $\mathbf{X'} \in R^{WH}$, which is equivalent to dividing each feature map into the $W \times H$ region, each block integrates the features of all channels.

$$\mathbf{z} = pooling1d(\mathbf{X'}) = \frac{1}{C}\sum \mathbf{X'} \qquad (5)$$

where $\mathbf{X'}$represents input $\mathbf{X}$ after resize.

Excitation: As with the SENet [10] method, two fully connected layers are used to form a bottleneck structure to model the correlations between regions, and output the same number of weights as the input features, the output features are obtained by multiplying the obtained weights with the input features. We first reduce the feature dimensions to 1/7 of the input, and activate it through *ReLu* and then return to the original dimension through a fully connected layer. As shown in Figure 3, using the idea of the squeeze-and-excitation module in SENet [10], we consider both the relationships between image channels and the relationships between regions.

$$\mathbf{s} = \sigma(\mathbf{W_2}\sigma(\mathbf{W_1 z})) \qquad (6)$$

where σ refers to the *ReLU* function, $\mathbf{W_2} \in R^{\frac{WH}{r}\times WH}$ and $\mathbf{W_1} \in R^{WH\times\frac{WH}{r}}$, r is set to 7.

$$\tilde{\mathbf{X}} = resize(\mathbf{X's}) \qquad (7)$$

where $\tilde{\mathbf{X}} \in R^{C\times W\times H}$ represents the final output of region attention.

3.3 Loss Function

We believe that after the same feature extraction network, the key channel features and key region features of image under the same label should be similar, so the discriminative features of the images under the same label should be similar. Based on this condition, this paper makes the distance of discriminative features of images of the same category is closer, shown as:

$$L_{dis} = MAE(GAP(\mathbf{dis\_f_m}), GAP(\mathbf{dis\_f_n})) \qquad (8)$$

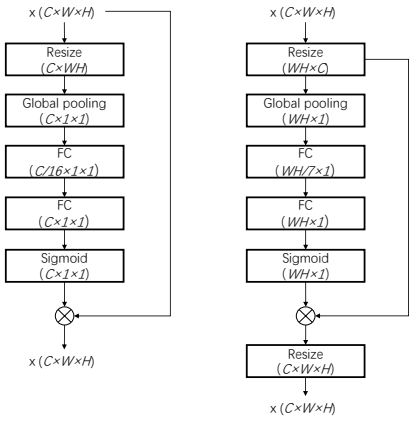

Figure 3: The figure shows the flow of the Squeeze and Excitation operation of the channels and the regions. The left is used to extract the relationships between channels, and the right is used to extract the relationships between regions.

where $\mathbf{dis\_f_m}$ and $\mathbf{dis\_f_n}$ respectively represent the two discriminative features after fusion under the same label, *GAP* represents global average pooling, *MAE* represents mean absolute deviation.

Regarding confusable features, this paper believes that such features are background features or features that easily confuse classification, that is, confusable features are not distinguishable or have a negative impact on fine-grained image classification tasks. Classifiers cannot accurately classify confusable features into a certain correct category but are randomly assigned to all categories. Based on this view, we design the loss function as follows as:

$$L_{con} = MAE(FC(GAP(\mathbf{con\_f_n})), T) \qquad (9)$$

where $\mathbf{con\_f_n}$ represents the confusable features after fusion, *GAP* represents global average pooling, *FC* represents a fully connected layer, T is a vector with a value of 1 for each dimension, and its dimension is the number of categories, *MAE* represents mean absolute deviation.

We send the final features to the classifier for classification, and use cross-entropy as the loss function to classify the loss.

$$L_{cls} = -\sum l \cdot log[C(GAP(\mathbf{F_5}))] \qquad (10)$$

where C represents a trainable classifier for final classification, *GAP* represents global average pooling, l represents image labels, and $\mathbf{F_5}$ represents the final features used for classification, that is the features of the fifth stage after fusion.

In our framework, we train the network in an end-to-end manner, specifically, we want to minimize the following objective:

$$L = L_{cls} + L_{dis} + L_{con} \qquad (11)$$

4 EXPERIMENTS

In this section, we present performance evaluations and analysis of our proposed method on three publicly available fine-grained classification datasets, and we explore the contribution of each proposed module.

Table 1: Comparison with state-of-the-art methods on CUB-200-2011 dataset

Methods	Anno.	1-Stage	Accuracy
DeepLAC [17]	✓	✓	80.3
NAC [25]	×	✓	81.0
Part-RCNN [40]	✓	×	81.6
PA-CNN [13]	✓	×	82.8
SENet-50 [10]	×	✓	83.0
B-CNN [18]	×	×	84.1
FCAN [19]	✓	✓	84.3
Kernel-Pooling [4]	×	✓	84.7
SPDA-CNN [39]	✓	✓	85.1
RA-CNN [5]	×	×	85.3
DT-RAM [16]	×	×	86.0
MAMC-CNN [28]	×	✓	86.2
DFB-CNN [32]	×	✓	87.4
Cross-X [21]	×	✓	87.7
DCL [2]	×	✓	87.8
CIN [7]	×	✓	**88.1**
LAFE(ResNet-50)	×	✓	87.6
LAFE(ResNet-101)	×	✓	**88.1**

Table 2: Comparison with state-of-the-art methods on Stanford Cars dataset

Methods	Anno.	1-Stage	Accuracy
DVAN [24]	×	×	87.1
FCAN [19]	✓	✓	89.1
SENet-50 [10]	×	✓	91.6
Kernel-Pooling [4]	×	✓	92.4
RA-CNN [5]	×	×	92.5
MAMC-CNN [28]	×	✓	93.0
DT-RAM [16]	×	×	93.1
DFB-CNN [32]	×	✓	93.8
WS-DAN [11]	×	×	94.5
DCL [2]	×	✓	94.5
CIN [7]	×	✓	94.5
Cross-X [21]	×	✓	94.6
LAFE(ResNet-50)	×	✓	94.8
LAFE(ResNet-101)	×	✓	**94.9**

4.1 Datasets and Baselines

To evaluate the effectiveness of our proposed model, we performed experiments on three broad and competitive datasets: Caltech-UCSD Birds (CUB-200-2011) [31], Stanford Cars [14], and FGVC-Aircraft [23]. Caltech-UCSD Birds contains 11,788 images of 200 types of birds, 5,994 for training and 5,794 for testing. Stanford Cars contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images. FGVC-Aircraft contains 10,000 images of 100 types of aircraft, 6,667 for training and 3,333 for testing. We compare with the following baselines, due to their state-of-the-art results. All the baselines are listed as follows:

Table 3: Comparison with state-of-the-art methods on FGVC-Aircraft dataset

Methods	Anno.	1-Stage	Accuracy
B-CNN [18]	×	×	84.1
Kernel-Pooling [4]	×	✓	85.7
RA-CNN [5]	×	×	88.2
SENet-50 [10]	×	✓	90.6
DFB-CNN [32]	×	✓	92.0
Cross-X [21]	×	✓	92.6
CIN [7]	×	✓	92.8
WS-DAN [11]	×	×	93.0
DCL [2]	×	✓	93.0
LAFE(ResNet-50)	×	✓	93.3
LAFE(ResNet-101)	×	✓	**93.6**

- **Part-RCNN** [40]: extends R-CNN [8] based framework by part annotations.
- **PA-CNN** [13]: part alignment-based method generates parts by using co-segmentation and alignment.
- **NAC** [25]: neural activation constellations find parts by computing neural activation patterns.
- **DVAN** [24]: a weakly-supervised iterative scheme, which shifts its center of attention to increasingly discriminative regions as it progresses, by alternating stages of classification and introspection.
- **FCAN** [19]: fully convolutional attention network adaptively selects multiple task-driven visual attention by reinforcement learning.
- **DeepLAC** [17]: deep localization, alignment and classification proposes to use a pose-aligned part image for classification.
- **MAMC** [28]: applies the multi-attention multi-class constraint in a metric learning framework, a novel attention-based convolutional neural network which regulates multiple object parts among different input images.
- **RACNN** [5]: recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way.
- **B-CNN** [18]: uses two separate feature extractors to capture pairwise feature interactions for classification.
- **DT-RAM** [16]: recurrent visual attention model that selects a sequence of regions through a dynamic continue/stop gating mechanism.
- **SPDA-CNN** [39]: semantic part detection and abstraction proposes to generate part candidates and extract features by detection/classification networks.
- **DFB-CNN** [39]: discriminant filter bank method for learning convolutional filter banks that capture specific class discriminant patches.
- **WS-DAN** [11]: proposes weakly supervised data augmentation network to explore the potential of data augmentation to improve the performance of fine-grained image classification.
- **Cross-X** [21]: a simple yet effective approach that exploits the relationships between different images and between different network layers for robust multi-scale feature learning.
- **DCL** [2]: a novel "Destruction and Construction Learning" method to enhance the difficulty of finegrained recognition and exercise the classification model to acquire expert knowledge.
- **CIN** [7]: proposes a channel interaction network, which models the channel-wise interplay both within an image and across images.

4.2 Implementation Details

We evaluate our proposed method on a widely used backbone network ResNet-50 and ResNet-101. The ResNet-50 and ResNet-101 are pre-trained on the ImageNet dataset. The input images are resized to a fixed size of 512×512 and randomly cropped to 448×448. We apply random rotation and random horizontal flips to data augmentation. In the third, fourth, and fifth stages, we add additional attention modules, as shown in section 3.2, to obtain the image channels and regions that have a positive effect on classification and the image channels and regions that play a negative role in classification. The former is called discriminative channel features and discriminative region features, and the latter is called confusable channel features and confusable region features. The discriminative channel features and the discriminative region features are combined to obtain the discriminative features, and the confusable channel features and the confusable region features are combined to obtain the confusable features. The shapes of the output feature map are $56 \times 56 \times 512$, $28 \times 28 \times 1024$, $7 \times 7 \times 2048$. The features of these three stages are fused according to the method in section 3.1 to obtain the final features for classification. No part or bounding box annotations are used during training.

We train the models using Stochastic Gradient Descent (SGD) with the momentum of 0.9, epoch number of 200, weight decay of 0.0005, and a mini-batch size of 16. In order to narrow the distance of discriminative features under the same label, we need to make the images under the same label appear in pairs during the training process, so we customize the sampling method so that the two adjacent images in each batch have the same label. The initial learning rate is set to 0.0015, with exponential decay of 0.1 after every 40 epochs.

4.3 Comparison with State-of-the-Art

Results on CUB-Birds: The classification results for CUB birds are presented in Table 1. Compared with previous methods, our method has better performance. For the two-stage method, Part-RCNN [40],PA-CNN [13], B-CNN [18], RA-CNN [5], and DT-CNN [16] achieve 81.6%,82.8%, 84.1%, 85.3%, and 86.0% accuracy respectively on the CUB-200-2011 dataset. Compared with them, the accuracy of our model is 6.5%, 5.3%, 4.0%, 2.8%, 2.1% higher than them. Two-stage learning is not end-to-end training, it is more difficult than one-stage training. For the one-stage model, researchers have proposed various novel approaches to study from various aspects. The kernel pool reaches 84.7%, and MAMC-CNN [28] which learns multiple feature maps by embedding OSME blocks into the metric learning framework, its accuracy reaches 86.2%. SPDA-CNN [39]

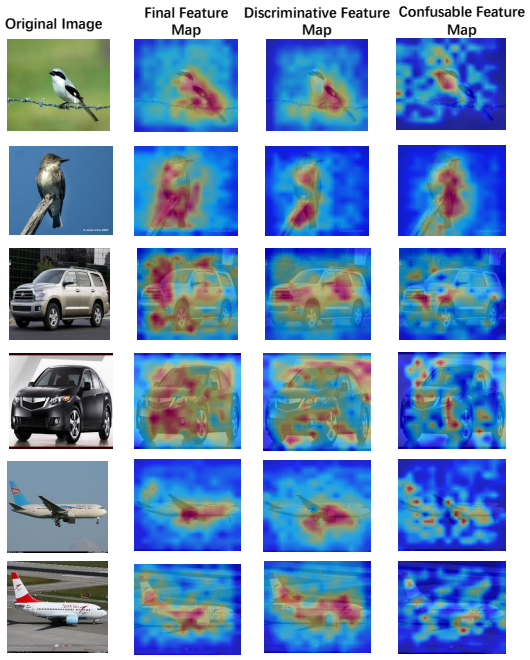

Figure 4: Visual feature maps. The first column is the original images, and the second column is the feature maps for classification which is obtained by the feature maps of the last convolutional layer of ResNet-50 plus discriminative features minus confusable features. The third column is the discriminative features obtained by the fifth stage of ResNet-50. The fourth column is the confusable features obtained by the fifth stage of ResNet-50.

proposed network has two sub-networks: one for detection and one for recognition. It has an accuracy of 85.1% on CUB-200-2011 dataset, but it uses manual annotation information. DFB-CNN [32] uses the discriminant filter bank method for learning convolutional filter banks that capture specific class discriminant patches, and its accuracy rate has reached 87.4%, but as can be seen from Table 1, the accuracy of our model LAFE is still better than them. The accuracy of LAFE is 0.4% higher than the recently proposed Cross-X [21] model. Compared with DCL [2], which proposes a novel "Destruction and Construction Learning" to acquire expert knowledge, the accuracy of LAFE is 0.3% higher than it. From the results on the Table 1, compared with CIN [7], which models the channel-wise interplay both within an image and across images. The accuracy of LAFE is equal to it on the CUB-200-2011 dataset, but the accuracy rate of LAFE on both other datasets exceeds CIN. We consider that it may be due to the small target of the CUB-200-2011 dataset. Compared with the large target on the Stanford Cars and FGVC-Aircraft datasets, LAFE would be more difficult to capture more detailed parts on the small target. This is what we need to improve in the future.

Results on Stanford Cars: The classification results for Stanford Cars are presented in Table 2. Our model achieves state-of-the-art performance on this dataset. For the two-stage models DVAN [24], RA-CNN [5], DT-RAM [16], and WS-DAN [11], their accuracy rates are 87.1%, 92.5%, 93.1%, and 94.5%, respectively. Compared with them, the accuracy of our model is 7.8%, 2.4%, 1.8%, 0.4% higher than them. WS-DAN [11] advocates focusing on other regions that are not critical and they explore the potential of data augmentation. They generate attention maps by performing weakly supervised learning on each training image, and enhance the images guided by these attention maps, including attention cropping and attention dropping. For the comparison of the one-stage models, LAFE still achieved the best results. Similarly, the accuracy of LAFE is 1.1%, 0.4%, and 0.3% higher than DFB-CNN, DCL [2], and Cross-X [21] in recent years. Cross-X [21] also uses squeeze and excitation operations, it exploits the relationships between different images and between different network layers for robust multi-scale feature learning. Compared with the recently proposed CIN [7], LAFE still achieves excellent performance. The accuracy of LAFE is 0.4% higher than that of CIN.

Table 4: An ablation study of proposed methods for recognition accuracy on three different datasets. The first line shows the experimental results of LAFE on three datasets. - Region Attention represents the model after removing the regional attention module on the LAFE network. - Feature Fusion represents the model after removing feature fusion at different stages on LAFE. - Feature Loss Function represents the removal of the discriminative features loss and the confusable features loss proposed in Section 3.3. - Confusable Loss Function represents the experimental results after only removing the confusable features loss function on LAFE, these experimental results better prove the effectiveness of the confusable features loss function. The last line shows the experimental results of the backbone network resnet-50 on three datasets.

Methods	CUB	CAR	AIR
LAFE	87.6	94.8	93.2
- Region Attention	87.2	94.6	92.9
- Feature Fusion	85.5	94.1	92.4
- Feature Loss Function	86.8	94.2	92.7
- Confusable Loss Function	87.0	94.5	92.9
resnet-50	84.5	92.9	90.3

Results on FGVC-Aircraft: The classification results for FGVC-Aircraft are presented in Table 3. Our model achieves state-of-the-art performance on this dataset. For the two-stage models, B-CNN [18], RA-CNN [5], WS-DAN [11], their accuracy rates are 84.1%, 88.2%, 93.0%, respectively. Compared with them, the accuracy of our model is 9.5%, 5.4%, 0.6% higher than them. Similarly, compared to the one-stage model, the accuracy of LAFE is still better than them. Compared with DCL [2] and Cross-X [21], the accuracy of LAFE is 0.6% and 1.0% higher than them, respectively. Compared with CIN, the accuracy of LAFE is 0.8% higher than the accuracy of CIN on the FGVC-Aircraft dataset.

4.4 Ablation Studies

We performed ablation studies to understand the different components in our proposed model. We designed different runs in the three datasets on ResNet-50 and reported the results in Table 4.

The Effect of Region Attention: Experiments show that using the region attention module is more effective than not using the module. After removing the regional attention modules, the accuracy of the CUB-200-2011 dataset, the Stanford dogs dataset, and the FGVC-Aircraft dataset decreased by 0.4%, 0.2%, and 0.3%, respectively, which indicates that the introduction of the region attention module proposed in 3.2 has played a positive role in the final classification.

The Effect of Feature Fusion: Similarly, the fusion of features at different stages also has a positive impact on the final classification. According to the data in Table 4, after removing the Feature Fusion operation, the accuracy of the CUB-200-2011 dataset, the Stanford dogs dataset, and the FGVC-Aircraft dataset decreased by 2.1%, 0.7%, and 0.8%, respectively. Feature Fusion operation has a greater impact on the CUB-200-2011 dataset. Because the CUB-200-2011 dataset has more classification categories, and less training data and smaller classification targets, the low-level convolution retains detailed information, so compared to other datasets, feature fusion at different stages has a greater impact on the CUB-200-2011 dataset.

The Effect of Feature Loss Function: In order to obtain the discriminative features and confusable features, the two loss functions proposed have played a positive role in LAFE. We induce channel attention modules and regional attention modules training by reducing the distance of discriminative features under the same label. Similarly, we use a classifier to map confusable features to a vector whose dimension is the number of categories with a value of 1 to induce channel attention modules and regional attention modules training. As shown in Table 4, after deleting these two loss functions, the accuracy of the three datasets has decreased significantly. The rates decreased by 0.8%, 0.6%, and 0.5%, respectively. As shown in Figure 5, the second column is the final feature map for classification. They add the discriminative features of the third column and subtract the confusable features of the fourth column. Discriminative features and confusable features are obtained through attention modules that explore relationships between channels and relationships between regions, respectively.

The Effect of Confusable Features Loss Function: In order to highlight the role of confusable features, we also design an experiment to remove the loss function of confusable features. As shown in Table 4, after removing the loss function of confusable features, the accuracy of the three datasets decreased by 0.6%, 0.3%, and 0.3%, respectively. These results further illustrate the significance of weakening the confusable features. Our confusable features loss function induce the attention module to learn the confusable features, and weakening the confusable features plays a positive role in final classification.

4.5 Visualization

Figure 5 shows the activation maps of 6 images from 3 datasets. We send the feature maps obtained by the last convolution of the backbone network to different attention modules to obtain discriminative features and confusable features. The visualization images of discriminative features and the visualization images of confusable features are represented by the third and fourth columns in Figure 5, respectively. It can be observed that the discriminative features maps and confusable features maps some have different attention parts on the image. The focus area of the discriminative features maps are mainly concentrated around the object or the object's key area, the attention area of the confusable features maps are relatively small and mainly focuses on the area details of the object. For example, the discriminative features maps of the bird in the first line focus on the bird's wings and bird's paws, while the confusable features maps focus on the abdomen of birds. However, some discriminative features and confusable features have the same attention parts, which means that the parts have both discriminative features and confusable features. For example, the area of the bird's neck in the second row contains both discriminative features and confusable features. The results show that the features obtained by locating the distinguishable parts only through the attention module are likely to contain confusable features. We try to eliminate such effects by subtracting the learned confusable features. We fuse discriminative features and confusable features to get the final classification features, as shown in the second column of Figure 5, that is, we add the discriminative features then subtract the confusable features to enhance the features that have a positive impact on the classification and weaken the features that have a negative impact on the classification.

5 CONCLUSION

Our method uses the squeeze-and-excitation module to learn the relationship between channels and the relationship between regions to obtain discriminative features and confusable features, and uses the feature fusion module to fuse features at different stages to enhance features, the final features used for classification is the original features of the last stage plus discriminative features minus confusable features. The confusable features of the images under the same label should be the same, and the confusable features are not important or easily confuse the final classification. Based on these two principles, we design two loss functions to reduce the distance between discriminative features under the same label and map confusable features to a vector with a value of 1 for each dimension, and its dimension is the number of categories. Our model is proven to be effective on three public datasets, and ablation experiments further prove the effectiveness of different models.

ACKNOWLEDGMENTS

This research work was partially supported by the National Natural Science Foundation of China under Grant No.61802029, the Open Project Funding of CAS-NDST Lab under Grant No.CASNDST202005, the National Natural Science Foundation of China (No.61801448), and the National Natural Science Foundation of China (Grant No.61802022).

REFERENCES

[1] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. 2020. Tiny Transfer Learning: Towards Memory-Efficient On-Device Learning. *CoRR* abs/2007.11622 (2020).

[2] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. 2019. Destruction and Construction Learning for Fine-Grained Image Recognition. In *CVPR*. 5157–5166.

[3] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge J. Belongie. 2018. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. In *CVPR*. 4109–4118.

[4] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge J. Belongie. 2017. Kernel Pooling for Convolutional Neural Networks. In *CVPR*. 3049–3058.

[5] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *CVPR*. 4476–4484.

[6] Chuang Gan, Tianbao Yang, and Boqing Gong. 2016. Learning Attributes Equals Multi-Source Domain Generalization. In *CVPR*. 87–97.

[7] Yu Gao, Xintong Han, Xun Wang, Weilin Huang, and Matthew Scott. 2020. Channel Interaction Networks for Fine-Grained Image Categorization. In *AAAI*. 10818–10825.

[8] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*. 580–587.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[10] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *CVPR*. 7132–7141.

[11] Tao Hu and Honggang Qi. 2019. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. *CoRR* abs/1901.09891 (2019).

[12] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel Dataset for Fine-Grained Image Categorization. In *CVPR*.

[13] Jonathan Krause, Hailin Jin, Jianchao Yang, and Fei-Fei Li. 2015. Fine-grained recognition without part annotations. In *CVPR*. 5546–5555.

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCV*. 554–561.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Commun. ACM*. 1106–1114.

[16] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. 2017. Dynamic Computational Time for Visual Attention. In *ICCV*. 1199–1209.

[17] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. 2015. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*. 1666–1674.

[18] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhransu Maji. 2015. Bilinear CNN Models for Fine-Grained Visual Recognition. In *ICCV*. 1449–1457.

[19] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. 2016. Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition. *CoRR* abs/1603.06765 (2016).

[20] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. 2018. Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification. In *CVPR*. 7834–7843.

[21] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry Davis, Jun Li, Jian Yang, and Ser-Nam Lim. 2019. Cross-X Learning for Fine-Grained Visual Categorization. In *ICCV*. 8241–8250.

[22] Jinna Lv, Wu Liu, Meng Zhang, He Gong, Bin Wu, and Huadong Ma. 2017. Multi-feature Fusion for Predicting Social Media Popularity. In *ACM Multimedia*. 1883–1888.

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. 2013. Fine-Grained Visual Classification of Aircraft. *CoRR* abs/1306.5151 (2013).

[24] Amir Rosenfeld and Shimon Ullman. 2016. Visual Concept Recognition and Localization via Iterative Introspection. In *ACCV*. 264–279.

[25] Marcel Simon and Erik Rodner. 2015. Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks. In *ICCV*. 1143–1151.

[26] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[27] Yicheng Song, Yong-Dong Zhang, Juan Cao, Tian Xia, Wu Liu, and Jin-Tao Li. 2012. Web Video Geolocation by Geotagged Social Resources. *IEEE Trans. Multimedia* 14, 2 (2012), 456–470.

[28] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. 2018. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. In *ECCV*. 834–850.

[29] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. 2019. Human Mesh Recovery From Monocular Images via a Skeleton-Disentangled Representation. In *ICCV*. 5348–5357.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report.

[32] Yaming Wang, Vlad I. Morariu, and Larry S. Davis. 2018. Learning a Discriminative Filter Bank Within a CNN for Fine-Grained Recognition. In *CVPR*. 4148–4157.

[33] Zhihui Wang, Shijie Wang, Pengbo Zhang, Haojie Li, Wei Zhong, and Jianjun Li. 2019. Weakly Supervised Fine-grained Image Classification via Correlation-guided Discriminative Learning. In *ACM Multimedia*. 1851–1860.

[34] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. 2018. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognit.* 76 (2018), 704–714.

[35] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. 2013. Hierarchical Part Matching for Fine-Grained Visual Categorization. In *ICCV*. 1641–1648.

[36] Ning Xu, Hanwang Zhang, An-An Liu, Weizhi Nie, Yuting Su, Jie Nie, and Yongdong Zhang. 2020. Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning. *IEEE Trans. Multimedia* 22, 5 (2020), 1372–1383.

[37] Shulin Yang, Liefeng Bo, Jue Wang, and Linda G. Shapiro. 2012. Unsupervised Template Learning for Fine-Grained Object Recognition. In *NIPS*. 3131–3139.

[38] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. 2018. Learning to Navigate for Fine-Grained Classification. In *ECCV*. 438–454.

[39] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed M. Elgammal, and Dimitris N. Metaxas. 2016. SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Recognition. In *CVPR*. 1143–1152.

[40] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. 2014. Part-Based R-CNNs for Fine-Grained Category Detection. In *ECCV*. 834–849.