

Received May 20, 2020, accepted May 31, 2020, date of publication June 3, 2020, date of current version June 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999722

Two-Level Progressive Attention Convolutional Network for Fine-Grained Image Recognition

HUA WEI^{1,2}, MING ZHU¹, BO WANG^{1,2}, JIARONG WANG^{1,2,3}, AND DEYAO SUN^{1,2}

¹Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³Changchun University of Science and Technology, Changchun 130022, China

Corresponding author: Ming Zhu (zhu_mingca@163.com)

ABSTRACT The learning of discriminative features is the key for fine-grained image recognition. To better extract effective features and improve the accuracy of fine-grained image recognition, we propose a two-level progressive attention convolutional network (TPA-CNN) for fine-grained image recognition. The model includes a multi-channel attention-fusion (MCAF) module and a cross-layer element-attention (CEA) module. The MCAF module is used to find distinctive feature map channels which significantly responds to specific regions. Inspired by Hierarchical Bilinear Pooling model, The CEA module is further assign weight values to feature map elements. From the perspective of attention visualization map, MCAF module can focused on one or more positive regions, CEA module further locates the most discriminative regions by interaction between the feature map elements. The model can dynamically search the discriminative region of the image, not limited to the boost or crop a selected region. Compared to previous models basing on attention mechanism, the model can extract non-correlated part features which spread over object foreground areas, further improving the recognition accuracy. Experimental results on CUB-200-2011, FGVC-Aircraft, and Stanford Cars datasets demonstrate that the proposed TPA-CNN achieves competitive performance.


INDEX TERMS Fine-grained image recognition, visual attention mechanism, convolution neural network, feature extracting.

I. INTRODUCTION

Image recognition is divided into generic and fine-grained image recognition [1]. Fine-grained image recognition differs from generic image recognition, which is to classify the object into a finer sub-class category, e.g., distinguishing wild birds, aircraft models, car models, and tank models. A successful fine-grained image recognition model can accurately identify the sub-class category without relying on expert experience. However, it is arduous to distinguish the categories precisely, where the sub-class categories invariably own same global appearance, and intra-class image features have large differences. Hence, the subtle and region features are the effective information for distinguishing subordinate categories.

Owing to the development of deep neural networks, great opportunities for computer vision are brought about. A large number of representative works have emerged in many computer vision tasks such as object detection [2]–[4], semantic segmentation [5], [6], recognition [7]–[9], and

some interdisciplinary subjects [10]. Convolutional neural networks have a powerful ability to extract feature, many researchers are attempting to use it for fine-grained recognition. In the early stage of fine-grained recognition researches, [11]–[13] have got some better results with supervised bounding box or part annotations, but the additional supervision information requires considerable resources, thus, the weak supervision model which only needs image category labels received much more attention in recent researches. In order to investigate the discriminative part for fine-grained image recognition, to the best of our knowledge, [7] early attempted to use the selective search to generate part proposals. Then, the methods [14]–[16] based on attention mechanism are designed from the perspective of visual attention mechanism, which can help researchers to suppress invalid areas and obtain effective areas in the fine-grained images. Thus, using visual attention mechanism to find discriminative regions has gained a series of researchers' preferences. Another method [17], [18] for extracting subtle and region features is based on the Bilinear Convolutional Neural Networks (B-CNNs). Those methods use a quiet concise model obtaining high-level features through

The associate editor coordinating the review of this manuscript and approving it for publication was Gulistan Raja .

the complementary sub-networks. All these works investigated the key influencing factors in fine-grained recognition which is searching for discriminative regions. However, current methods based on CNNs exist certain drawback. Some approaches, which use supervised bounding box or part annotations, got a higher accuracy but consumed huge human resources for labelling image datasets. Others just use a weakly supervised image label, but lack a mechanism to locate on the right regions, which usually leads to a decline in accuracy. Most effective work in fine-grained image recognition focuses on how to locate regions and extract region features. Due to the difficulty of the small differences among inter-class and large differences within intra-class, the task for fine-grained image recognition just using weakly supervised image label to focus on the right regions has been still an exceedingly further challenge.

In this work, inspired by above observations, we propose a two-level progressive attention convolutional network (TPA-CNN) from the perspective of feature map channels and elements for fine-grained image recognition. We observe that works based on B-CNNs taking notice of influence of mutual response of element between layers of neural network, and we regard this response as the degree to which feature map elements the model pays attention to. From the perspective of attention mechanism, the more important elements in a feature map should have greater responsiveness, in contrast, the less important elements should have less response. However, inter-layer interactions pour attention into all elements without differences. As we know, feature map has three dimensions: height, width, and channel. Accordingly, adding attention processing to the channels of the feature map before re-assigning feature map elements will effectively improve the ability to find the discriminative regions. We innovatively take the inter-layer interactions as cross-element attention (CEA) module to pay attention to elements of the feature map. Corresponding to CEA module, we propose a novel multi-channel attention-fusion (MCAF) module which pays attention to channels of the feature map. Visually, we firstly use the MCAF module to obtain one or more high response regions, further enhancing the response to the most discriminative features through CEA module.

The main contributions of our work are as following:

- We propose a novel model on account of the visual attention mechanism improving the accuracy of fine-grained image recognition. This model includes a MCAF module and a CEA module which re-assigns weights of a feature map from channels to elements, so our model can be regarded as a two-level progressive attention-based model.
- The MCAF module is proposed to re-assigns channel weights. By using different reduction ratio r and fusing features through numbers of filter kernels, a feature map group with channel attention is formed. Inspired by B-CNNs, we theoretically prove that inter-layer feature interactions can be regarded as an element attention method. According to the characteristics of element attention mechanism, CEA module is designed to pick import elements from a feature map.

Both of MCAF module and CEA module are modular and can be applied to others models or tasks.

- In order to verify the accuracy and universality of the proposed model, we conduct experiments on three most representative fine-grained datasets and demonstrate the effectiveness of our proposed model. Moreover, we also conduct ablation studies to carefully analyze the contribution of each module of our model and important parameters in each module.

II. RELATED WORK

Our work mainly involves two aspects of research: fine-grained image feature extraction and attention mechanism. We will present the recent research results in this section.

A. FINE-GRAINED IMAGE FEATURE EXTRACTION

As mentioned above, feature extraction for fine-grained image recognition is more challenging than generic image recognition tasks. In the technical report issued by CUB200-2011 dataset [19], the recognition accuracy baseline is only 10.3%. Early algorithms based on the artificial features, such as POOFs [20], those methods enhanced the ability of feature extraction in object recognition algorithms to further improve the accuracy. From this point, it can be found that more powerful feature extraction algorithm has a more significant impact on fine-grained image recognition algorithm. Deep convolutional neural network had confirmed this conclusion and brought about tremendous changes for feature extraction and improved the accuracy far more than previous methods in fine-grained image recognition. The works for fine-grained image recognition based on deep convolutional neural networks can be divided into two methods. The first method is part localization and learning, and the second method is feature encoding and decoding [21]. For the first method, Part R-CNN [7] is used to select discriminative regions by the intersection over union (IOU) of the candidate regions and the bounding box to extract subtle features. Part R-CNN [7] and these methods [22]–[24] based on CNNs improve the ability of feature extraction and take certain effects, but the use of bounding box or part annotations results in a high cost. The additional supervision information requires a lot of manpower and financial resources, thus, the weak supervision models, such as [25]–[27], only need image category labels which received much more attention in recent researches. In order to locate part and extract feature accurately, Mask-RCNN [11] transforms feature's localization into segmentation of head, body, and background, and then global and local image features are extracted. PDFS [16] selected the discriminative regions by deep filter response. DFL [28] designed a bank of convolutional filters to improve the mid-level feature learning ability of CNNs. NTS-net [29] proposed a self-supervision mechanism to locate the discriminative regions. The second representative method is B-CNNs [28], which performs the feature extraction through the coordination of two neural networks which have the

same architecture. However, due to a very high dimension, the bilinear feature is difficult to be applied in practice. In order to improve the application ability of the B-CNNs, [17], [30], [31] reduced the parameter dimension of B-CNNs, and [18], [32], [33] made some variants based on B-CNNs to improve the accuracy of fine-grained image recognition. These works have brought a lot of inspiration to our work.

B. ATTENTION MECHANISM

Attention mechanism [34] was firstly applied to machine translation. Reference [35] introduced it to computer vision and brought a lot of inspiration for computer vision tasks. Owing to the attention mechanism has the ability to select interesting areas, researchers dynamically have focused on the discriminative regions of the image, rather than directly processing all the information of an image. When designing the deep neural network, we imitate the human visual signal processing mechanism: quickly scan the global image to find the concerned target areas, and then suppress other useless information. The approach described in this section differs from that previous section in that the visual attention mechanism allows the model to focus on critical locations, ignoring unimportant locations using weakly supervised image label. References [36]–[38] are the earlier attempt to use visual attention for fine-grained recognition. Then, many typical works have been created, RACNN [39] uses a three-scale sub-network to gradually crop and enlarge the area through attention proposal network (APN). The class activation mapping [40] (CAM) provides a better way to visualize the attention effect. MAMC [14] used the attention mechanism and metric learning to screen out two unrelated discriminative features in a fine-grained image. MA-CNN [15] generates multiple attentional parts which can reinforce each other by clustering, weighting, and channel grouping. The visual attention mechanism has brought a lot of excellent exploration. However, it can be seen from the previous works that the CNNs design are often more complicated. From the simple and efficient purpose of network design, our work based on the following points: 1) when looking for the category of the interested object in an image, the information in the image is not all useful for recognition, some of them are useless, and may even result in interference; 2) there may be more than one information areas in an image, and two or more areas provide reliable features for fine-grained recognition. The number of discriminative regions may not be fixed, and can be set freely through the design neural network. To meet these demands, we present a novel framework that contains two attention modules which not only obtains discriminative regions directly, but also dynamically selects the number of discriminative regions. The two attention modules we designed are modular and can be applied to other fine-grained image recognition scenarios or a slice of other tasks.

III. MODEL

The main idea behind our TPA-CNN is to construct a modular design network that can effectively and flexibly find

discriminative parts features using soft-attention mechanism without bounding box or part annotation. Fig. 1 illustrates the architecture for fine-grained image recognition, which consists of three components: (1) multi-channel attention-fusion (MCAF) module; (2) cross-element attention (CEA) module; (3) Loss function. We briefly introduce the pipeline of the whole model for the first time. Then, we introduce MCAF module and explain the CEA mechanism by factorizing bilinear pooling through delving deep into formulation. Finally, we introduce the loss function.

A. APPROACH VIEW

For fine-grained image recognition, the parts which have positives for distinguishing are subtle and local. The foreground image of the object cannot be seen as a discriminative part, the subtle features which can fully express the object are not directly available. Our method rests on the assumption that for fine-grained recognition, the foreground image can be divided into many regions, these regions do not have the same value for recognition, and the truly distinguishing regions should have higher response. Based on this, we design the network with convolution, channel-wise, and element-wise as shown in Fig. 1. We pass an input image through a basic network such as ResNet-50 to obtain three high-level original feature maps. These three feature maps have the same dimension, and every feature map can be written as $x \in \mathbb{R}^{C \times H \times W}$. The first stage is to pass every feature map through MCAF module to get the fusion feature map. By operating the channels of feature map that can discover one or more discriminative part. The MCAF module uses different channel attention and feature fusion operation to facilitate the one original feature map $x \in \mathbb{R}^{C \times H \times W}$ to generate three sets of feature maps. The three sets of feature maps are concatenated to obtain the re-weight feature map written as $x \in \mathbb{R}^{4C \times H \times W}$. Assuming we have obtained a fused feature map written as $x \in \mathbb{R}^{4C \times H \times W}$ after passing through the MCAF module, we can see an activate map that activates several regions through convolution layer visualization operation. In our TPA-CNN, three sets of high-level original feature maps were selected to pass through the MCAF module, so we can obtain three fused feature maps. The second stage is using two different fusion feature maps to screen the most discriminative region through the CEA module. When the three set feature maps got from the first stage and one of them can be written as $x \in \mathbb{R}^{4C \times H \times W}$, these three sets of feature maps select different two for multiplication. There are three choices in total, the CEA module selects all three choices, so we can have three sets of feature maps with element attention through CEA module. Thus, when three original feature maps passed through the MCAF module, three feature maps with channel attention owing different activation responses have been obtained. Through the CEA module, feature map elements with higher response values will be prominent, and the weak responses will be correspondingly reduced. In the second and third sections, we will elaborate on the mathematical principles of implementation.

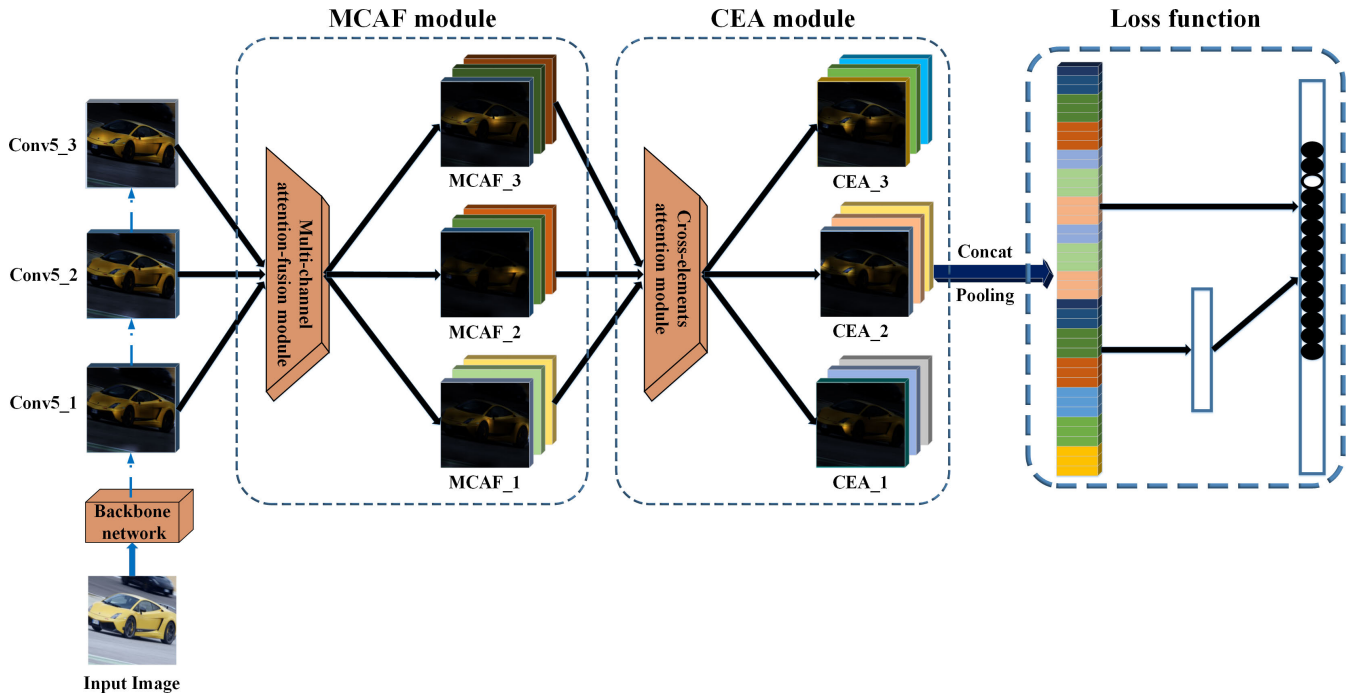


FIGURE 1. The overview of our model. The channel attention-fusion module makes a select of channels, and the cross-layer element-attention module makes a select of elements.

B. THE MAIN MODULE OF MODEL

For fine-grained image recognition, an important advantage of convolutional neural networks over traditional methods is that they have a powerful capability to acquire object’s features. A carefully designed neural network can obtain more effective feature maps to improve the accuracy of recognition. From the point to get more effective feature maps for fine-gained image recognition, we design two attention-based modules making up our main convolutional neural network model to improve the ability of extracting image features. Our model includes two main modules: MCAF module and CEA module. TPA-CNN firstly processes the feature map channels using MCAF module, and then processes the feature map elements via CEA module.

1) MCAF MODULE

According to the characteristics of the feature map, in this part, we designed MCAF module to dig more information concerning channel. In the next, we will introduce it in detail.

Suppose an image is entered in the CNNs and we obtain the output feature map written as $x \in \mathbb{R}^{c \times h \times w}$ with channel c , height h , and width w of the last convolutional layer, as shown in Fig. 2. Feature map $x \in \mathbb{R}^{c \times (h \times w)}$ can be described as $x = [x_1 \ x_2 \ \dots \ x_c] \in \mathbb{R}^{c \times h \times w}$. We use Global Average Pooling (GAP) to get a channel-wise descriptors $z = [z_1 \ z_2 \ \dots \ z_c] \in \mathbb{R}^c$ by reducing feature map x to its spatial dimension $H \times W$. The l - th element of z is computed by:

$$z_l = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_l(i, j) \tag{1}$$

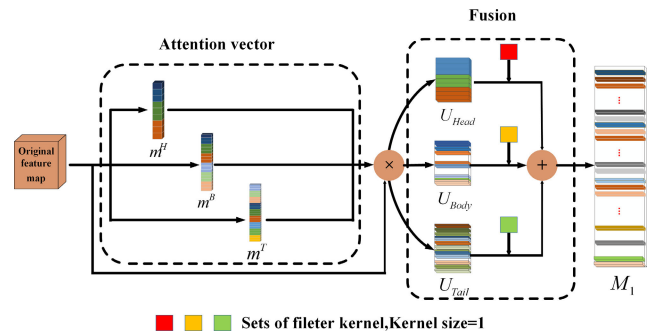


FIGURE 2. Overview of MCAF module. Firstly, we using three different reduction ratios obtain three different vectors m^H , m^B , and m^T with attention from one original feature map, then get three different feature maps using these three different vectors. Then, performing a 1×1 convolution operation on each feature maps and finally get one complete feature map M_1 .

where H represents the height of feature map, W represents the width of the feature map.

Next, we design a multi-reduction ratio excitation mechanism inspired by Senet [41]. we put z into the following formula for calculation and get a one-dimensional vector:

$$m = \sigma(W_2 \delta(W_1 z)) = [m_1 \ m_2 \ \dots \ m_c] \in \mathbb{R}^c \tag{2}$$

where δ and σ represents ReLU function and Sigmoid function, respectively, $W_1 \in \mathbb{R}_r^c \times c$ and $W_2 \in \mathbb{R}^{c \times \frac{c}{r}}$, here r represents the reduction ratio.

The parameter W is used to generate weights for each feature map’s channels, where the parameter W is learned to explicitly model the correlation between feature map’s channels. The reduction ratio r , which is related to parameter W , is an import parameter to determine the correlation

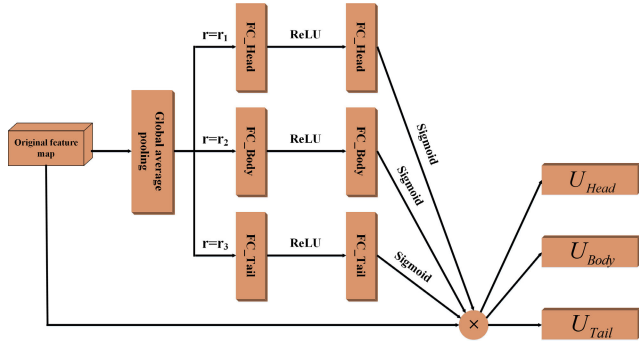


FIGURE 3. Overview of multi-channel attention (MCA) mechanism. We use three reduction ratios r_1 , r_2 , and r_3 respectively, to get the diversity of channels. U_{Head} , U_{Body} , and U_{Tail} represent the head, body, and tail part feature map of MCA output.

between channels. Using different reduction ratios r , different channel-wise weights can be obtained. As shown in Fig. 2, we use three different reduction ratios to get attention vectors with independent response information. and then feature map x can be re-assigned to multi-scale channel-wise attention feature map, respectively.

In the (2), we have the attention vector m^s . The re-weight feature map $U \in \mathbb{R}^{c \times h \times w}$ is obtained by re-calibrating the original feature map $x \in \mathbb{R}^{c \times h \times w}$:

$$U = [m_1 x_1 \quad m_2 x_2 \quad \dots \quad m_c x_c] \in \mathbb{R}^{c \times h \times w} \quad (3)$$

In order to seek a richer feature map channels attention information, we use multi-scale reduction ratio to get channel-wise as much as possible.

Here, we set three different reduction ratios, as shown in Fig.3. Using three different reduction ratios, an original feature map can form three feature maps. Because they will be concatenated as one complete feature map, we named them as head part feature map, body part feature map, and tail part feature map, respectively.

When $r = r_1$, we get excitation vector written as m^H :

$$m^H = \sigma \left(W_2^H \delta \left(W_1^H z \right) \right) = [m_1^H \quad m_2^H \quad \dots \quad m_c^H] \in \mathbb{R}^c \quad (4)$$

where $W_1 \in \mathbb{R}^{\frac{c}{r_1} \times c}$ and $W_2 \in \mathbb{R}^{c \times \frac{c}{r_1}}$, we get the re-weight feature map U_{Head} in (5)

$$U_{Head} = [m_1^H x_1 \quad m_2^H x_2 \quad \dots \quad m_c^H x_c] \in \mathbb{R}^{c \times h \times w} \quad (5)$$

We rewrite it as follow:

$$U_{Head} = [u_1^H \quad u_2^H \quad \dots \quad u_c^H] \in \mathbb{R}^{c \times h \times w} \quad (6)$$

When $r = r_2$, we get the original feature map can be re-weight as feature map U_{Body} in (7):

$$U_{Body} = [m_1^B x_1 \quad m_2^B x_2 \quad \dots \quad m_c^B x_c] \in \mathbb{R}^{c \times h \times w} \quad (7)$$

We rewrite it as follow:

$$U_{Body} = [u_1^B \quad u_2^B \quad \dots \quad u_c^B] \in \mathbb{R}^{c \times h \times w} \quad (8)$$

When $r = r_3$, we get last re-weight feature map U_{Tail} :

$$U_{Tail} = [m_1^T x_1 \quad m_2^T x_2 \quad \dots \quad m_c^T x_c] \in \mathbb{R}^{c \times h \times w} \quad (9)$$

We rewrite it as follow:

$$U_{Tail} = [u_1^T \quad u_2^T \quad \dots \quad u_c^T] \in \mathbb{R}^{c \times h \times w} \quad (10)$$

Thus, we get the re-weight feature maps U_{Head} , U_{Body} , and U_{Tail} with multi-channel attention information using three reduction ratios from one original feature map. In order to further obtain the feature maps diversity, we perform a 1×1 convolution operation on U_{Head} , U_{Body} , and U_{Tail} , respectively.

For the head part feature map U_{Head} , we perform a 1×1 convolution operation on it, the mathematical derivation process is as follows:

Let $V = [v_1 \dots v_l \dots v_i] \in \mathbb{R}^{i \times c \times 1 \times 1}$ represent the learned set of filter kernels, where $v_l \in \mathbb{R}^{c \times 1 \times 1}$ refers to the l -th filter vector. Here, c represents the channel number of feature map U_{Head} . Thus, the head part output feature map of our MCAF module can be written as $M_H = [m_1 \dots m_l \dots m_i] \in \mathbb{R}^{i \times h \times w}$, where $m_l = v_l * U_{Head}$.

Here $*$ denotes convolution, and we set the number of filter kernels i equal to the number of channels c of feature map U_{Head} . Thus, the head part feature map $M_H \in \mathbb{R}^{c \times h \times w}$ of MCAF module gets the correlation between the channels.

Similarly, we get the tail part feature map $M_T \in \mathbb{R}^{c \times h \times w}$ of MCAF module.

Different from $M_H \in \mathbb{R}^{c \times h \times w}$ and $M_T \in \mathbb{R}^{c \times h \times w}$, we let $V = [v_1 \dots v_c \dots v_{2c}] \in \mathbb{R}^{2c \times c \times 1 \times 1}$, the body part feature map of MCAF module can be written as:

$$M_B = [m'_1 \quad \dots \quad m'_c \quad \dots \quad m'_{2c}] \in \mathbb{R}^{2c \times h \times w} \quad (11)$$

Thus, we can get the complete output of our MCAF module from one original feature map by (12):

$$M_1 = \text{concat} (M_H, \quad M_B, \quad M_T) \in \mathbb{R}^{4c \times h \times w} \quad (12)$$

Through above operation, we obtained a feature map M_1 with channel attention by inputting an original feature map. We use the same operations for another two original feature maps as above and obtained feature map M_2 and M_3 . As shown in Fig. 1, we have selected three output feature map of basic network convolutional layers, and used the MCAF module to obtain three feature maps with channel attention. So we have three re-weight feature maps M_1 , M_2 , and M_3 , respectively.

2) CEA MODULE

Bilinear convolution neural networks (B-CNNs) have been applied to many visual tasks, such as visual question-answering (VQA), segmentation, and fine-grained image recognition. Factoring bilinear pooling using Hadamard product is an effective way to improve visual task. Yu et al. [18] developed hierarchical B-CNNs to get fine part features. Here, we consider cross-layer bilinear pooling to be an attention enhancing method for the elements of feature map, and regard it as one part of TPA-CNN. We describe the factoring bilinear pooling using Hadamard product for feature map elements making a choice of weights.

The B-CNNs model performs feature fusion utilizing the last layer of output from two identical convolutional neural networks. Here we have a convolutional network output feature map $X \in \mathbb{R}^{c \times h \times w}$ with channel c , height h , and width w . Feature map $Y \in \mathbb{R}^{c \times (h \times w)}$ can be described as $X = [x_1 \ x_2 \ \dots \ x_c]^T \in \mathbb{R}^{c \times (h \times w)}$, and $Y \in \mathbb{R}^{c \times (h \times w)}$ can be described as $Y = [y_1 \ y_2 \ \dots \ y_c]^T \in \mathbb{R}^{c \times (h \times w)}$. The B-CNNs model can be defined by:

$$f_i = X^T W_i Y \quad (13)$$

where f_i is the output of bilinear pooling model and $W_i \in \mathbb{R}^{c \times c}$ is a weight matrix. In order to get the output $F = [f_1 \ f_2 \ \dots \ f_i \ \dots \ f_o]^T \in \mathbb{R}^o$ of model, where o represents the number of image categories, we need to get the value of $W = [W_1 \ W_2 \ \dots \ W_o]^T \in \mathbb{R}^{c \times c \times o}$. In order to reduce the rank of W_i , according to matrix factorization, the weight matrix W_i can be factorized into two one-rank vectors, the output of B-CNNs model can be rewritten as:

$$f_i = X^T W_i Y = X^T U_i V_i^T Y = I^T (U_i^T X \circ V_i^T Y) \quad (14)$$

where $I \in \mathbb{R}^d$ is a column vector of ones, $U \in \mathbb{R}^{c \times d}$ and $V \in \mathbb{R}^{c \times d}$ are projection matrices. Meanwhile, d is a hyper-parameter deciding the dimension of features and \circ is Hadamard product. The output vector $F \in \mathbb{R}^o$ can be written as:

$$F = P^T (U^T X \circ V^T X) \quad (15)$$

where we replace $I \in \mathbb{R}^d$ with $P \in \mathbb{R}^{d \times o}$, which is the classification matrix. In our method, we let $U \in \mathbb{R}^{c \times d}$ and $V \in \mathbb{R}^{c \times d}$ be identity matrices, then the output vector can be written as:

$$F = P^T (X \circ X) \quad (16)$$

We can see from (16), the feature map X has a Hadamard product with itself to perform element reweighting, which is the self-element attention (SEA) method. SEA method reweights elements using single feature map while completely ignoring the interaction information between cross-layer feature map [18]. We selected two different feature maps X and Y which come from different convolution layer and own the same dimension. Through the above mathematical reasoning, the cross-layer bilinear pooling can be regarded as the product of corresponding positions elements of two different feature maps X and Y . After the Hadamard product between two feature maps coming from different layers, the weights of the element with the larger value will be increased, and the weights of the element with the smaller value will be weakened. Compared with SEA module, CEA module has much more sensitive to dig more information about feature map elements.

In the 1) part, we have three re-weighted feature maps M_1 , M_2 , and M_3 , respectively. And we combine MCAF module and CEA module. Let M_1 , M_2 , and M_3 have a Hadamard

product in any two of them, there are three options. The final output vector $Z \in \mathbb{R}^o$ of TPA-CNN can be obtained by (17):

$$Z = P^T \text{concat} \left(U^T M_1 \circ V^T M_2, U^T M_1 \circ V^T M_3, U^T M_2 \circ V^T M_3 \right) \in \mathbb{R}^o \quad (17)$$

According to the (16), the (17) can be simplified as follows:

$$Z = P^T \text{concat} (M_1 \circ M_2, M_1 \circ M_3, M_2 \circ M_3) \in \mathbb{R}^o \quad (18)$$

where P is the classification matrix, M_1 , M_2 , and M_3 are three re-weighted feature maps through MCAF module which have the same dimension.

In the 1) part, we designed a module to dig more information concerning channel. In this part, taking inspiration from work in [16] and [31], we proposed CEA module and verified it through mathematical derivation. The CEA module aims to set multiple weights for feature map elements through the interaction of cross-layer elements. Therefore, by connecting the MCAF module in front of the CEA module, our model use channel by pixel attention mechanism to locate the discriminative parts.

C. LOSS FUNCTION

Center loss is designed for face recognition. Inspired by face recognition, we introduce it into our model as an auxiliary loss function for fine-grained recognition. For fine-grained image recognition, one of the most difficult is that there are large differences in inter-class and small differences in intra-class. In order to solve this problem, at the end of our network model, we design two parallel fully connected layers, one fully connected layer is designed for the Softmax loss function, and another fully connected layer is design for center loss function [42]. We utilize a joint loss function to minimize the distance in the intra-class, and maximum the distances in the inter-class of the deep features.

Thus, our joint loss function can be written as:

$$L = L_s + \lambda L_c \quad (19)$$

where L_s represents Softmax loss function, L_c donates center loss function, and we set the $\lambda = 0.3$.

IV. EXPERIMENTS

A. DATASET OVERVIEWS

We evaluate TPA-CNN on three widely used datasets for fine-grained image recognition, including Caltech-UCSD Birds (CUB-200-2011) [19], FGVC-Aircraft [47], and Stanford Cars [48]. The details of the three datasets are shown in Table 2, and we use the same data splits displayed in the table.

B. IMPLEMENTATION DETAILS

We remove the fully-connected layers from the basic network, such as ResNet-50 and wide-ResNet-50, and add our design layers to make up our model. We firstly train the layers we designed and then fine-tune the whole model to get the

TABLE 1. Comparison of results on CUB-200-2011, FGVC-Aircraft, and Stanford cars datasets.

Method	Anno.	CUB-200-2011	Cars datasets	FGVC-Aircraft
		Accuracy	Accuracy	Accuracy
SPDA-CNN [43]	✓	85.1%	-	-
DeepLAC [44]	✓	80.3%	-	-
MasK-CNN [11]	✓	87.3%	-	-
HS-net [13]	✓	87.5%	93.9%	-
PDFS [16]	×	84.5%	-	-
GMNet [45]	×	86.5%	93.5%	90.5%
DFL-CNN [28]	×	87.4%	93.8%	91.7%
B-CNNs [46]	×	85.1%	90.6%	86.9%
CBP [17]	×	84.0%	-	-
LRBP [30]	×	84.2%	90.9%	87.3%
HBP [18]	×	87.1%	93.7%	90.3%
FCAN [38]	×	84.5%	89.1%	-
RA-CNN [39]	×	85.3%	92.5%	88.2%
MA-CNN [15]	×	86.5%	92.8%	89.9%
TPA-CNN (wide-ResNet-50)	×	87.8%	94.1%	91.3%
TPA-CNN (ResNet-50)	×	88.0%	94.0%	91.7%

TABLE 2. Statistic of three public datasets.

Dataset	Images	Classes	Train	Test
CUB-200-2011 [19]	11788	200	5994	5794
FGVC-Aircraft [47]	10000	100	6667	3333
Stanford Cars [48]	16185	196	8144	8041

best parameters for our TPA-CNN. In all our experiments, the input images are resized to $600 * 600$ and cropped to $448 * 448$ randomly, then we flip the image horizontally with a probability of 0.5. When testing the model, we just resized the image to $600 * 600$ and cropped to $448 * 448$ by center. We use the standard stochastic gradient descent method. We train on each dataset for 200 epochs; the batch size is set as 16, the weight decay is set as $1 * 10^{-5}$ and the momentum is set as 0.9. The learning rate is initially set to be 1.0 when training the layers we designed while ResNet-50's parameters are fixed. We set the learning rate is 0.01 when fine-tune the whole model. Then, learning rate reduced by a factor of 10 every 50 epochs. We implemented our experiment using Torch. When we trained the layers we designed and fine-tune the whole model, we spend 24 hours, 28 hours, and 22 hours on the CUB-200-2011, FGVC-Aircraft, and Stanford Cars datasets, respectively.

C. QUANTITATIVE EVALUATION

We evaluate TPA-CNN on the CUB-200-2011, FGVC-Aircraft, and Stanford Cars datasets, respectively. The results are displayed in Table 1. It is observed that compared with those state-of-the-art method, our method achieved the competitive result. In Table 1, from top to bottom, those methods can divide into two groups according to whether supervised bounding box or part annotation is used.

TABLE 3. Effect of attention module.

Attention module	CUB-200-2011	Stanford Cars	FGVC-Aircraft
Base: FT ResNet-50	83.3%	92.7%	90.3%
Base+SEA	85.4%	91.4%	90.8%
Base+CEA	87.4%	93.3%	91.1%
Base+CEA+MCAF	88.0%	94.0%	91.7%

Note:

CEA: conv5_1*conv5_1+conv5_2*conv5_2+conv5_3*conv5_3

SEA: conv5_1*conv5_2+conv5_1*conv5_3+conv5_2*conv5_3

Here, Anno. represents whether the model uses the supervised bounding box or part annotation. We compared our model with two types of baselines based on whether they use bounding box or part annotation or not. When we train our model, we only use the weakly supervised image label information. The models which only use weakly supervised image label can be divided into three categories, (1) part feature extracting and learning; (2) models based on B-CNNs; and (3) methods based on attention mechanism. We take two convolutional networks as backbone network, compared our model based on ResNet-50 with those state-of-the-art method stated in Table 1. In CUB-200-2011 dataset, Mask-CNN [11] uses the supervision with part annotation. B-CNNs [46], DeepLAC [44], SPDA-CNN [43] and HS-net [13] using bounding box and part annotation, compared with those models using strong supervision information, we get 0.5% higher than the HS-net [13] model, which is the best performing model between these models. Compared with weak supervision methods, such as PDFS [16], GMNet [45], and DFL-CNN [28] using part feature extracting and learning, our model gets 3.5%, 1.5%, and 0.6% higher accuracy than these models, respectively. Our work is based on the idea of

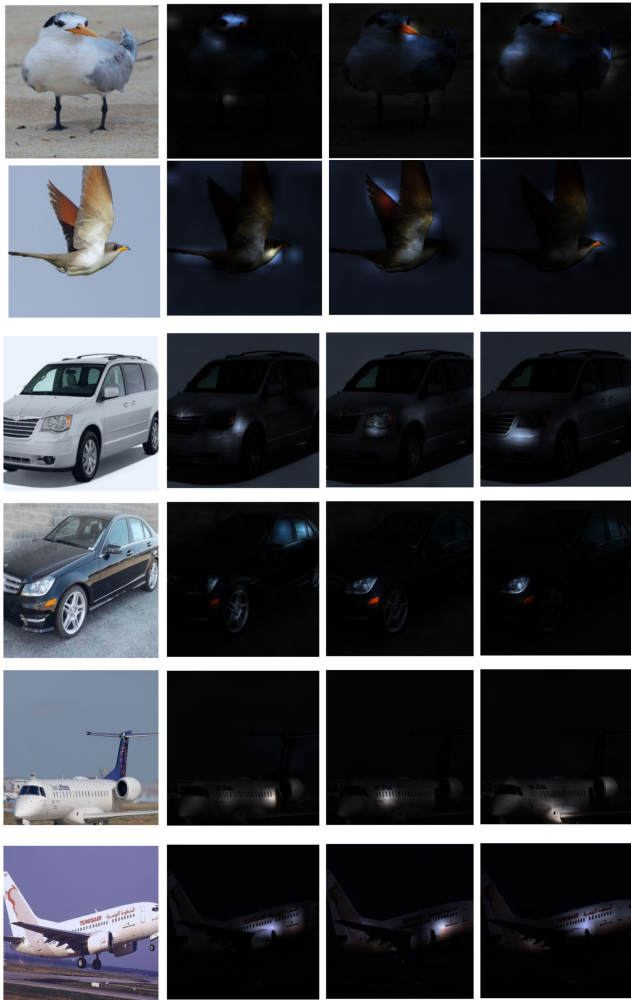


FIGURE 4. Visualization of the model output on three datasets. The first column is the original input image, the second column to fourth column are the visualization of the three feature maps generated in (18) before they concatenating.

attention mechanism, compared with works based on visual attention mechanism, such as RA-CNN [39] and MA-CNN [15], our model gets 2.7% and 1.5% higher accuracy respectively. Compared with works such as CBP [17], LRBP [30], and HBP [18], our work gets 4%, 3.8%, and 0.9% higher accuracy, respectively.

We also evaluate our model on the Stanford Cars [48] and FGVC-Aircraft [47]. Compared with state-of-the-art method in Table 1, TPA-CNN get 0.1% higher than HS-net [13] which has the highest accuracy on the Stanford Cars datasets and get the same accuracy as the DFL-CNN [28] which has the highest accuracy on FGVC-Aircraft dataset.

D. ABLATION STUDY

In this section, we provide an analysis and ablation experiment to understand the components and variants of our proposed TPA-CNN on three validation set.

1) AVAILABILITY OF ATTENTION MODULE

In order to verify the availability of TPA-CNN, we selected three datasets to evaluate the attention module. In Table 3,

TABLE 4. Effect of layers on cub200-2011 dataset.

Method	CEA1	CEA2	CEA3	CEA4	CEA (ours)
Accuracy	76.1%	82.9%	84.4%	83.4%	88.0%

Note:

CEA1: conv4_0*conv4_1+conv4_0*conv4_2+conv4_1*conv4_2
 CEA2: conv4_2*conv4_3+conv4_2*conv4_4+conv4_3*conv4_4
 CEA3: conv4_3*conv4_4+conv4_3*conv4_5+conv4_4*conv4_5
 CEA4: conv4_0*conv4_3+conv4_0*conv4_5+conv4_3*conv4_5
 CEA (ours): conv5_1*conv5_1+conv5_2*conv5_2+conv5_3*conv5_3

TABLE 5. Effect of MCAF module parameter *r* setting on cub200-2011 dataset.

Reduction ratio	Accuracy
<i>r</i> = 16	87.7%
<i>r</i> = 32	87.8%
<i>r</i> = 64	87.8%
<i>r</i> = mix(16,32,64)	88.0%

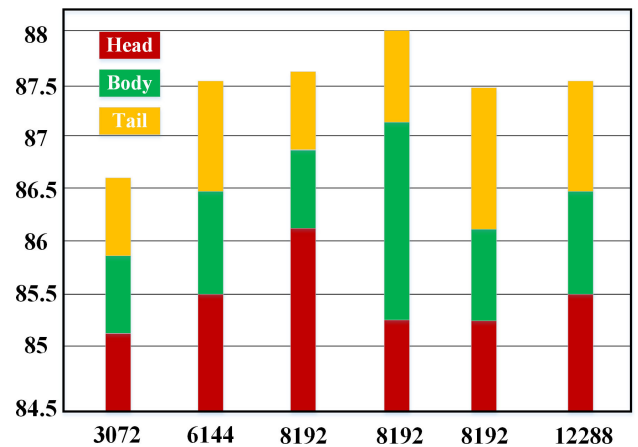


FIGURE 5. Overview of the effect of parameter number filter setting on cub200-2011 dataset.

we provide a detailed analysis on different key module setting. Given a base network, we investigate the contribution of each module. Taking ResNet-50 as the backbone network, after fine-tuning, ResNet-50 achieved 83.3%, 92.7%, and 90.3% accuracy on the CUB-200-2011, Stanford Cars datasets, and FGVC-Aircraft, respectively. Our CEA module is designed based on B-CNNs, and the CEA module is designed for the element of the feature map. Before re-weighting the element of the feature map, we added a channel re-weighting module that is MCAF module to form our two-level progressive attention model. Firstly, we compared the SEA module with CEA module, the SEA and CEA are indicated in Table 3 comments. We compared the impact of each module on the classification accuracy. we only added the SEA module on backbone network, and get 2.1% and 0.5% higher accuracy on CUB-200-2011 and FGVC-Aircraft dataset but 1.3% lower accuracy on Stanford Cars. When we replaced the SEA module with the CEA module, the effect was 2 %, 1.9%, and 0.3% higher than ResNet-50 with SEA module on three datasets. when we add the MCAF module

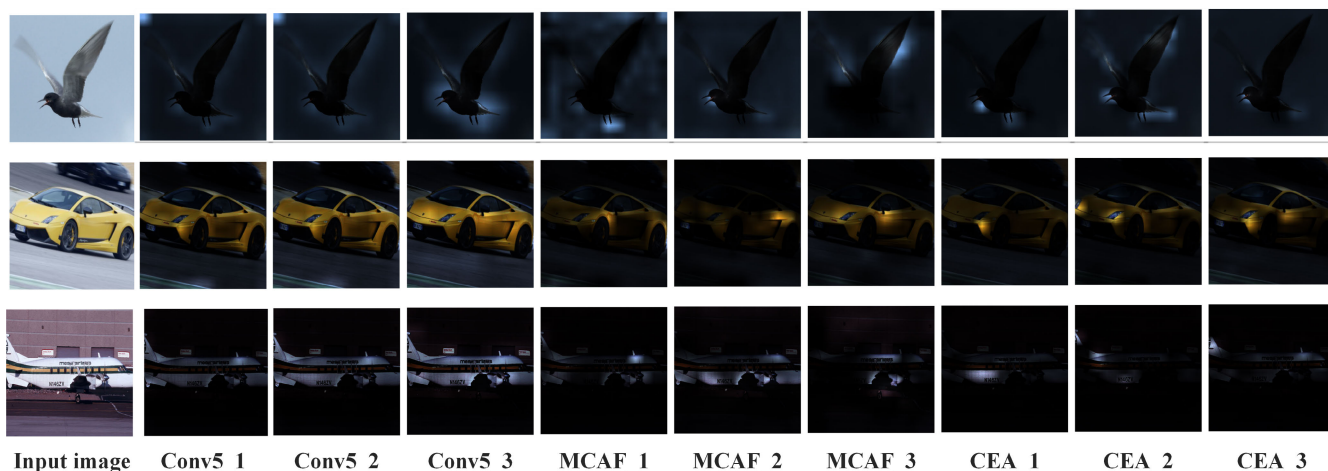


FIGURE 6. Visualization of the model’s response to different convolutional outputs on CUB-200-2011, Stanford Cars and FGVC-Aircraft datasets respectively. The first column is the original image, the second to fourth columns are response of the basic network convolution, the fifth to seventh columns are columns are the response of the convolution through the MCAF module, and the last three columns are the response of the convolution through the CEA module.

on ResNet-50 with using CEA module, and we achieved a 4.7 %, 1.3%, and 1.4% higher accuracy than fine-tuned base network.

Table 3 manifests the two attention modules can effectively improve the fine-grained recognition accuracy. To demonstrate the location and number of regions which the model final activated, we further visualize the final convolution output of the model. As shown in Fig. 4, the left column is the original image, and the right three columns are the visualization of the three feature maps generated in (18) before they concatenating. Here we randomly selected two images from three datasets, respectively. For the birds, the model randomly activates the head, mouth, feet or other parts. For the cars, the lights, wheels, and so on, are activated. For the aircrafts, the engine room and wings are activated. From the above analysis, it can be clearly found that our model can search one or more unrelated discriminative regions without setting the location and number of discriminant regions artificially, which are obtained by the model dynamically.

2) EFFECT OF CEA MODULE PARAMETERS SETTING

In our model, we select the ResNet-50’s conv5_1, conv5_2, and conv5_3 as the original input feature maps in Fig 1. In order to investigate different layers performance, we choose other layers in ResNet-50, such as conv4_0~conv4_5 which have same dimensional. The results are shown in the Table 4, If we use the feature maps conv4_0, conv4_1, and conv4_2, the accuracy only can reach 76.1%. When use a little high-level feature maps, the accuracy becomes higher gradually. The results demonstrate that the performance gain of our designed framework come from the high-level features such as conv5_1~conv5_3. Since our model is inspired by B-CNNs which utilize the high-level features of the final output of the convolutional network, the middle-level features have a very little help for fine-feature learning. Our work is a further design on B-CNNs. The experience in Table 4 has proved our model

utilizes the high-level features to make the accuracy of fine-grained image recognition higher.

3) EFFECT OF MCAF MODULE PARAMETERS

In the Table 3, we test the availability of attention module and compare the effect of the SEA module and the CEA module. For the MCAF module, both the parameters r in channel-wise attention operation and number of filter kernels determining the output feature map channel in channel fusion operation have a certain effect on the accuracy for fine-grained recognition. Therefore, in order to find the best hyper-parameters, we conduct multiple sets of comparative experiments on CUB-200-2011.

a: EFFECT OF PARAMETERS R

In section III, we use (4)-(10) to elaborate three different reduction ratios r . In Table 5, when other parameters of TPA-CNN were fixed, the results display the effect of the three different reduction ratios r . Through comparison experiments, the mix reduction ratios r get 0.2% or 0.3% higher accuracy than just use one reduction ratio. Experiments prove that using three different reduction ratios r in MCAF module can achieve the best result.

b: EFFECT OF PARAMETERS NUMBER FILTER

In section III, we use (11) to illustrate the channel fusion. In Fig. 5, in order to evaluate the influence of number of filter kernels, we fixed the parameter r , and set $r = 16, 32, 64$ for M_H, M_B , and M_T , respectively.

In Fig. 5, the length ratio of the lines represents the distribution the number of filter kernels determining the output feature map channel. When performing $1 * 1$ kernel filter for convolution in channel fusion operation, we use different number of filter kernels for U_{Head}, U_{Body} , and U_{Tail} . We can see from Fig. 5, when the number of filter kernels is halved or double number of input feature map channel, their accuracies are reduced. When we set the number i of filter kernels for

M_H , M_B , and M_T are 2048, 4096, and 2048, respectively, the model reaches the optimal performance.

4) VISUALIZATION AND ANALYSIS

To better observe the model, we visualized the model's response to different convolutional outputs on the CUB-200-2011, Stanford Cars, and FGVC-Aircraft in Fig. 6, respectively. We obtain the activation response graph by superimposing the average value of the output of the convolution layer with the original image. We randomly select three images from that three datasets respectively.

The visualization shows that output of the three convolutional layers after ResNet-50's conv5_1, conv5_2, and conv5_3. It can be clearly observed that in the original convolutional layers' visualization, the discriminative areas even the foreground areas are not activated or only locate extremely few parts. After through the MCAF module, the visualization of feature maps demonstrates that MCAF module can ignore the complex and messy background region and effectively focus on the foreground region. Finally, the visualization of feature maps after through the CEA module demonstrates that our model can find one or several discriminative regions before softmax loss function classification. And we can also find that the different convolutional layers activate the same or different discriminative regions. From left to right in Fig. 6, it can be seen that our module gradually finds the discriminative regions and increases the diversity of attention areas spreading over object foreground areas. Our model uses a two-level attention module to gradually focus on the most discriminative areas from coarse-grained to fine-grained can effectively improve the accuracy of recognition.

V. CONCLUSION

For fine-grained image recognition, the discriminative feature is crucial but also is subtle and regional. In order to better locate and extract correct region features for fine-grained image recognition, in this study, we propose a two-level progressive attention convolutional network (TPA-CNN) without using bounding box or part annotations. Our model includes MCAF module and CEA module. Firstly, the MCAF module obtains a re-weight feature map group by making a channel-wise and channel fusion. Then, the CEA module re-weights the elements of the feature map we obtained in the first step. Our model is a two-level progressive model which can effectively focus on the discriminative areas from coarse-grained to fine-grained. We have achieved competitive results on the three popular datasets.

However, there are still some issue worthy of further discussion in our research. We only rethought about two important dimensions of the feature map: channel and element, but feature map still has other information to mine. In future we will attempt to exploit more information regarding feature maps which may have better performance for fine-grained image recognition.

REFERENCES

- [1] J.-H. Luo and J.-X. Wu, "A survey on fine-grained image categorization using deep convolutional features," *Acta Autom. Sin.*, vol. 43, no. 8, pp. 1306–1318, 2017.
- [2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [4] W. Cao, J. Yuan, Z. He, Z. Zhang, and Z. He, "Fast deep neural networks with knowledge guided training and predicted regions of interests for real-time video object detection," *IEEE Access*, vol. 6, pp. 8990–8999, 2018.
- [5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [6] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional Nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [7] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2014, pp. 834–849.
- [8] Z. Ma, D. Chang, J. Xie, Y. Ding, S. Wen, X. Li, Z. Si, and J. Guo, "Fine-grained vehicle classification with channel max pooling modified CNNs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3224–3233, Apr. 2019.
- [9] R. Li, W. Lu, H. Liang, Y. Mao, and X. Wang, "Multiple features with extreme learning machines for clothing image recognition," *IEEE Access*, vol. 6, pp. 36283–36294, 2018.
- [10] Y. Rivenson, Y. Wu, and A. Ozcan, "Deep learning in holography and coherent imaging," *Light, Sci. Appl.*, vol. 8, no. 1, p. 85, Sep. 2019.
- [11] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, Apr. 2018.
- [12] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Augmenting strong supervision using Web data for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2524–2532.
- [13] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as HSnet search for informative image parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2520–2529.
- [14] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 805–821.
- [15] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5217.
- [16] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1134–1142.
- [17] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 317–326.
- [18] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 574–589.
- [19] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Dept. Comput. Neural Syst., California Inst. Technol., Pasadena, CA, USA, Tech. Rep. TR-200-2011, 2011.
- [20] T. Berg and P. N. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 955–962.
- [21] X.-S. Wei, J. Wu, and Q. Cui, "Deep learning for fine-grained image analysis: A survey," 2019, *arXiv:1907.03069*. [Online]. Available: <http://arxiv.org/abs/1907.03069>
- [22] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.
- [23] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1153–1162.

- [24] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2399–2406.
- [25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
- [26] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4683–4695, 2020.
- [27] R. Du, D. Chang, A. K. Bhunia, J. Xie, Y.-Z. Song, Z. Ma, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," 2020, *arXiv:2003.03836*. [Online]. Available: <https://arxiv.org/abs/2003.03836>
- [28] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.
- [29] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 420–435.
- [30] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 365–374.
- [31] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6510–6519.
- [32] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 947–955.
- [33] T.-Y. Lin, S. Maji, and P. Koniusz, "Second-order democratic aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 620–636.
- [34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [35] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 2048–2057.
- [36] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.
- [37] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," 2014, *arXiv:1412.7054*. [Online]. Available: <http://arxiv.org/abs/1412.7054>
- [38] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," 2016, *arXiv:1603.06765*. [Online]. Available: <http://arxiv.org/abs/1603.06765>
- [39] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4438–4446.
- [40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
- [41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [42] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 499–515.
- [43] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.
- [44] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1666–1674.
- [45] J. Liang, J. Guo, X. Liu, and S. Lao, "Fine-grained image classification with Gaussian mixture layer," *IEEE Access*, vol. 6, pp. 53356–53367, 2018.
- [46] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.
- [47] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*. [Online]. Available: <http://arxiv.org/abs/1306.5151>
- [48] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.



HUA WEI received the B.S. degree in automation from Shandong University. She is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include object detection and fine-grained recognition.



MING ZHU is currently a Research Fellow and a Ph.D. Supervisor with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include digital image processing, television tracking, and automatic target recognition technology.



BO WANG received the B.S. degree in microelectronics from Jilin University, in 2016. He is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. His research interests include object detection and 3D object detection.



JIARONG WANG was born in Changchun, Jilin, China, in 1989. She received the B.S. degree in optical engineering from the Changchun University of Science and Technology, and the M.S. degree in circuits and systems from Jilin University. She is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include 2D and 3D object detections and stereo vision.



DEYAO SUN received the B.E. degree in testing instrument and technology from the Harbin Institute of Technology, in 2014. He is currently pursuing the Ph.D. degree in mechatronics engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. His research interests include object detection in remote sensing images and 3-Dimension images, scene understanding, and semantic segmentation.

• • •