Mixed Attention Mechanism for Small-Sample Fine-grained Image Classification

Xiaoxu Li*, Jijie Wu*, Dongliang Chang†, Weifeng Huang‡, Zhanyu Ma† and Jie Cao*
* Lanzhou University of Technology, Lanzhou, China
E-mail: xiaoxulilut@gmail.com
† Beijing University of Posts and Telecommunications, Beijing, China
E-mail: mazhanyu@bupt.edu.cn
‡ South-to-North Water Diversion Middle Route Information Technology Co., Ltd., China
E-mail: huangweifeng@nsbd.cn

***Abstract*—Fine-grained image Classification is an important task in computer vision. The main challenge of the task are that intra-class similarity is large and that training data points in each class are insufficient for training a deep neural network. Intuitively, if we can learn more discriminative features and more detailed features from fined-grained images, the classification performance can be improved. Considering that channel attention can learn more discriminative features, spatial attention can learn more detailed features, this paper proposes a new spatial attention mechanism by modifying Squeeze-and-Excitation block, and a new mixed attention by combining the channel attention and the proposed spatial attention. Experimental results on two small-sample fine-grained image classification datasets demonstrate that on both VGG16 network and ResNet-50 network, the proposed two attention mechanisms achieve good performance, and outperform other referred fine-grained image classification methods.**

I. Introduction

With rapid development of deep learning, Convolutional Neural Networks (CNNs) are widely used in the task of fine-grained image classification which is to distinguish one subordinate categories from others among the same superordinate category [1]. Fine-grained image classification based on CNNs have obtained impressive performance either by replacing hand-crafted features with CNN features or by adopting an end-to-end fashion. However, there still exists big challenges since intra-class similarity is large and training data points in each class are insufficient in fine-grained images [2] [3] [4].

The works of fine-grained image classification based on CNNs mainly focus on learning more subtle and more discriminative features. Some works improved network structure [5], [6], [4], some works proposed a new loss [3], and some works improved fine-grained classification by introducing attention mechanism [7], [8], [9]. The Attention Mechanisms in Neural Networks are derived from the visual attention mechanism found in humans. Human visual attention focuses on a certain region of an image with "high resolution" while perceiving the surrounding image in "low resolution", and then adjusting the focal point over time [10]. Fine-grained classification with attention mechanism could learn more delicate difference than other methods[11].

There are three types of attention mechanisms, e.g. channel attention, such as the SE (Squeeze-and-Excitation) block [12] , the spatial attention, such as the Spatial Transformer [13], and mixed attention, such as two-level Attention Models [8] and Recurrent attention model [10]. The channel attention aims to learn more discriminative features. SE (Squeeze-and-Excitation) block [12] is a classical channel attention method, which focuses on the channel relationship and adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. The spatial attention aims to learn more detailed features. The Spatial Transformer [13], a new learnable and differentiable module, which explicitly allows the spatial manipulation of data and can be inserted into existing convolutional architectures. The method is conditional on the feature map itself and can learn invariance to scale, rotation and so on. Residual Attention Network [14] is built by stacking Attention Modules which generate attention-aware features.

The mixed attention aims to learn more discriminative and more detailed features simultaneously. Two-level Attention Model [8] combines three types of attention: the bottom-up attention, the object-level top-down attention, and the part-level top-down attention, which are responsible for proposing candidate patches, selecting relevant patches to a certain object, and localizing discriminative parts, respectively, to find object parts and extract discriminative features. Recurrent attention model [10] is a recurrent neural network model that is capable of extracting information from an image by adaptively selecting a sequence of regions or locations and only processing the selected regions at high resolution. Compared with convolutional neural networks, the model greatly reduces the amount of computation.

Intuitively, if we can learn more discriminative features and more detailed features from fined-grained images, the classification performance can be improved. Therefore, this paper builds on the existing mixed attention works, proposes a new spatial attention mask by modifying Squeeze-and-Excitation block, and a new mixed attention method by combining the channel attention and the proposed spatial attention. In order to evaluate the proposed two attention methods, we use two widely used networks, VGG16 and ResNet-50, and select two small-sample fine-grained image classification datasets, the Stanford Cars-196 dataset and the FGVC-Aircraft

dataset. Experimental results show that the proposed spatial and mixed attention mechanisms achieve good performance, and outperform other referred fine-grained image classification methods.

II. Proposed the Mixed Attention

Attention mechanism can help the model to assign different weights to different parts of input X, extract more critical and more discriminative information, and make a model do more accurate precision. At the same time, it does not bring additional to the calculation and storage of the model. In this section, we will simply introduce channel attention mechanism, then propose a new spatial attention mechanism by modifying Squeeze-and-Excitation block, and propose our mixed attention mechanism which is based on these two mechanisms.

A. Channel Attention

The channel attention mechanism [12] can be regarded as a computational unit, which adds corresponding weight to each channel in the output features of the convolution network, making the neural network pays more attention to the channel that can improve the classification results. Specifically, referring to Fig 1, in order to efficiently calculate channel weights, [12] fetch an input image through a series of convolution and pooling operation F_{tr} to obtain a feature map marked as $U = [u_1, u_2, ..., u_c]$, which size is $H \times W \times C$, and then uses function F_{sq} and F_{ec} compute the channel-wise weights of U. Here F_{sq} operation generates channel-wise statistic by using global average pooling and F_{ec} operation generates channel-wise attention by using ReLU [15] activation function, two fully connected (FC) layers and Sigmoid activation function. Follow [12], through function F_{scale} can get channel attention feature maps $\widetilde{X}_c = [\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_c]$

$$\widetilde{x}_c = F_{scale}(u_c, s_c) = u_c \cdot s_c, \tag{1}$$

where s_c represents the weight of each c channel.

Generating such weighted feature maps to train network will make the classifier pays more attention to the information, which is useful to the image classification results, and ignore or weaken the irrelevant information.

B. Spatial Attention

Different from channel attention which mainly focuses on what is meaningful in the input image, spatial attention mainly focuses on some details. In order to calculate spatial attention, we also need to use function F_{sq} and F_{et} to generate a group of weights, here the function F_{sq} same with previous section, and the function F_{et} replaces the last activation function of function F_{ec} with Softmax. Thus the sum of the weights generated by F_{ex} is one. Just like the channel attention mechanism, the temp feature maps $\widetilde{X}_t$ can be calculated by:

$$\widetilde{x}_t = F_{scale}(u_c, s_c) = u_c \cdot s_c. \tag{2}$$

Here $\widetilde{X}_t = [\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_C]$, and $s_c = [s_1, s_2, ..., s_C]$ represents a weight of c channel.

After the operation F_{flat} is finished, the feature maps are flattened along channel direction, and the spatial attention weights, V, is obtained. The spatial attention feature maps $\widetilde{X_s}$ are obtained by multiplying the spatial attention weights with U. The function F_{mul} is defined as follows:

$$\widetilde{x_s} = F_{mul}(u_{s,c}, v_s), \tag{3}$$

where s ranges over all spatial positions and v_s represents a weight of one of the spatial positions, $u_{s,c}$ represents the feature on spatial position corresponding to each channel of U.

C. Mixed Attention

Based on the channel attention mechanism and the proposed new spatial attention mechanism, we propose a mixed attention mechanism, which combines two attention modules, e.g. channel attention and spatial attention, and concatenates the feature maps of the two attention modules.

The channel attention focuses on what the images show, and the spatial attention focuses on the spatial information of images, that is, where the objects in the images are. Considering that both channel and spatial attention only learn one aspect of the image, and both of them are crucial to image classification, we propose a new mixed attention mechanism, and the equation of the proposed mechanism is as follow:

$$\widetilde{X}_{cs} = F_{cat}(\widetilde{X}_c, \widetilde{X}_s), \tag{4}$$

where F_{cat} represents concatenation operation of two feature maps, $\widetilde{X}_c$ and $\widetilde{X}_s$. After the operation F_{cat}, we use the convolution operation F_{merge} to get $\widetilde{X}$, a feature map with the size of $H \times W \times C$.

III. Experimental Results

A. Datasets

The Stanford Cars-196 dataset[16] contains 16,185 images of 196 classes of cars, and categories of this dataset are mainly divided based on the brand, model and year of the car. The dataset is split into 8,144 training images and 8,041 testing images, in which each class is split roughly 50-50.

The FGVC-Aircraft dataset[17] contains 10,000 images of aircraft, with 100 images for each of 100 different aircraft model variants, most of which are airplanes. The image resolution is about 1-2 Mpixels. Although images of the planes were taken over a period of decades, their quality is still good. In this paper, the FGVC-Aircraft dataset is split into 6,667 training images and 3,333 testing images. Both datasets are suitable for small-sample fine-grained image classification.

Considering that the influence of feature quality on image classification performance, we resize each image to 224×224, before the images are fed into the neural network, and we normalize the images. Meanwhile, we do not use part or bounding box (BBox) annotations in the above two datasets.

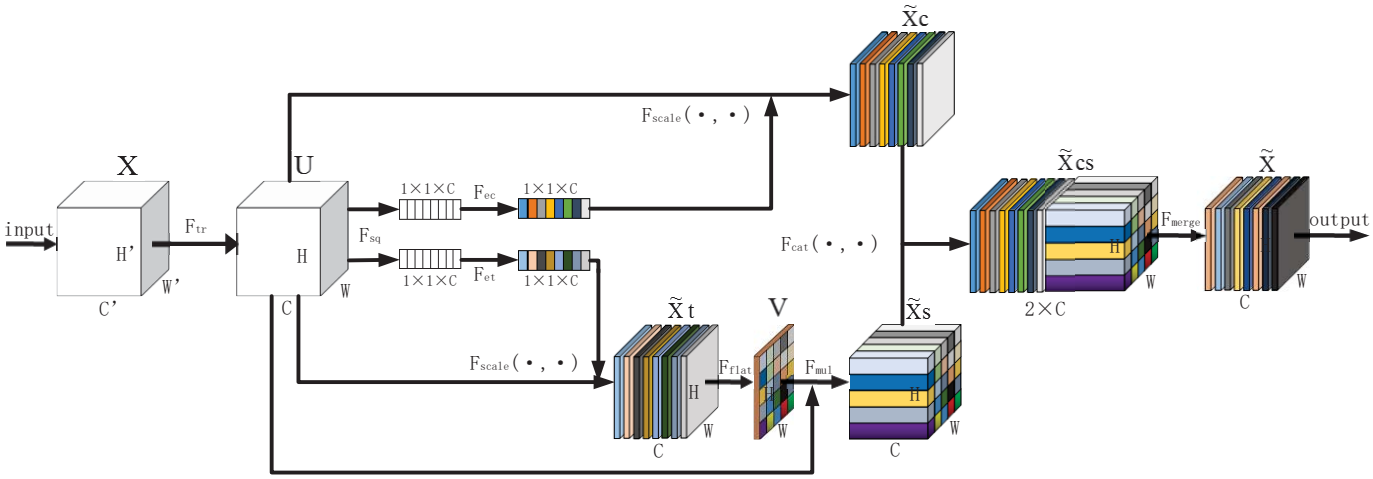

Fig. 1. Model architecture of the our mixed attention mechanism.

TABLE I
Comparison of our methods and recent results on FGVC-Aircraft without extra annotation. SA-Ours: the proposed spatial attention mechanism. MA-Ours: the proposed mixed attention mechanism. The black-bold number represents the best result.

Method	Base Model	Accuracy
BoT(+BBox)[19]	VGG16	88.4
B-CNN[20]	VGG16	84.1
Low-Rank B-CNN[21]	VGG16	87.3
Kernel-Activation[21]	VGG16	88.3
Kernel-Pooling[22]	VGG16	86.9
SA-Ours	VGG16	88.5
MA-Ours	VGG16	**88.6**
Kernel-Pooling[22]	ResNet-50	85.7
SA-Ours	ResNet-50	**87.1**
MA-Ours	ResNet-50	86.7

B. Implementation Details

Our experiment uses the publicly available deep network VGG16 with batch normalization and ResNet-50 pre-trained on the ImageNet dataset. Both VGG16 and ResNet-50 here are publicly available from Pytorch [18].

For pre-trained VGG16, we replaced the classifier layer trained on the ImageNet dataset with a randomly initialized FC layer with a batch normalization layer, and an ELU layer as the activation function. The numbers of the hidden nodes were 512 for the Stanford Cars-196 dataset and the FGVC-Aircraft dataset. A randomly initialized K-way softmax layer was adopted, where K is the number of classes in the dataset.

For pre-trained ResNet-50, we removed the last global average pool layer and FC layer in the ResNet-50 features, and the changes to the classifier are the same as those made on pre-trained VGG16 model.

We embedded channel attention module layer and spatial attention module layer in the basic model respectively, which also served as the comparison schemes of the proposed mechanism. In terms of mixed attention model architecture, we added channel attention module layer and spatial attention module layer in the last layer of modified basic models, and channel attention module, spatial attention module and VGG16 initial learning rates is set to 0.1, 0.1, 0.01 respectively. The learning rate decreases with cosine [23], the batch size is set to 32, and the value of epoch set to 300. The loss function uses the commonly used cross entropy loss. For pre-trained ReNet-50, we also added channel and spatial attention layer, setting the same parameters as those set on VGG16. Each of models is trained on the Stanford Cars-196 dataset and the FGVC-Aircraft dataset.

C. Classification Performance

In order to quantitatively analyze the performance of the classification in this paper, the classification accuracy of test is used as the performance measurement index of the experiment. Table II shows that when trained with VGG16 model and based on the Stanford Cars-196 dataset, the accuracy of baseline, CA, SA-Ours and MA-Ours is 87.99%, 88.75%, 88.61% and 88.89% respectively. And on the FGVC-Aircraft dataset, the accuracy of baseline, CA, SA-Ours and MA-Ours is 86.50%, 86.98%, 88.51% and 88.63 % respectively. Among them, no matter what kind of dataset, the accuracy of our proposed two mechanisms are better than others.

The ResNet-50 model also shows the similar performance. Obviously, both our mixed attention mechanism and our spatial attention mechanism have a higher classification performance than other mechanisms.

According to Fig 2 which is shown the classification accuracy and cross entropy loss on the Stanford Cars-196 dataset trained with ResNet-50 model. When the epoch is greater than 30, the loss curve of our mixed attention is significantly lower than the others, we can see that our test accuracy of mixed attention also began to show a slight advantage. In addition for our spatial attention, both loss and accuracy are slightly better than baseline and channel attention.

Then, the comparisons with the recent results trained with the FGVC-Aircraft dataset are also presented in Table I. Several methods that performed well on this dataset are the BoT(+BBox) method (with accuracy 88.4%) and the B-

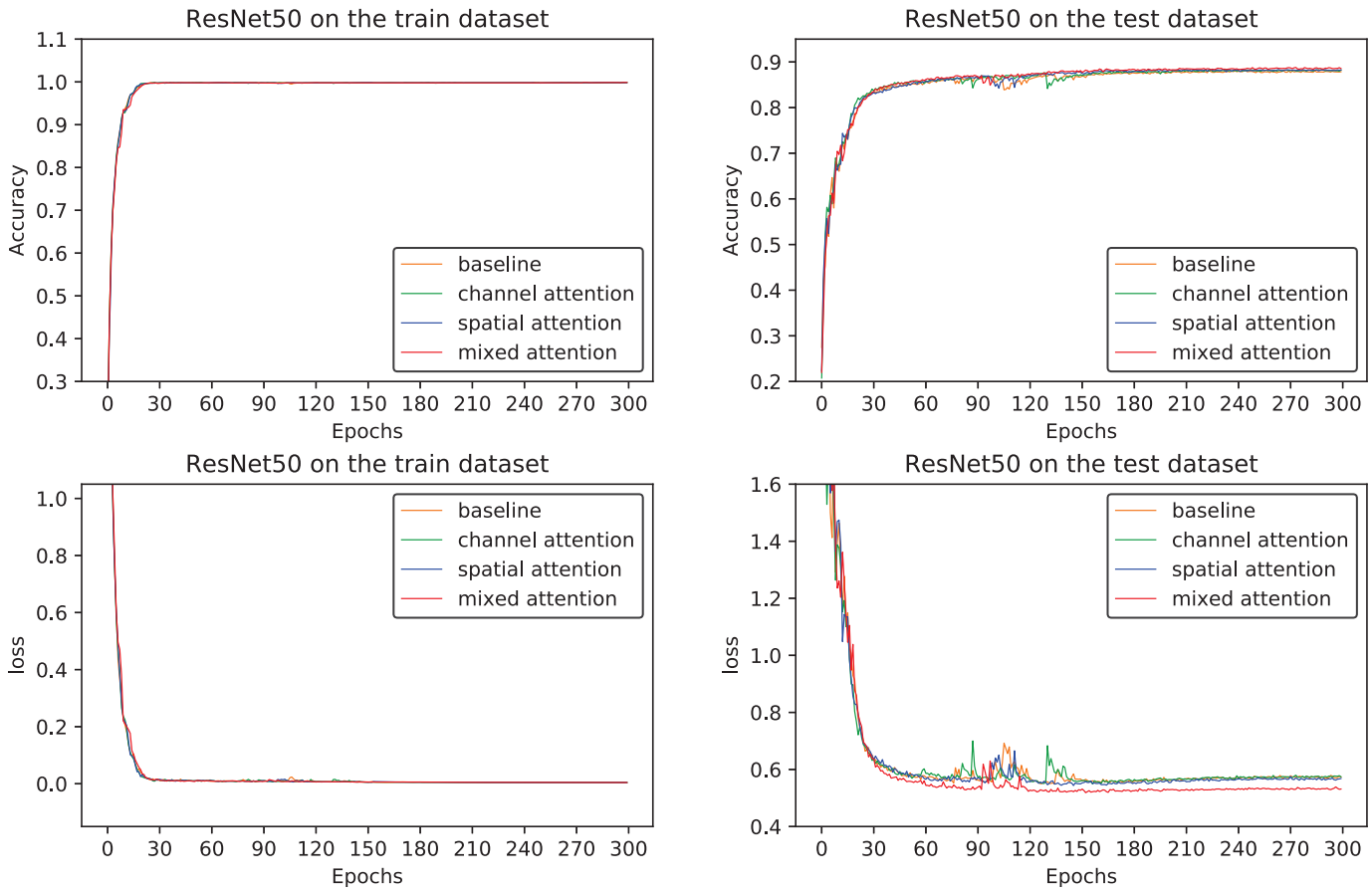

Fig. 2. The loss and accuracy obtained by ResNet-50 with channel attention, ResNet-50 with the proposed spatial attention and ResNet-50 with the proposed mixed attention on the Stanford Cars-196 dataset.

TABLE II
The test accuracy on the Stanford Cars-196 dataset and the FGVC-Aircraft dataset based on the VGG16 and the ResNet-50 model respectively. Base Model: publicly available the VGG16 with batch normalization and ResNet-50 network pre-trained on the ImageNet dataset. Baseline: base model. CA: base model with channel attention mechanism [12]. SA-Ours: base model with the proposed spatial attention mechanism. MA-Ours: base model with the proposed mixed attention mechanism.

Dataset	Stanford Cars-196		FGVC-Aircraft	
Base Model	VGG16	ResNet-50	VGG16	ResNet-50
Baseline	87.9866	88.1358	86.4986	86.8887
CA	88.7452	88.2850	86.9787	86.8287
SA-Ours	88.6084	88.4218	88.5089	**87.1587**
MA-Ours	**88.8944**	**88.7949**	**88.6289**	87.0987

CNN method (with accuracy 84.1%). No matter what kind of base model, our spatial attention mechanism and our mixed attention mechanism all achieve the best accuracy (spatial attention mechanism is 88.5% and mixed attention mechanism is 88.6% based on pre-trained VGG16 model) among all the referred methods.

D. Ablation Study

Since the proposed mechanism is a combination of the output features of channel attention module and our spatial attention module, in order to prove the classification effectiveness of the mechanism we proposed, we simply removed spatial attention module and channel attention module respectively and conducted two groups of experiments as the comparative ablation experiments with our mixed attention mechanism. In other words, the CA experiment refers to the ablation experiment after removing the spatial attention module from our mechanism, the SA-Ours experiment refers to the ablation experiment after removing the channel attention module. And baseline method refers to the base model without any attention mechanism.

As show in Fig 2, for pre-trained VGG16 model, the accuracy curves of both modified channel attention and our spatial attention are slightly lower than those of the our mixed attention. As shown in Table I, on the Stanford Cars-196 dataset, the accuracy of MA-Ours is 88.89%, which is 0.90% higher than baseline and 0.14% higher than CA and 0.28% higher than SA-Ours. On the FGVC-Aircraft dataset, the accuracy of MA-Ours is 88.63%, which is 2.13% higher than baseline and 1.65% higher than CA and 0.12% higher than SA-Ours. The similar improvement could be found when

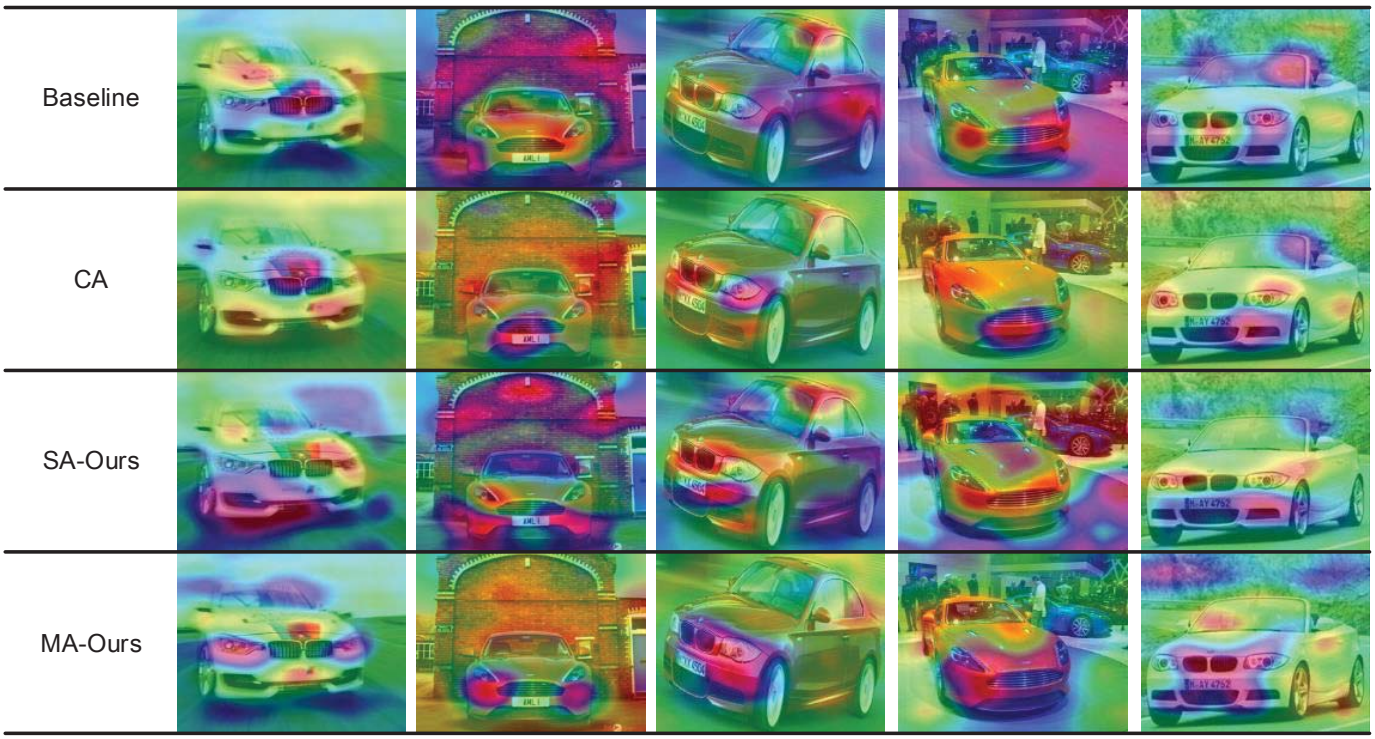

Fig. 3. Visualization of the energy distribution of feature maps.

the pre-trained ResNet-50 model is well.

E. Discussion

From the experimental results on the Stanford Cars-196 dataset, compared with basic VGG16, channel attention and spatial attention mechanism, the proposed mixed attention mechanism has batter performance. During the training process, back propagation of network can influence the feature map obtained by the convolution layer. By drawing the heatmap of the feature map obtained after network training, we can understand which parts of the image play a key role in the image classification, and at the same time locate the position of objects in the image. As shown in Fig 3, compared with heatmaps of other methods, the proposed mechanism can capture more discriminative regions of images.

Hence, in order to determine whether the feature map obtained by network which we proposed is effective in corresponding image classification, it is necessary to draw the heatmap of the high-level part feature. Firstly, we trained modified VGG16 network on the Stanford Cars-196 dataset, and then input images into the trained network to extract the highest-level feature map of VGG16 model for visualization. As can be seen from Fig 3, in the last column, when trained with our mixed attention module, the energy distribution of heatmaps was mostly focused on key parts such as car lights. Therefore, after our mixed attention module training, the energy distribution is refined and has significantly beneficial discrimination on the feature maps.

IV. Conclusions

In this paper, we proposed a new spatial attention and mixed attention mechanism for improving the performance of the fine-grained images classification. The mechanism combined channel attention features and spatial attention features as the whole feature representation of images. Experimental results on the Stanford Cars-196 dataset and the FGVC-Aircraft dataset show the effectiveness of proposed attention mechanism, and confirmed the network with the proposed attention mechanisms can obtain better classification performance.

Acknowledgement

This work was partly supported by the National Key Research and Development Program of China under Grant 2018YFC0807205, the National Natural Science Foundation of China (NSFC) grant No.61563030, No.61763028, No.61773071, No.61922015 and No.61906080, the Natural Science Foundation of Gansu Province, China, grant No.17JR5RA125, the Beijing Nova Program Interdisciplinary Cooperation Project No.Z181100006218137, Beijing Nova Program No.Z171100001117049 and by the Hong-liu Outstanding Youth Talents Foundation of Lanzhou University of Technology.

References

[1] Yaming Wang, Jonghyun Choi, Vlad Morariu, and Larry S Davis. Mining discriminative triplets of patches for fine-grained classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1172, 2016.

[2] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152, 2016.

[3] Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao. Dual cross-entropy loss for small-sample fine-grained vehicle classification. *IEEE Transactions on Vehicular Technology*, 2019.

[4] Shaoyong Yu, Yun Wu, Wei Li, Zhijun Song, and Wenhua Zeng. A model for fine-grained vehicle classification based on deep learning. *Neurocomputing*, 257:97–103, 2017.

[5] Jie Fang, Yu Zhou, Yao Yu, and Sidan Du. Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture. *IEEE Transactions on Intelligent Transportation Systems*, 18(7):1782–1792, 2016.

[6] Krassimir Valev, Arne Schumann, Lars Sommer, and Jurgen Beyerer. A systematic evaluation of recent deep learning architectures for fine-grained vehicle classification. In *Pattern Recognition and Tracking XXIX*, volume 10649, page 1064902. International Society for Optics and Photonics, 2018.

[7] Zhanyu Ma, Dongliang Chang, Jiyang Xie, Yifeng Ding, Shaoguo Wen, Xiaoxu Li, Zhongwei Si, and Jun Guo. Fine-grained vehicle classification with channel max pooling modified cnns. *IEEE Transactions on Vehicular Technology*, 2019.

[8] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[9] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

[10] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[11] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3973–3981, 2015.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[14] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition)*, July 2017.

[15] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[17] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[18] Pytorch. https://github.com/pytorch/pytorch.

[19] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2399–2406, 2015.

[20] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1449–1457, 2015.

[21] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 365–374, 2017.

[22] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017.

[23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.