Weakly Supervised Fine-Grained Image Recognition Based on Multi-Channel Attention and Object Localization

Sibo Li
School of Computer and Information Engineering
Jiangxi Normal University
Nanchang, China 330022
lisibo@jxnu.edu.cn

Jianming Liu*
School of Computer and Information Engineering
Jiangxi Normal University
Nanchang, China 330022
liujianming@jxnu.edu.cn

Simin Chen
School of Computer and Information Engineering
Jiangxi Normal University
Nanchang, China 330022
simin_chen@jxnu.edu.cn

ABSTRACT
The task of fine-grained image classification recognizes similar subcategories that belong to the same superclass. The subtle differences between inter classes and intra-class diversity make it a challenging task. The existing methods focus on how to locate the most discriminative parts, which usually ignored the whole feature of objects. In this paper, we propose a fine-grained image recognition method based on multi-channel attention and object localization, which mainly includes two parts: multi-channel attention (MCA) module is used to learn different discriminative regions, attention object location (AOL) module can locate the object from the input image. The method we proposed can be trained in an end-to-end manner without adding bounding box/part annotation. At the same time, we build a new application-oriented fine-grained dataset: Poyang Lake birds, which contain 370 species of birds in Poyang Lake. We conducted extensive experiments on our dataset and three commonly used fine-grained datasets (CUB-200-2011, Stanford Cars, and FGVC-Aircraft). Extensive experiments show that our model shows compelling performance whether it is on our dataset or the other existing datasets.

CCS Concepts
• Computing methodologies→Artificial intelligence→ Computer vision→Computer vision problems→Object recognition.

Keywords
Weakly supervised learning; fine-grained visual categorization; visual attention; object region localization.

1. INTRODUCTION
Traditional image recognition only needs to classify general objects (for example a bird, a flower or a dog), and the goal of fine-grained visual categorization is to identify subcategories from the same category (for example birds [6], cars [2], aircraft [4]). Fine-grained image recognition has been regarded as a challenging task for a long time due to the subtle differences between fine-grained image classes, which attracted the attention of a large number of researchers.

ICCPR 2020, October 30-November 1, 2020, Xiamen, China

ACM ISBN 978-1-4503-8783-5/20/10…$15.00
https://doi.org/10.1145/3436369.3436459

Some of the early works [5,7,8] need additional manual annotation, and these methods are based on strongly supervised learning. However, the strongly supervised learning requires much workforce for manual annotation, which is limited to its development to a certain extent. With the rapid development of deep learning, there have been many excellent neural network models, such as Resnet [9]. Therefore, more and more researchers have shifted from strong supervision to weak supervision. Although weakly supervised learning only has an image-level label, it also achieves excellent performance on fine-grained image recognition tasks. In this paper, our method only uses the image-level label and does not use other annotations.

Most of the existing methods [10,11,12] capture the discriminative features in fine-grained images by designing a subnetwork or navigation network to locate the feature of the parts, and compared with the method using the entire image only, they can obtain better performance. However, these methods also have some problems. Firstly, although some of their networks can locate part features, this recognition method based on part features cannot identify all the discriminative parts and will lose some useful information, and it will cause the performance of the model is not good enough. Secondly, some methods combine multiple models for staged training, which makes the training process complex, and the training parameters are very large, which requires a lot of GPU resources.

In this paper, we propose a novel weakly supervised fine-grained image recognition framework to solve these problems. This novel network is mainly composed of a backbone network and two essential modules. The backbone network is mainly used to generate basic feature maps. The two modules are multi-channel attention module and attention object location. First, we use multi-channel attention to generate the attention maps, so that the discriminative regions we find are distributed in different regions as much as possible. Finally, the attention object location module's primary function is to extract the object to be recognized from the attention maps and then input the extracted object image into the backbone network for more refined learning.

2. RELATED WORK
2.1 Fine-Grained Image Recognition
In recent years, thanks to the ImageNet competition, many researchers have proposed many excellent convolutional neural network models on the issue of large-scale image classification, such as Resnet [9], Densest [15]. This also makes the study of fine-grained image classification enter a new stage, transfer to weakly supervised learning with the only image-level label. It is

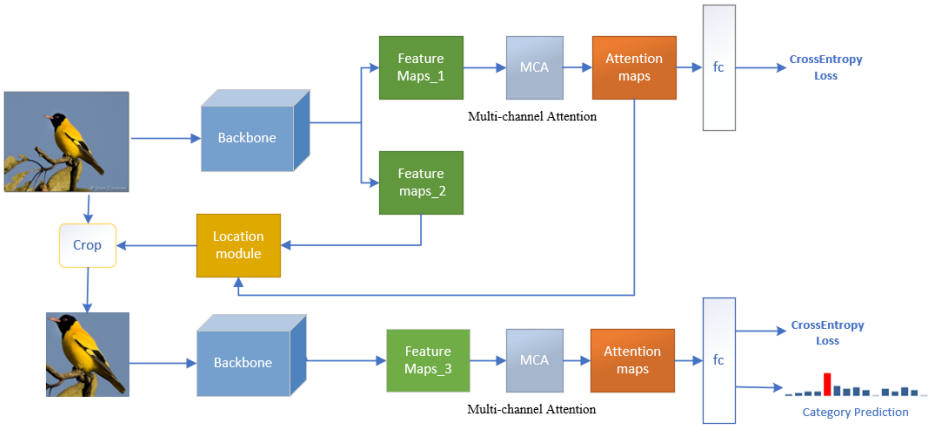

Figure 1. An overview of our framework for fine-grained image recognition. There are two important modules: (a) MCA module is used to learn different discriminative regions so that the generated attention maps can contain a variety of discriminative features. (b) The Location module is used to cut out the object to be recognized from the original image.

no longer like the earlier works [5,7,8] which require additional bounding box/part annotation as additional information for training neural networks. The classic B-CNN [16] uses two parallel CNN models to extract different features, respectively, then performs the outer product operation on the two features to get the final image descriptor. Zheng et al. [12] propose MA-CNN, which generates multiple parts by clustering and pooling from spatially related channels, and classifies each part. RA-CNN [17] utilizes a recursive method to predict the location of the region that needs attention and extracts the corresponding features, then combines various features to predict the final category. Yang et al. [10] design a novel training mechanism, which mainly includes the Navigator network and the Teacher network, so that the Navigator network can detect discriminate information under the teacher network's guidance.

2.2 Object Region Location

Many previous works have tried to locate the part information of the object in the image. Jaderberg et al. [19] propose a space conversion network, that allows the spatial transformation of data in the network, but the network is more difficult to train. Zhang et al. [20] propose an adversarial complementary learning method. This method discovers the entire object by training two classifiers, but there are only two complementary regions in the network, which limits its accuracy. The trilinear attention sampling network (TASN) [21] uses attention to sample multiple regions of interest in the input image and then learns it again through the network. S3N [22] performs selective semantic sampling on the information region to enhance the representation of features. Although the joint global and local features are highly representative, these methods require multi-stage training based on parts and detectors, which complicates the training and testing process.

2.3 Attention Mechanism

We know that the convolutional neural network is a feature extractor. Still, it cannot capture the subtle differences effectively in fine-grained image classification, so some researchers try to use the attention mechanism to discover more detailed information. The attention mechanism is mainly composed of two parts. First, determine the region that needs attention, and second, extracts features from the region that we focus on to obtain the necessary information. Fu et al. [17] propose R-CNN, which uses recursive methods for local region localization and fine-grained feature learning. Jetley et al. [23] propose a cascading attention mechanism to guide the different layers of CNN and connect them to obtain more details as the final linear classifier's input. Hu et al. [24] consider that the importance of different channels may be different. They redistribute different weights according to the importance of each channel by using the attention mechanism for the channels. Woo et al. [25] consider both the importance of different channels and the importance of the same channel in different spatial locations. They propose a method that combines spatial and channel attention mechanisms.

3. METHODS

In this section, we will introduce the weakly supervised fine-grained image recognition framework based on multi-channel and object localization. An overview of our framework for fine-grained image recognition is illustrated in Fig. 1. Our method mainly includes two branches, one is the original image classification branch, and the other is the object image classification branch. The backbone network is used to extract mid-level features. the AOL module is used to locate the object from discriminative regions, and The MCA module can learn different discriminative regions. Our final prediction results come from the object classification branch.

3.1 Attention Object Location (AOL) Module

To obtain a complete object for more detailed learning, first, we can get attention maps $F \in R^{N\times W\times H}$ that contain multiple discriminative features. We are inspired by the SCDA [28] algorithm; after that, aggregation map A is obtained, as shown in Equation 1.

$$A = \sum_{n=0}^{N-1} f_n \tag{1}$$

The location region where the fine-grained object appeared response in most channels of the feature tensor after the convolution. The sum of the convolutional features in the depth direction can highlight the position of the fine-grained object. Furthermore, determining the position of the object in the obtained aggregation map A. $Mask_A$ of the object can be obtained by

$$M_{(x,y)} = \begin{cases} 1, & \text{if } A_{(x,y)} > \bar{a} \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

where $\bar{a}$ represents the mean value of H×W elements in the two-dimensional aggregation map A, (x, y) represents the position coordinated in the two-dimensional map A. The mean $\bar{a}$ as the threshold to determine whether the position in A is the target object or not.

Based on previous experience [28], it can be known that if $Mask_A$ generated only by attention maps F, the positioning performance is not very well. Therefore, Zhang et al. [29] propose a method to solve this problem. They use feature maps F_1 and F_2 to generate $Mask_1$ and $Mask_2$ respectively, and perform an intersection operation on these two masks to obtain $Mask_{12}$ for location. We use a similar method. First, we generate $Mask_A$ through the attention maps F, second generate $Mask_B$ by obtaining the feature maps F_2 from ResNet-50 [9], and finally, do an intersection operation based on $Mask_A$ and $Mask_B$ to obtain an accurate $Mask_{AB}$. We resize $Mask_{AB}$ to the same size as the input image and overlay it on the input image to learn the object location of the second step. The experimental results also prove that our method is effective for improving object location accuracy.

3.2 Multi-Channel Attention (MCA) Module

To extract the salient features of the image, we introduce an attention mechanism, which can assign different weights to important regions. The current mainstream attention modules are Squeeze-and-Excitation (SE) [24] and Convolutional Block Attention Module (CBAM) [25], in this paper, we use the SE module. We have conducted many experiments and have found that the CBAM module is not as good as the SE module in our model performance. We speculate that the Spatial Attention Module (SAM) in CBAM may destroy the channel features of our multi-channel attention loss function.

The role of the attention module is that it can be used to guide the network to focus on meaningful features and extract the features we want. We represent the feature maps extracted from the backbone network as $F_1 = [f'_1, f'_2, \dots, f'_n] \in R^{N\times W\times H}$. The first is the squeeze operation. We do a global average pooling on F_1 to generate the channel-wise descriptor D, where D is formulated as $D = [d_1, d_2, \dots, d_n] \in R^N$. Then D through two consecutive gating mechanisms, this Excitation operation can be expressed as

$$\mathrm{k^n} = \delta\left(\mathrm{W_2^n}\sigma(\mathrm{W_1^n D})\right) = [\mathrm{k_1^n}, \mathrm{k_2^n}, \dots, \mathrm{k_N^n}] \in \mathrm{R^N}. \quad (3)$$

Where σ and δ represent the Relu and Sigmoid functions, respectively. The function to get the attention maps is presented as

$$\mathrm{F} = [\mathrm{k_1^n f'_1}, \mathrm{k_2^n f'_2}, \dots, \mathrm{k_N^n f'_n}] \in \mathrm{R^{N\times W\times H}}. \quad (4)$$

Although the SE module can extract discriminative regions of an image, how to make these discriminative features diverse and complete is still a challenge. Some existing methods [11,26] use a loss function to optimize this problem. The loss function draws features from the same SE module closer, and features from different SE modules as far away as possible. In this paper, we also try to optimize such a problem by adding a loss function.

To make our model find as many discriminative regions as possible to improve the accuracy of location, we introduce the attention loss function in [27]. The loss function is mainly composed of two parts, one is the discriminative component, which can discover discriminative regions, and the other is the diversity component, which can keep the discriminative regions in different positions. First, we use $F \in R^{N\times W\times H}$ to denote the attention maps, where N, H, and W to denote the number of channels, height, and width, respectively. Then we split the value of N into $c \times \zeta$, c and ζ represent the number of categories of the dataset and the number of channels required for each category. Therefore, the N-th feature maps of F is represented as $F_n \in R^{WH}$, the feature maps of the i-th channel group are expressed as $F_i \subseteq R^{\zeta\times WH}$ $(\mathrm{i} = 0,\ 1,\ \dots,\ \mathrm{c}-1)$ represent all channels. The discriminative loss function can be expressed as

$$\mathrm{L_{discr}(F)} = \mathrm{L}\left(\mathrm{y},\ \underbrace{\frac{\left[\mathrm{e^{k(F_0)}}, \mathrm{e^{k(F_1)}}, \dots, \mathrm{e^{k(F_{c-1})}}\right]^{\mathrm{T}}}{\sum_{\mathrm{i=0}}^{\mathrm{c-1}} \mathrm{e^{k(F_i)}}}}_{\mathrm{Softmax}}\right) \quad (5)$$

$$\mathrm{k(F_i)} = \underbrace{\frac{1}{\mathrm{HW}}\sum_{\mathrm{k=1}}^{\mathrm{HW}}}_{\mathrm{GAP}} \underbrace{\max_{\mathrm{j}=1,2,\dots,\zeta}}_{\mathrm{CCMP}} \underbrace{\left[\mathrm{M_i}\cdot \mathrm{F_{i,j,k}}\right]}_{\mathrm{CWA}}, \quad (6)$$

Where GAP, CCMP, and CWA stand for Global Average Pooling, Cross-Channel Max Pooling, and Channel-Wise Attention, respectively. And M_i is formulated as $M_i = diag(Mask_i)$, where diag(·) is the operation to find the diagonal matrix, $Mask_i \in R^\zeta$ is a vector with random $\left\lceil\frac{\zeta}{2}\right\rceil$ elements of value 0 and $\left\lceil\frac{\zeta}{2}\right\rceil$ elements of value 1. The y is the real label, L is the cross-entropy loss function. The role of CWA is a random channel discarding process, and CCMP is to do max pooling between channel groups. Second, the diversity loss function can be expressed as

$$\mathrm{L_{diver}(F)} = \frac{1}{\mathrm{c}}\sum_{\mathrm{i=0}}^{\mathrm{c-1}} \mathrm{u(F_i)} \quad (7)$$

$$\mathrm{u(F_i)} = \sum_{\mathrm{k=1}}^{\mathrm{HW}} \underbrace{\max_{\mathrm{j}=1,2,\dots,\zeta}}_{\mathrm{CCMP}} \underbrace{\left[\frac{\mathrm{e^{F_{i,j,k}}}}{\sum_{\mathrm{k'=1}}^{\mathrm{HW}} \mathrm{e^{F_{i,j,k'}}}}\right]}_{\mathrm{Softmax}} \quad (8)$$

The function of the diversity loss function is mainly used to constrain the features of each channel in a feature channel-group F_i which is different, as shown in Equation 7, u(·) is defined as Equation 8. And softmax function is used to normalize in the spatial dimension.

3.3 Classification Loss

We use cross entropy loss as the classification loss. L_{CE1} and L_{CE2} represent the original image classification loss and the object image classification loss respectively. L_{att_1} and L_{att_2} represent the loss function of attention maps_1 and attention maps_2 respectively. Our total loss function can be summarized as:

$$\mathrm{L_{total}} = \alpha\mathrm{L_{CE1}} + \beta\mathrm{L_{CE2}} + \gamma\mathrm{L_{att\_1}} + \upsilon\mathrm{L_{att\_2}} \quad (9)$$

Where $\alpha, \beta, \gamma, \upsilon$ are hyper-parameters.

4. EXPERIMENTS AND RESULTS

4.1 Datasets and Implementation Details

4.1.1 Datasets Poyang Lake Birds

We contribute a fine-grained Poyang Lake Birds dataset, which is a complete dataset currently containing birds in Poyang Lake National Nature Reserve. There are 381 species of birds listed on the Birds List of Poyang Lake National Nature Reserve. We use search engines (such as Google and Baidu) to grab images and merge remarkably similar classes by manual validation. Finally, our dataset contains a total of 370 species of birds, the number of images in each category is less than 120, with 32,795 images. Since we use the image-level label, we do not annotate the collected images. This dataset has two advantages. First, it contains a relatively wide variety of birds in Poyang Lake and is specifically for the Poyang Lake bird. Second, compared with the

CUB-200-2011 dataset that is also about birds, our dataset has a larger size and more images for each category.

We measure the performance of our model on our Poyang Lake Birds dataset and three publicly available fine-grained image datasets, including Caltech-UCSD birds (CUB-200-2011) [6], Stanford Cars [2], and FGVC-Aircraft [4]. The statistics of the four datasets are shown in Table 1. We use top-1 accuracy as our evaluation metric.

Table 1. Statistics of four datasets

Datasets	Category	Train Set	Test Set
CUB-200-2011[6]	200	5994	5794
Stanford Cars [2]	196	8144	8041
FGVC-Aircraft [4]	100	6667	3333
Poyang Lake Birds (ours)	**370**	**16489**	**16306**

4.1.2 Implementation Details

In the next experiment, we use PyTorch to implement our method and use ResNet-50 pre-trained on ImageNet as the backbone network. In addition to the image-level label, no other labels are used. We resize the input image to a size of 448×448. Stochastic gradient descent (SGD) is used as an optimizer, and batch normalization is used as a regular. The hyper-parameters are set to 1, 1, 0.001 and 0.001 respectively. The momentum is set to 0.9, the weight decay is 0.0001, and is limited by GPU resources, our batch size is 9, the initial learning rate is 0.001, the maximum training epoch is 200, and the GeForce GTX 1080Ti GPU with a memory size of 11G is used.

4.2 Comparisons Against Baselines

In this section, we conduct experiments of our model on the Poyang Lake Birds, CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets, and compare with baselines, respectively.

4.2.1 Comparisons Against Baselines on Poyang Lake Birds Dataset

We first conduct four experiments on our Poyang Lake Birds dataset. We compare our method with 3 existing methods, including: ResNet-50 [9], ResNet-101 [9] and HBP [14]. The comparison results are shown in Table 2. We can find that our method achieves the best performance.

Table 2. Accuracy comparison with different methods for fine-grained image recognition on Poyang Lake Birds dataset

methods	Accuracy (%)
ResNet-50 [9]	89.5
ResNet-101 [9]	90.4
HBP [14]	91.4
ours	**92.5**

4.2.2 Comparisons Against Baselines on CUB-200-2011 Dataset

Then, we compare our method with 11 existing methods, including 3 training with annotation methods: MG-CNN [8], FCAN [7], and Mask-CNN [5]; and 8 training without annotation methods: ResNet-50, RA-CNN [17], MA-CNN [12], MAMC [11], PC [18], HBP, DFL-CNN [3] and NTS-Net [10]. As reported in Table 3, our method consistently achieves the best performance. Specifically, the two methods of MG-CNN and FCAN added annotation when training the network, and their classification accuracy reached 83.0% and 84.3%, respectively. Compared with these two methods, our method can achieve 87.1% without the help of manual annotation, and compare with the benchmark model ResNet-50, our proposed method is 3.5% higher than it. As for RA-CNN, MA-CNN, MAMC, and HBP, these are methods based on the attention mechanism. After removing the object location module, the network we proposed only retains the attention mechanism, the model's performance can still exceed them. Compared with the state-of-the-art AE-AN method, the accuracy of our method is 88.0%, although it is only 0.2% higher than it, on other fine-grained datasets, our model performance is equally excellent.

4.2.3 Comparisons Against Baselines on Stanford Cars Dataset

Next, we train our model on the Stanford Cars dataset. We compare our method with several methods, including 2 training with annotation methods: R-CNN [1] and FCAN, and 8 training without annotation methods: RA-CNN, MA-CNN, PC, MAMC, HBP, DFL-CNN, AE-AN and NTS-Net. Table 3 shows that accuracy comparison with different methods for fine-grained image recognition on Stanford Cars dataset. Our method achieves the best performance compared to all the baselines on the dataset. AE-AN locates discriminative regions using confrontational tactics, but this typical way is easy to destroy the replacement of objects, and our experimental result is 0.7% higher than it. Comparing MA-CNN and NTS-Net, they are first located to different discriminative regions, and then extract features from these discriminative regions and combine them for classification. Our classification accuracy is 1.7% and 0.6% higher than them.

4.2.4 Comparisons Against Baselines on FGVC-Aircraft Dataset

Finally, we train our model on the FGVC-Aircraft dataset. We compare our method with 9 methods including MG-CNN, B-CNN [16], PC, RA-CNN, MA-CNN, HBP, AE-AN, NTS-Net and DFL-CNN. As shown in Table 3, whether it is compared with the training with annotation or the training without annotation methods, our method always achieves compelling performance on FGVC-Aircraft dataset. Compared with PC, they use the more advanced DenseNet-161 as the backbone network, our method can still surpass it, compared with the suboptimal method (NTS-Net and DFL-CNN), we still achieve a 1.0% improvement in accuracy, which proves that our method has a strong ability to discover features and learn to discriminative features.

4.3 Ablation Study

We provide ablation studies of our method in terms of three modules, i.e., (i) object location module, (ii) attention mechanism, and (iii) attention loss function. We conduct experiments on CUB-200-2011 dataset and the experimental results are represented in Table 4. We can see that when we removed the object location module and only used the attention mechanism and the attention loss function, our model's recognition accuracy is achieved 86.1%. When we removed the attention mechanism and only used the object location module and the attention loss function, the accuracy is achieved at 87.8%. When we removed the attention loss function and only used the object location module and the attention mechanism, the accuracy is achieved at 87.1%. Only jointly using all three modules leads to the best performance.

Table 3. Accuracy comparison with different methods for fine-grained image recognition on CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets. Train Anno. Represents using bounding box in training

Methods	Train Anno.	CUB-200-2011 Accuracy (%)	Stanford Cars Accuracy (%)	FGVC-Aircraft Accuracy (%)
MG-CNN [8]	√	83.0	-	86.6
R-CNN [1]	√	-	88.4	-
B-CNN [16]		-	-	84.1
FCAN [7]	√	84.3	91.3	-
ResNet-50 [9]		84.5	-	-
Mask-CNN [5]	√	85.4	-	-
RA-CNN [17]		85.3	92.5	88.2
MA-CNN [12]		86.5	92.8	89.9
MAMC [11]		86.5	93.0	-
PC [18]		86.9	92.9	89.2
HBP [14]		87.1	93.7	90.3
DFL-CNN [3]		87.4	93.8	92.0
NTS-Net [10]		87.5	93.9	91.4
AE-AN [13]		87.8	93.8	91.3
Ours		**88.0**	**94.5**	**93.0**

Table 4. Impact of object location module, attention mechanism and attention loss function on recognition accuracy

object location module	attention mechanism	attention loss function	Accuracy (%)
	√	√	86.1
√		√	87.8
√	√		87.1
√	**√**	**√**	**88.0**

5. CONCLUSION

In this paper, we proposed a weakly supervised fine-grained image recognition based on multi-channel attention and object localization. We used the attention object location module to find the complete object region. Then we used multi-channel attention to discover discriminative features, the network we proposed can be trained in an end-to-end manner, and apart from the image-level label, no other labels are used. Furthermore, we contributed a Poyang Lake Birds dataset. On the Poyang Lake Birds, CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets, our experimental results show that our framework performs well and is comparable with state-of-the-art methods. In future work, we will study how to sample the attention maps to obtain different part features and further improve the performance of the model.

6. ACKNOWLEDGMENTS

This work was financially supported by the Natural Science Foundation of China (No. 61662034), the Youth Science Foundation of Education Department of Jiangxi Province (GJJ12213).

7. REFERENCES

[1] Girshick, R., Donahue, J., Darrell, T., & Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. (Jun. 2014), 580-587. DOI= http://dx.doi.org/10.1109/cvpr.2014.81.

[2] Krause, J., Jin, H., Yang, J., and Fei-Fei, L. 2015. Fine-grained recognition without part annotations. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (Jun. 2015), 5546-5555. DOI= http://dx.doi.org/10.1109/cvpr.2015.7299194.

[3] Wang, Y., Morariu, V. I., & Davis, L. S. 2018. Learning a Discriminative Filter Bank Within a CNN for Fine-Grained Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (Jun. 2018), 4148-4157. DOI= http://dx.doi.org/10.1109/cvpr.2018.00436.

[4] Maji, S., Rahtu, E., Kannala, J., Blaschko, M., & Vedaldi, A. 2013. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151.

[5] Wei, X. S., Xie, C. W., Wu, J., & Shen, C. 2018. Mask-CNN: Localizing Parts and Selecting Descriptors for Fine-Grained Bird Species Categorization. Pattern Recognition. 76, (Apr. 2018), 704–714. DOI= http://dx.doi.org/10.1016/j.patcog.2017.10.002.

[6] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

[7] Liu, X., Xia, T., Wang, J., Yang, Y., Zhou, F., & Lin, Y. 2016. Fully convolutional attention networks for fine-grained recognition. arXiv preprint arXiv:1603.06765.

[8] Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., and Zhang, Z. 2015. Multiple Granularity Descriptors for Fine-Grained

Categorization. 2015 IEEE International Conference on Computer Vision (ICCV). (Dec. 2015), 2399-2406. DOI= http://dx.doi.org/10.1109/iccv.2015.276.

[9] He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (Jun. 2016), 770-778. DOI= http://dx.doi.org/10.1109/cvpr.2016.90.

[10] Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., & Wang, L. 2018. Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV). 420-435.

[11] Sun, M., Yuan, Y., Zhou, F., & Ding, E. 2018. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. Lecture Notes in Computer Science. 834–850. DOI= http://dx.doi.org/10.1007/978-3-030-01270-0_49.

[12] Zheng, H., Fu, J., Mei, T., & Luo, J. 2017. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. 2017 IEEE International Conference on Computer Vision (ICCV). (Oct. 2017), 5219-5227. DOI= http://dx.doi.org/10.1109/iccv.2017.557.

[13] Ji, J., Jiang, L., Zhang, T., Zhong, W., & Xiong, H. 2020. Adversarial erasing attention for fine-grained image classification. Multimedia Tools and Applications. (Jan. 30, 2020), 1-23. DOI= http://dx.doi.org/10.1007/s11042-020-08666-3.

[14] Yu, C., Zhao, X., Zheng, Q., Zhang, P., & You, X. 2018. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. Lecture Notes in Computer Science. 595–610. DOI= http://dx.doi.org/10.1007/978-3-030-01270-0_35.

[15] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (Jul. 2017), 2261-2269. DOI= http://dx.doi.org/10.1109/cvpr.2017.243.

[16] Lin, T.Y., RoyChowdhury, A., & Maji, S. 2015. Bilinear CNN Models for Fine-Grained Visual Recognition. 2015 IEEE International Conference on Computer Vision (ICCV). (Dec. 2015), 1449-1457. DOI= http://dx.doi.org/10.1109/iccv.2015.170.

[17] Fu, J., Zheng, H., & Mei, T. 2017. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (Jul. 2017), 4476-4484. DOI= http://dx.doi.org/10.1109/cvpr.2017.476.

[18] Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., & Naik, N. 2018. Pairwise Confusion for Fine-Grained Visual Classification. Lecture Notes in Computer Science. 71–88. DOI= http://dx.doi.org/10.1007/978-3-030-01258-8_5.

[19] Jaderberg, M., Simonyan, K., & Zisserman, A. 2015. Spatial transformer networks. In Advances in neural information processing systems. 2017-2025.

[20] Zhang, X., Wei, Y., Feng, J., Yang, Y., & Huang, T. 2018. Adversarial Complementary Learning for Weakly Supervised Object Localization. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (Jun. 2018), 1325-1334. DOI= http://dx.doi.org/10.1109/cvpr.2018.00144.

[21] Zheng, H., Fu, J., Zha, Z.-J., & Luo, J. 2019. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (Jun. 2019), 5007-5016. DOI= http://dx.doi.org/10.1109/cvpr.2019.00515.

[22] Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., & Jiao, J. 2019. Selective Sparse Sampling for Fine-Grained Image Recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). (Oct. 2019), 6598-6607. DOI= http://dx.doi.org/10.1109/iccv.2019.00670.

[23] Jetley, S., Lord, N. A., Lee, N., & Torr, P. H. 2018. Learn to pay attention. arXiv preprint arXiv:1804.02391.

[24] Hu, J., Shen, L., & Sun, G. 2018. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (Jun. 2018), 7132-7141. DOI= http://dx.doi.org/10.1109/cvpr.2018.00745.

[25] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. Lecture Notes in Computer Science. 3–19. DOI= http://dx.doi.org/10.1007/978-3-030-01234-2_1.

[26] Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L., Li, J., and Lim, S.-N. 2019. Cross-X Learning for Fine-Grained Visual Categorization. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). (Oct. 2019), 8241-8250. DOI= http://dx.doi.org/10.1109/iccv.2019.00833.

[27] Chang, D., Ding, Y., Xie, J., Bhunia, A. K., Li, X., Ma, Z., … Song, Y. Z. 2020. The Devil is in the Channels: Mutual-Channel Loss for Fine-Grained Image Classification. IEEE Transactions on Image Processing. 29, 4683–4695. DOI= http://dx.doi.org/10.1109/tip.2020.2973812.

[28] Wei, X.-S., Luo, J.-H., Wu, J., & Zhou, Z.-H. 2017. Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. IEEE Transactions on Image Processing. 26, 6 (Jun. 2017), 2868–2881. DOI= http://dx.doi.org/10.1109/tip.2017.2688133.

[29] Zhang, F., Zhai, G., Li, M., & Liu, Y. 2020. Three-branch and Mutil-scale learning for Fine-grained Image Recognition (TBMSL-Net). arXiv preprint arXiv:2003.09150.