2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) | 978-1-7281-6338-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/MLSP52302.2021.9596481

GRAD-CAM GUIDED CHANNEL-SPATIAL ATTENTION MODULE FOR FINE-GRAINED VISUAL CLASSIFICATION

Shuai Xu[1], *Dongliang Chang*[1], *Jiyang Xie*[1], *Zhanyu Ma*[1,2,*]

[1] Pattern Recognition and Intelligent System Lab.,
Beijing University of Posts and Telecommunications, Beijing, China
[2] Beijing Academy of Artificial Intelligence,
5th Floor, Saier Building, 1 Zhongguancun East Road, Haidian District, Beijing, China

ABSTRACT

Fine-grained visual classification (FGVC) is becoming an important research field, due to its wide applications and the rapid development of computer vision technologies. The current state-of-the-art (SOTA) methods in the FGVC usually employ attention mechanisms to first capture the semantic parts and then discover their subtle differences between distinct classes. The existing attention modules have significantly improved the classification performance but they are poorly guided since part-based detectors in the FGVC depend on the network learning ability without the supervision of part annotations. As obtaining such part annotations is labor-expensive, some visual localization and explanation methods, such as gradient-weighted class activation mapping (Grad-CAM), can be utilized for supervising the attention mechanism. In this paper, we propose a Grad-CAM guided channel-spatial attention module for the FGVC, which employs the Grad-CAM to supervise and constrain the attention weights by generating the coarse localization maps. To demonstrate the effectiveness of the proposed method, we conduct comprehensive experiments on three popular FGVC datasets, including CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets. The proposed method outperforms the SOTA attention modules in the FGVC task. In addition, visualizations of the feature maps demonstrate the superiority of the proposed method against the SOTA approaches.

Index Terms— Fine-grained visual classification, gradient-weighted class activation mapping, channel-spatial attention mechanism

1. INTRODUCTION

Fine-grained visual classification (FGVC) aims to distinguish fine-grained classes under the same coarse class labels, *e.g.*, birds [1], airplanes [2], and cars [3] *etc.* The main challenge of the FGVC task is the tiny inter-class difference along with significant intra-class variance. For example, it is difficult to distinguish a redhead woodpecker from a pileated woodpecker and a downy woodpecker caused by highly similar sub-categories, but with the adjustment of poses, scales, and rotations, the redhead woodpecker can be photographed in a very different visual view. In order to generate discriminative features more precisely, we better have the ability to capture the key characteristics of the red head and ignore the background and other irrelevant regions, which is an obvious way for overcoming the challenge.

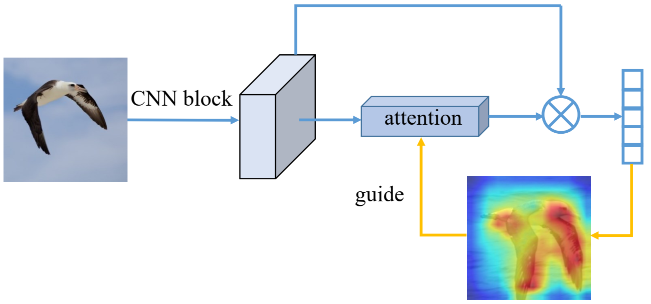

Fig. 1. Motivation of the Grad-CAM guided channel-spatial attention module. The blue part is the general pipeline of the previous attention mechanisms. The yellow line is our proposed supervision mechanism that the weights obtained by the gradient backpropagation in the Grad-CAM are used for the guidance of the attention weights, with which the attention mechanisms focus on parts that contribute significantly to classification.

The existing approaches can be roughly divided into two classes: (1) searching the informative regions that contribute the most to the classification task [4, 5, 6] and (2) paying more attention to extract high-order features from the images [7, 8, 9, 10, 11]. For the former one, previous approaches [4, 7] usually employed the prior location information such as part-level bounding boxes and segmented masks to generate the discriminative parts. Meanwhile, for the latter one, the powerful deep networks [8, 9] were employed for feature ex-

[*] CORRESPONDING AUTHOR

978-1-7281-6338-3/21/$31.00 ©2021 IEEE

traction, and different loss functions [10, 11] were designed for constraining these networks to improve the discrimination of the extracted features. Recently, the attention mechanisms [12, 13, 14, 15], which only require image labels for training, have gradually replaced the manual annotation methods, since part annotations are time-consuming and laborious that limits the flexibility and versatility of the real-world FGVC applications. Compared with the approaches that introduce complex structures, the attention-based methods add fewer learned parameters for model training, which efficiently reduces the computation cost.

The attention mechanisms fully simulate the observation habits of human eyes, which always concentrate on the most distinctive regions when seeing images. For example, we can easily pay attention to the head and the wings of a bird and ignore the other common regions to identify its species. Inspired by this, many methods have been proposed by utilizing the attention mechanisms to detect the discriminative information from the images, including channel attention [5, 12], spatial attention [14], and channel-spatial attention [16]. Specifically, SENet [12] introduced "squeeze-and-excitation" (SE) blocks to adaptively recalibrate the feature maps in channel-wise by modeling the interactions between channels. The trilinear attention sampling network [5] generated attention maps by integrating feature channels with their relationship matrix and highlighted the attended parts with high resolution. The recurrent attention convolutional neural network (RA-CNN) [14] introduced attention proposal network (APN) to capture the region relevance information based on the extracted features, and then amplified the attention region crops to make the network gradually focus on the key areas. The convolutional block attention mechanism (CBAM) [16] is a channel-spatial attention method that utilizes both the channel-level and region-level information. It can effectively improve the characteristic expression ability of the networks. The existing methods [5, 12, 14, 16] usually utilize different attention mechanisms to generally adjust the distributions of the attention weights for balancing the contributions of feature maps extracted from each part. Although these methods for obtaining the weights are different, they are all constructed based on the original feature maps only, without part information supervision. Obviously, if the feature maps focus on the non-significant parts such as backgrounds and distractions, the attention mechanism is meaningless under the unsupervised conditions.

In Convolutional neural networks (CNNs), as each channel of the feature maps can be also considered as a semantic part [13], supervision on discriminative parts can be transferred to that on channels. Grad-CAM [17] is usually introduced to illustrate attentions of the networks with heat maps and visualize the attentions in each part by weighted averaging channels, so we can use it to guide the networks to focus on the parts which have specific characteristic information, such as the head and the beak of a bird. Therefore, in this paper, we propose a Grad-CAM guided channel-spatial attention module to focus on more efficient parts and discard the redundant information for the classification. In our module, the channel weighted feature maps are pooled along with the channel dimensions and multiplied by the original feature maps to obtain the channel-spatial attention maps. Meanwhile, a Grad-CAM guided channel-spatial attention mechanism loss (GGAM-Loss) is applied for guiding the learning process of the channel weights and forcing the attention module to focus on the parts that contribute most to the classification. As shown in Figure 1, we employ the channel weights obtained from the gradient backpropagation in the Grad-CAM to constrain the channel weights of the forward propagation.

Our contributions can be summarized as follows:

- We address the challenge of the FGVC by proposing a Grad-CAM guided channel-spatial attention module, which constrains the channel-spatial attention mechanism to focus on the parts that contribute most to the classification. Moreover, it is not limited to a specific network architecture.
- We conduct comprehensive experiments on the three commonly used FGVC datasets, *i.e.*, CUB-200-2011 [1], FGVC-Aircraft [2] and Stanford Cars [3] datasets. The results show the effectiveness of the proposed method.

2. METHODOLOGY

2.1. Channel-spatial Attention Mechanism

The channel-spatial attention is a module that combines both the spatial attention and the channel attention. Specifically, as shown in Figure 2, the input image is processed through a series of convolution and pooling operations F_{cp} and feature maps denoted as $A = [a_1, a_2, ..., a_C] \in R^{C\times W\times H}$ are obtained, with height H, width W, and channel number C, respectively. Then we apply a global average pooling F_{cg} to downsample each feature map in A and a two-layer fully connected (FC) network F_{cr} with softmax function to calculate the weights of each channel as the channel attention weights $S = [s_1, s_2, ..., s_C] \in R^C$, according to [12]. We rescale the original feature maps A by S, which obtains the weighted feature maps $B = [b_1, b_1, ..., b_c] \in R^{C\times W\times H}$ by F_{cm} as

$$b_c = F_{cm}(a_c, s_c) = a_c \cdot F_{cr}\left(F_{cg}(A)\right)_c, \qquad (1)$$

where $c = 1, \cdots, C$.

After gaining the channel attention-weighted feature maps B, spatial attention is undertaken. Through the operation F_{fa}, which combines a channel-wise summation and a 2D softmax function, the feature maps in B are flattened along the channel dimension to obtain the spatial attention weights $T \in R^{W\times H}$. Then the channel-spatial attention-weighted feature maps $D = [d_1, d_1, ..., d_c] \in R^{C\times W\times H}$ are obtained by rescaling A with T as

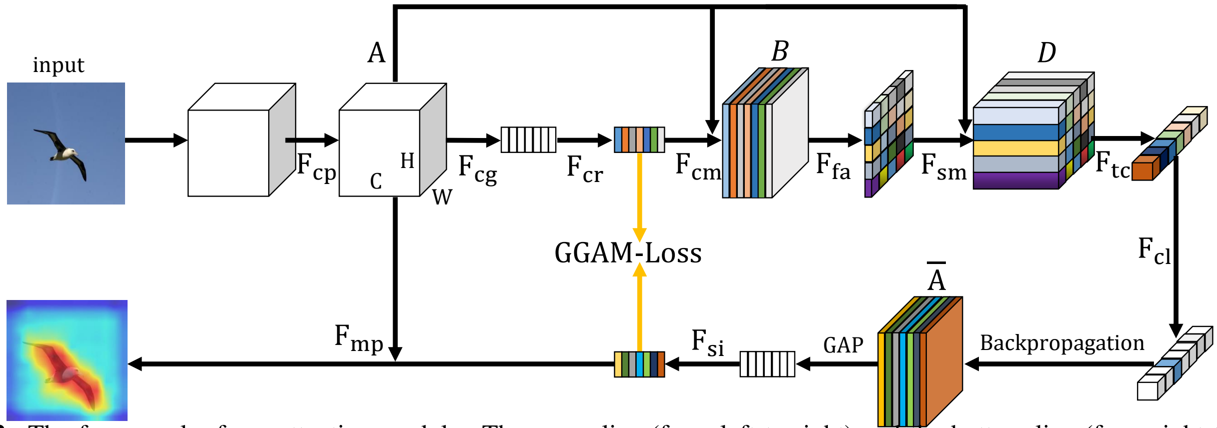

Fig. 2. The framework of our attention module. The upper line (from left to right) and the bottom line (from right to left) present the forward and the gradient backpropagation processes, respectively. A symmetrical Kullback-Leibler (KL) divergence between the weights of each channel in forward propagation and the weights of each feature map in the Grad-CAM is utilized as the loss function in backpropagation to supervise the channel-spatial attention.

$$d_c = F_{sm}(a_c, T) = a_c \odot F_{fa}(B), \tag{2}$$

where $\odot$ is Hadamard product and

$$T = F_{fa}(B) = \frac{\sum_{c=1}^{C} b_c}{\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{c=1}^{C} b_{c,i,j}}. \tag{3}$$

Then the classification is undertaken according to D. F_{tc} is the classifier with multiple FC layers and a softmax function.

2.2. Grad-CAM

The Grad-CAM uses the class-specific gradient information and it flows into the final convolutional layer of a CNN to generate a heat map, which shows the main concentrated regions of the CNN. Specifically, as illustrated in Figure 2, given an input image, we obtain the score y^k for the predicted class k before the last softmax function. Then, y^k is propagated back to the elements of A through the upper line and we gain the gradient $\frac{\partial y^k}{\partial A_{c,i,j}}$. The weights β_c^k of the Grad-CAM, which represent the importance of feature map c with the predicted class k, can be defined as

$$\beta_c^k = \underbrace{\frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}}_{GAP}\frac{\partial y^k}{\partial A_{c,i,j}}. \tag{4}$$

2.3. Grad-CAM guided channel-spatial attention loss

In the FGVC, attention mechanism is introduced to ensure the CNN focus on more effective parts mainly, so as to improve the classification accuracy. As mentioned above, Grad-CAM can extract the key parts of the input image. In this section, we follow the same motivation and propose the Grad-CAM guided channel-spatial attention loss to enhance discriminative part searching and feature extraction abilities of CNNs in the FGVC.

Table 1. The statistics of the three FGVC datasets. #Class, #Training, and #Test are class number, training sample number, and test sample number, respectively.

Dataset	#Class	#Training	#Test
CUB-200-2011 [1]	200	5994	5794
FGVC-Aircraft [2]	100	6667	3333
Stanford Cars [3]	196	8144	8041

In Figure 2, after the F_{cr} operation, we can obtain the weights S of each channel in A. Through the operation F_{si}, we apply a sigmoid function for $\beta_c^k, c = 1, \cdots, C$, to scale their intervals and obtain $\tilde{\beta}^k = [\tilde{\beta}_1^k, \cdots, \tilde{\beta}_C^k] \in R^C$, where $\tilde{\beta}_c^k = \text{sigmoid}(\beta_c^k)$. As $\tilde{\beta}^k$ can reflect the contribution of each channel to the classification, we constrain the channel attention weights S with it. Here, we propose the Grad-CAM guided channel-spatial attention mechanism loss, GGAM-Loss in short, to construct the regularization term. The GGAM-Loss (L_{GGAM}), which performs as a symmetrical Kullback-Leibler (KL) divergence between S and $\tilde{\beta}^k$, can be defined as

$$L_{\text{GGAM}} = \frac{1}{2}\left(\text{KL}(S||\tilde{\beta}^k) + \text{KL}(\tilde{\beta}^k||S)\right), \tag{5}$$

where $\text{KL}(x||y)$ is the KL divergence from x to y.

Moreover, as we use the original cross-entropy (CE) loss L_{CE} for training the model as well, the total loss function $Loss$ of the whole network can be defined as

$$Loss = L_{\text{CE}} + \lambda L_{\text{GGAM}}, \tag{6}$$

where λ is a nonnegative multiplier.

Table 2. Classification accuracies (%) on the CUB-200-2011, the FGVC-Aircraft, and the Stanford Cars datasets. The best results on each dataset are in **bold**, and the second best results are in underline.

Datasets	Base Model	CUB-200-2011	FGVC-Aircraft	Stanford Cars
RA-CNN (CVPR17 [14])	VGG19	85.30	88.20	92.50
MA-CNN (ICCV17 [13])	VGG19	84.92	90.35	92.80
SENet (CVPR18 [12])	VGG19	84.75	90.12	89.75
SENet (CVPR18 [12])	ResNet50	86.78	91.37	93.10
CBAM (ECCV18 [16])	VGG19	84.92	90.32	91.12
CBAM (ECCV18 [16])	ResNet50	86.99	91.91	93.35
DFL (CVPR18 [18])	ResNet50	87.40	91.73	93.11
NTS (ECCV18 [6])	ResNet50	87.52	91.48	93.90
TASN(CVPR2019 [5])	VGG19	86.10	90.83	92.40
TASN(CVPR2019 [5])	ResNet50	87.90	92.56	93.80
DCL(CVPR2019 [19])	ResNet50	87.80	93.00	94.50
ACNet(CVPR2020 [20])	ResNet50	88.10	92.40	**94.60**
Ours	VGG19	87.34	91.55	93.32
Ours	ResNet50	**88.45**	**93.42**	94.41

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Datasets

We evaluate our method on three challenging FGVC datasets, including CUB-200-2011 [1], FGVC-Aircraft [2], and Stanford Cars [3] datasets. The statistics of the datasets mentioned above, including class numbers and the training/test sample numbers are shown in Table 1. We followed the same training/test splits as presented in the Table 1. For model training, we did not use artificially marked bounding box or part annotation.

3.2. Implementation Details

In order to compare the proposed method with other attention mechanisms, we resized every image to 448×448, which is standard in the literatures [19, 20]. The backbones we used for extracting features were VGG19 and ResNet50 which were pre-trained on the ImageNet dataset. We used stochastic gradient descent optimizer. The weight dacay value and the momentum were kept as 5×10^{-4} and 0.9, respectively, with 100 epochs. The learning rate of the FC layers was initially set at 0.1 and we used the cosine anneal schedule update strategy [21]. The learning rate of the pre-trained feature extaction layers was one-tenth of the FC layers.

3.3. Experimental Results

According to the aforementioned implementation details, the detailed results are listed in Table 2. Our method achieves significant performance improvement on all the three datasets and the evaluation results can be summarized as follows:

- On the CUB-200-2011 dataset, our method achieves the best result on both VGG19 and ResNet50, respectively, comapred with their corresponding referred methods. Our method exceeds the second best method, TASN, by 1.24% with the VGG19. In addition, compared with the leading result achieved by the ACNet, our method has improved the accuracy by 0.35% with the ResNet50.

Table 3. Ablation study of our method on classification accuracies (%). Key modules of the proposed method, including the channel attention, the spatial attention, and the Grad-CAM are compared. "✓" represents the module contained, otherwise "×". The best result is in **bold**.

Spatial attention	Channel attention	GGAM-Loss	Accuracy
×	×	×	85.10
✓	×	×	85.61
×	✓	×	85.39
✓	✓	×	86.86
×	×	✓	85.30
✓	×	✓	86.58
×	✓	✓	86.26
✓	✓	✓	**88.45**

- On the FGVC-Aircraft dataset, our method also obtains the best accuracy of 93.42% with the ResNet50, around 0.4% improvement than the DCL. With the VGG19, the result of our method also improves slightly.
- On the Stanford Cars dataset, our method outperforms the most compared methods, especially with the same VGG19 backbone. The accuracy of the ACNet with the ResNet50 turns out 0.19% better than ours. Note that the ACNet depends mainly on the addition of the network parameters and the complex training process to improve the accuracy, which is much more complex than ours.

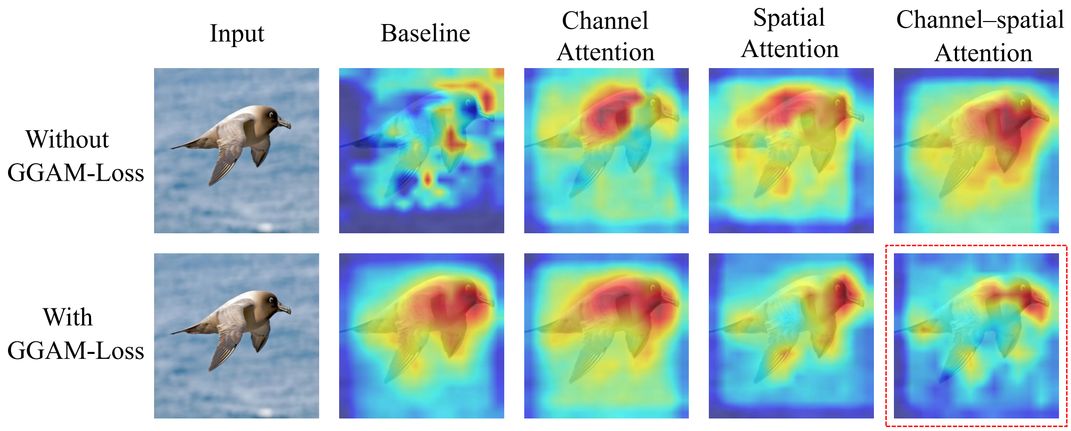

Fig. 3. Visualizations of the ablation models in Section 3.4. The first column represents the original image. The following four columns show visualization results of the baseline, the channel attention, the spatial attention, and the channel-spatial attention, respectively. The top row is trained without the GGAM-Loss, while the bottom row is trained with the GGAM-Loss.The red box indicates the visualization result of our proposed method.

3.4. Ablation Study

Attention mechanisms and Grad-CAM are major modules of our method, and the attention mechanisms include channel and spatial attention mechanisms. We analyze the influence of each module by the experimental results. The ablation experiments are all conducted on the CUB-200-2011 dataset and we use the ResNet50 as the base model if not particularly mentioned. The experimental results are shown in Table 3.

- **Effectiveness of the attention mechanisms.** Compared with the baseline model, the spatial attention can improve performance by 0.51% and the channel attention also has a slight promotion. In particular, the combination of channel and spatial attention obtains a 1.76% increase on accuracy. This enhancement is obvious and shows that the channel-spatial attention is useful for the FGVC.

- **Effectiveness of the GGAM-Loss.** It can be seen that the classification accuracy of each attention mechanism model is improved after adding the GGAM-Loss as the constraint for the attention mechanism. The above results demonstrate the effectiveness of the GGAM-Loss.

3.5. Visualizations

In order to better explain the improvement of our method, Figure 3 shows the visualizations of each model in Section 3.4, which were generated by the Grad-CAM. The baseline cannot clearly focus on the right region of the object. With the addition of the attention mechanisms, the models tend to pay attention on the beak and the neck of the bird, which are discriminative parts. After adding the GGAM-Loss, the models can focus on more accurate discriminant characteristics and pay less attention to background information.

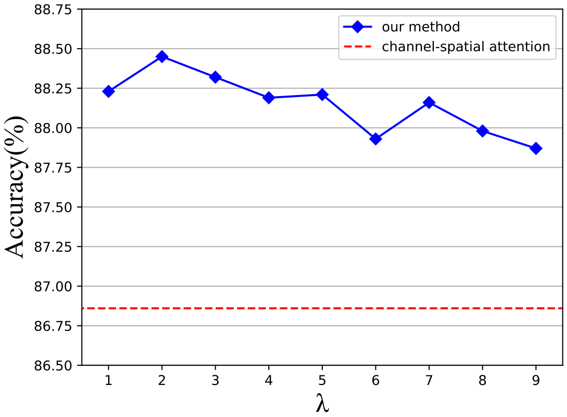

Fig. 4. Sensitivity study of λ for our model on the CUB-200-2011 dataset.

3.6. Sensitivity Study of λ

In order to evaluate the robustness of our method, we conduct the sensitivity study of the hyperparameter λ to see whether the network performance changes a lot with a change of λ. We conduct this study on the CUB-200-2011 dataset and we use the ResNet50 as the base model. We run the proposed model set with λ varying from 1 to 9 with step size of 1. The classification accuracies are shown in Figure 4. From Figure 4, it can be observed that the performance of our method has always been better than the channel-spatial attention (without the GGAM-Loss) and does not change much by varying the value of λ, which proves the effectiveness and robustness of our method.

4. CONCLUSIONS

In this paper, we proposed a Grad-CAM guided channel-spatial attention module for the FGVC task to focus on the most discriminative parts in the images. Note that the proposed module can be also applied to other network architectures. The performance of the proposed method is evaluated in the FGVC task and superior performance is achieved on three FGVC datasets (CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets). The effectiveness of the key modules of the proposed method were also evaluated. Visualizations of the feature maps illustrate the validity of the porposed method.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant No. 2020AAA0105200, 2019YFF0303300, Subject II No. 2019YFF0303302, in part by National Natural Science Foundation of China (NSFC) under Grant 61922015, 61773071, U19B2036, in part by Beijing Natural Science Foundation Project under Grant Z200002, in part by the scholarship from China Scholarship Council (CSC) under Grant CSC No. 202006470036, and in part by the BUPT Excellent Ph.D. Students Foundation No. CX2020105.

6. REFERENCES

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” *California Institute of Technology*, 2011.

[2] S. Maji, E. Rahtu, J. Kannala, M. B. Blaschko, and A. Vedaldi, “Fine-grained visual classification of aircraft,” *CoRR*, vol. abs/1306.5151, 2013.

[3] J. Krause, M. Stark, J. Deng, and F.-F. Li, “3D object representations for fine-grained categorization,” in *Proceedings of the International Conference on Computer Vision*, 2013.

[4] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based R-CNNfs for fine-grained category detection”,” in *Proceedings of the European Conference on Computer Vision*, 2014.

[5] H. Zheng, J. Fu, Z. Zha, and J. Luo, “Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition,” in *Proceedings of the Computer Vision and Pattern Recognition*, 2019.

[6] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, “Learning to navigate for fine-grained classification,” in *Proceedings of the European Conference on Computer Vision*, 2018.

[7] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, “Hierarchical bilinear pooling for fine-grained visual recognition,” in *Proceedings of the European Conference on Computer Vision*, 2018.

[8] H. Zheng, J. Fu, Z. Zha, and J. Luo, “Learning deep bilinear transformation for fine-grained image representation,” in *Advances in Neural Information Processing Systems*, 2019.

[9] R. Du, D. Chang, A. Kumar Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, “Fine-grained visual classification via progressive multi-granularity training of jigsaw patches,” in *Proceedings of the European Conference on Computer Vision*, 2020.

[10] M. Sun, Y. Yuan, F. Zhou, and E. Ding, “Multi-attention multi-class constraint for fine-grained image recognition,” in *Proceedings of the European Conference on Computer Vision*, 2018.

[11] 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 Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020.

[12] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the Computer Vision and Pattern Recognition*, 2018.

[13] H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning multi-attention convolutional neural network for fine-grained image recognition,” in *Proceedings of the International Conference on Computer Vision*, 2017.

[14] J. Fu, H. Zheng, and T. Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition,” in *Proceedings of the Computer Vision and Pattern Recognition*, 2017.

[15] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, “Learning rich part hierarchies with progressive attention networks for fine-grained image recognition,” *IEEE Transactions on Image Processing*, vol. 29, pp. 476–488, 2020.

[16] J. Woo, S.and Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision*, 2018.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the International Conference on Computer Vision*, 2017.

[18] Y. Wang, V. I. Morariu, and L. S. Davis, “Learning a discriminative filter bank within a CNN for fine-grained recognition,” in *Proceedings of the Computer Vision and Pattern Recognition*, 2018.

[19] Y. Chen, Y. Bai, W. Zhang, and T. Mei, “Destruction and construction learning for fine-grained image recognition,” in *Proceedings of the Computer Vision and Pattern Recognition*, 2019.

[20] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. Huang, “Attention convolutional binary neural tree for fine-grained visual categorization,” in *Proceedings of the Computer Vision and Pattern Recognition*, 2020.

[21] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, “Snapshot ensembles: Train 1, get M for free,” in *Proceedings of the International Conference on Learning Representations*, 2017.