Feature Channel Adaptive Enhancement for Fine-Grained Visual Classification

Dingzhou Xie, Cheng Pang(✉), Guanhua Wu, and Rushi Lan

Guilin University of Electronic Technology, Guilin 541004, China
pangcheng3@guet.edu.cn

Abstract. Fine-grained classification poses greater challenges compared to basic-level image classification due to the visually similar sub-species. To distinguish between confusing species, we introduce a novel framework based on feature channel adaptive enhancement and attention erasure. On one hand, a lightweight module employing both channel attention and spatial attention is designed, adaptively enhancing the feature expression of important areas and obtaining more discriminative feature vectors. On the other hand, we incorporate attention erasure methods that compel the network to concentrate on less prominent areas, thereby enhancing the network's robustness. Our method can be seamlessly integrated into various backbone networks. Finally, an evaluation of our approach is conducted across diverse public datasets, accompanied by a comprehensive comparative analysis against state-of-the-art methodologies. The experimental findings substantiate the efficacy and viability of our method in real-world scenarios, exemplifying noteworthy breakthroughs in intricate fine-grained classification endeavors.

Keywords: Fine-grained Visual Classification · Feature Channel Enhancement · Attention Erasure

1 Introduction

Fine-grained visual classification represents a substantial and intricate challenge within the realm of computer vision. Unlike traditional object classification tasks, fine-grained classification requires precise differentiation among objects with similar appearances but belonging to different subcategories. This task holds practical value in various domains such as animal recognition [20], plant classification [28], and car vehicle recognition [22]. However, As shown in Fig. 1, fine-grained classification presents a delicate balance between similarity and difference. Objects with similar appearances exhibit subtle local differences that often encompass crucial features determining their categories. Conventional classification methods struggle to capture these minute differences, resulting in poor performance. To overcome this challenge, extensive research has introduced innovative methods and techniques for fine-grained classification. Examples include local feature extraction [12,13,30], key part localization [6,24,36,37,40], metric learning [5,7,32,43], and attention-based mechanisms [19,26,41]. These approaches aim to extract discriminative features from local or

H. Lu et al. (Eds.): ACPR 2023, LNCS 14408, pp. 185–198, 2023.
https://doi.org/10.1007/978-3-031-47665-5_16

key details, thereby enhancing classification performance by learning subtle differences between categories. Despite the progress made in this area, fine-grained classification still faces several challenges. On one hand, accurate classifiers often require large amounts of fine-grained data and annotations [2,16,23,38] due to the nuanced differences among categories. This limits the scope and scalability of fine-grained classification methods. On the other hand, traditional fine-grained classification methods are often sensitive to image changes and noise, resulting in a decline in classification performance in complex scenes. Therefore, this paper proposes a comprehensive method named "Channel Feature Adaptive Enhancement". Our approach will focus on the extraction of local details and the learning of discriminative features, while exploring how to reduce intra-class variability and improve classifier robustness. At the outset, the extraction of features from both shallow and deep layers is facilitated through the utilization of a convolutional neural network (CNN). Our designed feature enhancement mechanism learns to emphasize key detail features, thereby improving classification performance. Additionally, an attention mechanism is introduced to highlight regions with significant fine-grained differences. This combination of feature augmentation and attention enables us to capture subtle differences more accurately and achieves significant performance gains in fine-grained classification tasks. Secondly, to address confusion among objects with similar appearances, we employ attention erasure. This technique erases highly discriminative areas, reducing intra-class differences, weakening features with high similarity to other categories, and encouraging the network to learn from previously unattended areas. This approach enhances classification accuracy and robustness, effectively reducing the risk of misclassification. In summary, the following constitute the principal contributions of this work:

1. We propose a feature enhancement module that accurately enhances discriminative regions in images.
2. We improve the channel attention mechanism by doubling the value of channels greater than the mean value calculated from the feature map after global average pooling. This assigns higher scores to important channels.
3. We introduce an informative mining module that masks a strong feature and facilitates the learning of complementary features.

2 Related Work

In the realm of fine-grained visual classification tasks, researchers have introduced various methods to enhance, suppress, and diversify features, aiming to improve classification performance. We describe several commonly employed techniques:

2.1 Local Feature Enhancement

Local feature enhancement methods aim to extract crucial local information from objects to enhance classification performance. For instance, Spatial Transformer

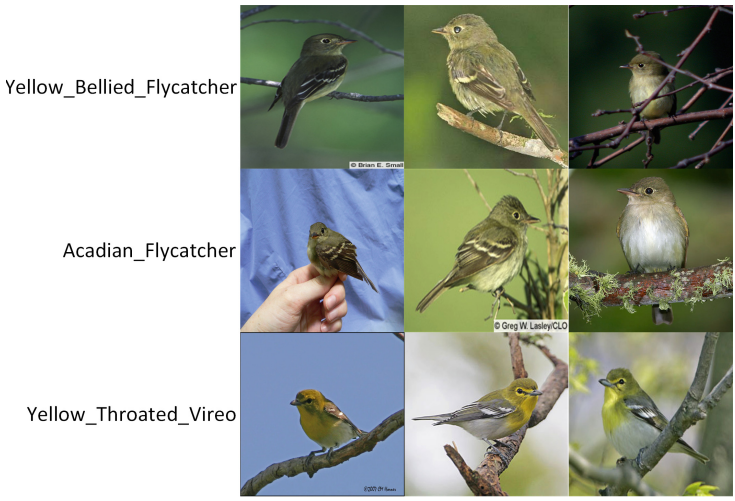

Fig. 1. In FGVC images, the same subclass exhibits significant appearance variations, whereas different categories demonstrate subtle appearance disparities.

Networks (STN) [18] uses spatial transformer modules to adaptively learn local regions of interest of images to improve the representation ability of key local features. In addition, Part-based Convolutional Neural Networks (PCNN) [29] divides the image into different parts, and classifies each part independently, and then combines the classification results of each part.

2.2 Global Feature Enhancement

Global feature augmentation methods concentrate on capturing the overall feature representation of the entire object. For example, Multi-scale CNNs [35] employ convolution kernels of varying scales to extract feature representations at different levels. Additionally, Spatial Pyramid Pooling (SPP) [9] divides the image into multiple spatial levels and performs pooling operations on the features within each level, allowing the capture of global information at different scales.

2.3 Attention Mechanism

The attention mechanism plays a pivotal role in fine-grained classification by automatically learning to focus on key regions or features. For instance, Squeeze-and-Excitation Networks (SENet) [11] employ a gating mechanism to adaptively adjust the importance of feature channels, thereby enhancing key features that contribute to classification. Moreover, the attention mechanism can generate region-specific attention heatmaps, offering interpretable explanations.

2.4 Feature Suppression

Feature suppression methods aim to reduce noise or irrelevant information that may interfere with classification tasks. For instance, DropBlock [8] randomly drops features in specific regions during training to minimize the impact of redundant information on classification. Additionally, Feature Dropout [31]enhances the model's learning by randomly discarding feature channels during training, promoting the exploration of other relevant features.

2.5 Feature Diversification

Feature diversification methods aim to generate multiple features with different transformations to enhance the robustness of classification models. For example, Cutout [3] randomly occludes a portion of the image, forcing the model to learn more resilient feature representations. Furthermore, the utilization of data augmentation techniques, including rotation, scaling, and translation, can facilitate the generation of a diverse array of features.

By comprehensively employing these methods, the performance of fine-grained visual classification tasks can be further enhanced, resulting in increased accuracy and reliability in classification results. However, several challenges, such as class imbalance, occlusion, and pose variation, remain unresolved and merit further exploration and resolution in future research.

3 Method

In this section, a comprehensive depiction of the proposed method is presented, where in Fig. 2 offers an illustrative overview of the framework. Our model consists of two lightweight modules: (1) Feature Channel Adaptive Augmentation module (FCAE), which focuses on learning multiple discriminative part-specific representations with maximum diversity. (2) Information Sufficient Mining Module (ISMM), which randomly erases highly discriminative components to guide the network in learning complementary information.

3.1 Feature Channel Adaptive Enhancement Module

The implementation of the FE method is shown in Fig. 3. The feature map $X_i \in R^{C\times W\times H}$, derived from the final three layers of the backbone network, with C, W, H, and i denoting the channel count, width, height, and layer index, respectively. Drawing inspiration from [30], we adopt a simple approach of taking the maximum value of each pixel along the channel dimension, resulting in $A_i \in R^{1\times W\times H}$:

$$A_i = \max\left(X_i\right) \in \mathrm{R}^{1\times \mathrm{W}\times \mathrm{H}} \tag{1}$$

Subsequently, we calculate a score for each pixel. Multiplying A_i by a hyper-parameter λ and adding it to X_i, we obtain a new feature map F_i:

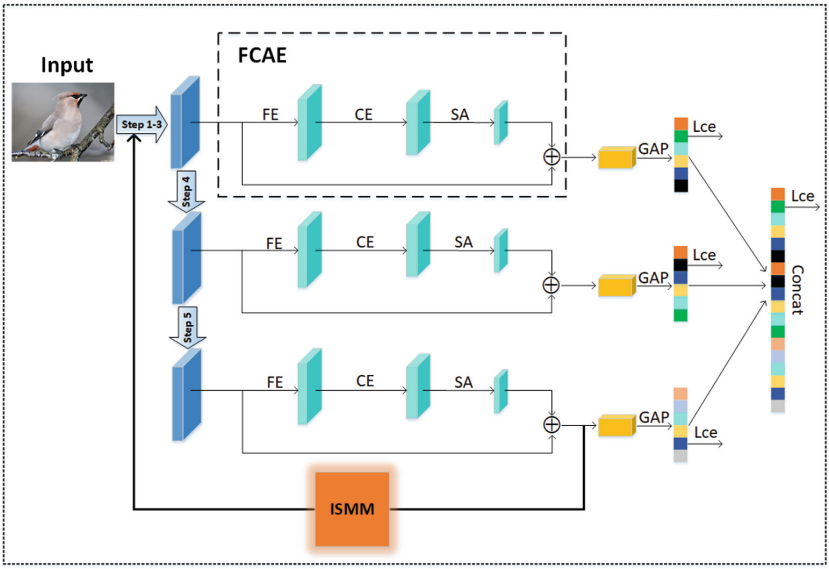

Fig. 2. The general framework of our model is composed of two major modules: (1) The Feature Channel Adaptive Augmentation module (2) The Information Sufficient Mining Module. (FCAE is comprised of three sub-methods, with the intricate process of ISMM illustrated in Fig. 4.)

$$F_i = ((softmax(A_i)) * X_i * \lambda + X_i) \tag{2}$$

At this stage, we aim to enhance globally relevant information. To emphasize discriminative local regions, we employ an attention mechanism inspired by [34]. The feature map F_i undergoes channel attention enhancement (CE), which determines the importance of each channel. We perform global average pooling on all channels and calculate the mean value across all channels.

$$mean = GAP\left(CA\left(F_i\right)\right) \tag{3}$$

If a specific channel's value exceeds the mean, it is doubled; otherwise, it remains unchanged. Upon the application of a sigmoid activation function, the resulting values are subjected to multiplication with F_i:

$$F_i = \begin{cases} 2 * F_i, if F_i > mean \\ F_i, \quad otherwise \end{cases} \tag{4}$$

Subsequently, the spatial attention (SA) mechanism is employed to capture relationships between different positions in the feature map, assigning varying weights to each position.

$$F_i = \left(SA\left(F_i\right)\right) + F_i \tag{5}$$

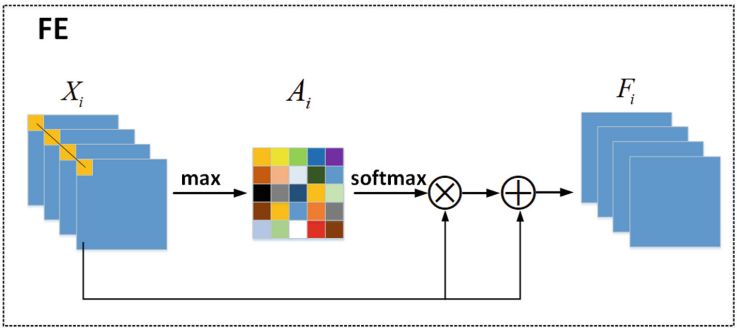

Fig. 3. The feature map X_i, generated by the backbone network, undergoes a process where the maximum value is extracted from the channel dimension to derive A_i. Subsequently, the score of each pixel value is computed and added to X_i, resulting in the feature map F_i after global feature enhancement.

3.2 Information Sufficient Mining Module

The implementation of the ISMM method is shown in Fig. 4. In order to further mine information and utilize attention resources, we take the last layer feature map of the backbone through the FCAE attention map as the attention activation map to extract the key information in the feature map. To fuse the features, we utilize a 1×1 convolution. Subsequently, the attention weight for each group of feature map channels is calculated, we employ global average pooling to select the top k sheets (where k is equal to the batch size). Randomly choosing one image, we apply bilinear interpolation [21] to upsample it to the original image's size. Subsequently, the introduction of a random threshold enables the classification of pixel values, whereby values below the threshold are identified as 0 and those surpassing the threshold as 1. Ultimately, an element-wise multiplication is executed between the acquired mask and the original image, resulting in the creation of the erased image. By employing attention erasure, we can filter out important stimuli that have already been learned and focus on less significant stimuli that may not have been well-learned. By adopting this approach, the efficiency of information processing is effectively heightened, while concurrently optimizing the allocation and utilization of attention resources.

3.3 Network Design

Various convolutional neural network structures can readily incorporate our method. ResNet [10] serves as an example, with its feature extraction process comprising five stages, where the spatial size of the feature map is halved after each stage. Given the abundance of semantic information within the deep feature maps, we opt to insert the Feature Channel Adaptive Enhancement module (FCAE) at the conclusion of the third, fourth, and fifth stages. Our method is very flexible and can adapt to classification tasks of different granularities by

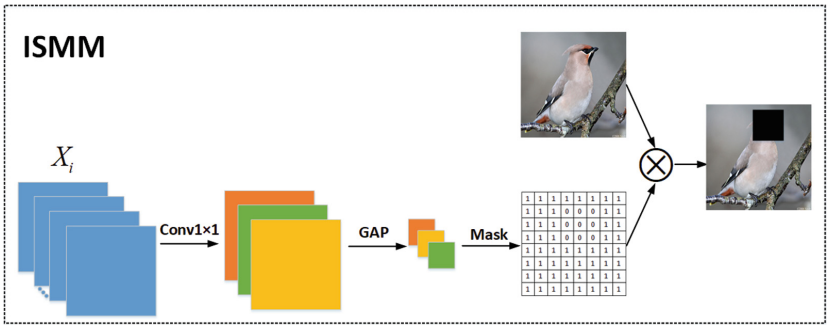

Fig. 4. The attention map is reduced in dimensionality through conv1×1, followed by global average pooling along the channel dimension. A mask is then randomly selected and multiplied with the original image.

directly adjusting the number of FCAEs, and can be customized according to the needs of specific tasks.

4 Experiments

In this section, the performance of the proposed method is evaluated across three fine-grained classification datasets: Caltech UCSD-Birds (CUB) [20], Stanford Cars (CAR) [22], and FGVC-Aircraft (AIR) [27], with detailed information provided in Table 1. The implementation details of our method are extensively discussed in Sect. 4.1, while Sect. 4.2 presents a comparative analysis of our method's accuracy performance against state-of-the-art approaches, showcasing its competitive advantage. Sections 4.3 and 4.4 encompass comprehensive ablation experiments and visualization analyses, aiming to highlight the distinctive features and efficacy of our method. By presenting both experimental results and visualizations, a comprehensive understanding of the robustness and efficiency of our method is provided. Through meticulous evaluations and extensive analyses, the superior performance of our method in fine-grained classification tasks is conclusively demonstrated.

Table 1. The statistics information of the three widely used Fine-Grained Visual Categorization datasets.

Dataset	Catagory	Train	Test
CUB-200-2011 [20]	200	5994	5974
FGVC-Aircraft [27]	100	6667	3333
Stanford Cars [22]	196	8144	8041

4.1 Implementation Details

All experiments were conducted using PyTorch [17], with a version higher than 1.12, on an NVIDIA A100 GPU. The performance evaluation of our method employed the widely adopted ResNet50 as the backbone network, which had been pretrained on the ImageNet dataset. The incorporation of the FCAE module was performed at the conclusion of phases 3, 4, and 5 in our methodology, while the integration of the ISMM module occurred at the termination of phase five. During the training process, the input images were resized to dimensions of 550×550 and subsequently randomly cropped to dimensions of 448×448 to augment the training set. Furthermore, random horizontal flipping was implemented as an auxiliary augmentation technique. For testing purposes, the input images were resized to 550×550 and cropped from the center to dimensions of 448×448. In our approach, a hyperparameter λ=0.5 was set. To optimize the model, we utilized the stochastic gradient descent (SGD) optimizer with a momentum of 0.9. The model underwent training for 300 epochs, employing a weight decay of 0.00001 and a batch size of 16. The learning rate of the backbone layer was set to 0.002, while the learning rate of the new layer was set to 0.02. We adjusted the learning rate using a cosine annealing scheduler [25].

4.2 Compared with State-of-the-Art Methods

Table 2 showcases a comparative analysis between our method and recent fine-grained classification approaches across the CUB 2002011, Stanford Cars, and FGVC-Aircraft datasets. DeepLAC [23] and Part-RCNN [38] are both component positioning-based methods. DeepLAC [23] initially employs a convolutional neural network to extract image features, followed by the introduction of a local localization module to identify important regions within object images. Part-RCNN [38] first generates candidate object areas in the image using the candidate frame generation algorithm. Subsequently, for each candidate box, the object is decomposed into parts, and features are extracted for each part. Among the attention-based methods, RA-CNN [6], MA-CNN [40], API-Net [43], PCA [39], AC-Net [19], and AKEN [14] are noteworthy. RA-CNN [6] incorporates a circular attention mechanism, allowing the network to dynamically focus attention on the intricate details crucial for fine-grained classification. MA-CNN [40] integrates multiple attention modules, each responsible for producing an attention map for a distinct region of the image, directing the network's focus toward regions with higher discriminative properties. API-Net [43] introduces a pairing interaction mechanism, where selected feature regions undergo pairwise interactions with each other. PCA [39] leverages a co-attention mechanism to learn the association and importance between different regions in an image. The co-attention network consists of a local feature extraction module and a progressive attention module. AC-Net [19] employs an attention mechanism to select and weigh key regions in feature representations. By incorporating a convolutional binary neural tree structure, the model sequentially performs feature selection and learning, progressively focusing on meaningful areas from a

global to local perspective. AKEN [14] integrates attention mechanisms and kernel encoding to extract and encode fine-grained features. Attention mechanisms select and weigh important features, aiding in distinguishing between different fine-grained categories. Kernel encoding captures specific feature information by encoding selected features. Guided Zoom [1] and S3N [4] fall under the category of local area amplification methods. Guided Zoom [1] utilizes the interplay of local area amplification and model feedback to acquire insights into model decision-making and confidence. This information serves as guidance for improved decision-making. S3N [4] employs an adaptive sparse sampling strategy to selectively sample local regions within the image. HOI [33] proposes a high-level interaction method, where the high-level interaction module leverages associations between different features in the image to enhance and integrate features. Specific weights and combinations are learned to reinforce and highlight features relevant to fine-grained classification tasks. SPS [15] randomly selects a subset of training samples and exchanges them with other samples, generating new sample pairs. CIN [7] introduces a channel interaction module that facilitates interaction and information transfer between channels. This module utilizes convolution operations and attention mechanisms to enable information exchange and joint feature learning across channels. LIO [42] utilizes unlabeled data for self-supervised learning. By designing self-supervised tasks, the model learns structural information about objects. Each of the above methods possesses its unique advantages and characteristics. However, our proposed method exhibits superior performance across the three datasets, which can be attributed to the exceptional components we have introduced.

Table 2. Comparison results on CUB-200-2011, FGVC-Aircraft and Stanford Cars datasets. "-" means no data

]Method	Backbone	CUB-200-2011	FGVC-Aircraft	Stanford Cars
DeepLAC [23]	VGG	80.3	–	–
Part-RCNN [38]	VGG	81.6	–	–
RA-CNN [6]	VGG	85.3	88.1	92.5
MA-CNN [40]	VGG	86.5	89.9	92.8
SPS [15]	ResNet50	87.3	92.3	94.4
API-Net [43]	ResNet50	87.7	93.0	94.8
HOI [33]	ResNet50	89.5	92.8	**95.3**
PCA [39]	ResNet50	88.3	92.4	94.3
AC-Net [19]	ResNet50	88.1	92.4	94.6
Guided Zoom [1]	ResNet50	87.7	90.7	93.0
CIN [7]	ResNet50	87.5	92.6	94.1
LIO [42]	ResNet50	88.0	92.7	94.5
S3N [4]	ResNet50	88.5	92.8	94.7
FBSD [30]	ResNet50	89.3	92.7	94.4
AKEN [14]	ResNet50	86.2	93.3	92.6
Ours	ResNet50	**89.6**	**93.3**	95.1

4.3 Ablation Experiment

A series of ablation experiments were conducted to evaluate the effectiveness of each module for fine-grained classification. Initially, we conducted classification experiments using a baseline model that solely comprised the backbone network. Subsequently, we gradually introduced the two modules we designed and integrated them with the backbone network individually. Finally, we compared the performance of using a single module versus using both modules simultaneously. The experimental results are presented in Table 3. When employing only the backbone network, the classification accuracy achieved a benchmark level. Upon introducing the FCAE module, the accuracy improved by 2.7% on the bird dataset1.8% on the airplane dataset, and 4.4% on the Stanford car dataset. The introduction of the ISMM module resulted in respective accuracy improvements of 3%, 1.3%, and 4.8% on the three datasets. Remarkably, the best classification results were obtained when both modules were introduced simultaneously, resulting in an improvement of 4.1%, 3%, and 5.3% compared to the baseline. Utilizing the combination of both modules yielded the highest accuracy, significantly outperforming the usage of a single module.

Table 3. Ablation Study on Three Benchmark Datasets

Method	CUB-200-2011	FGVC-Aircraft	Stanford Cars
Resnet50	85.5	90.3	89.8
Resnet50+FCAE	88.2	92.1	94.2
Resnet50+ISMM	88.5	91.6	94.6
Resnet50+FCAE+ISMM	89.6	93.3	95.1

4.4 Visualization

To visually showcase the effectiveness of our proposed method, the utilization of GradCAM is employed. As depicted in Table 4, when compared to the baseline, a tendency of the baseline to learn global features is observed. In scenarios where the target and background exhibit similarity, the baseline is prone to capturing background noise. Conversely, our method progressively captures global information during the initial stages, which encompasses the assimilation of learned background noise. However, as the network delves deeper into subsequent stages, the learned features become more localized, resulting in precise and targeted focus on the target. The method exhibits remarkable consistency in addressing different categories, emphasizing its robustness and reliability in making classification decisions. This consistency highlights the network's sensitivity to capturing fine-grained features (Fig. 5).

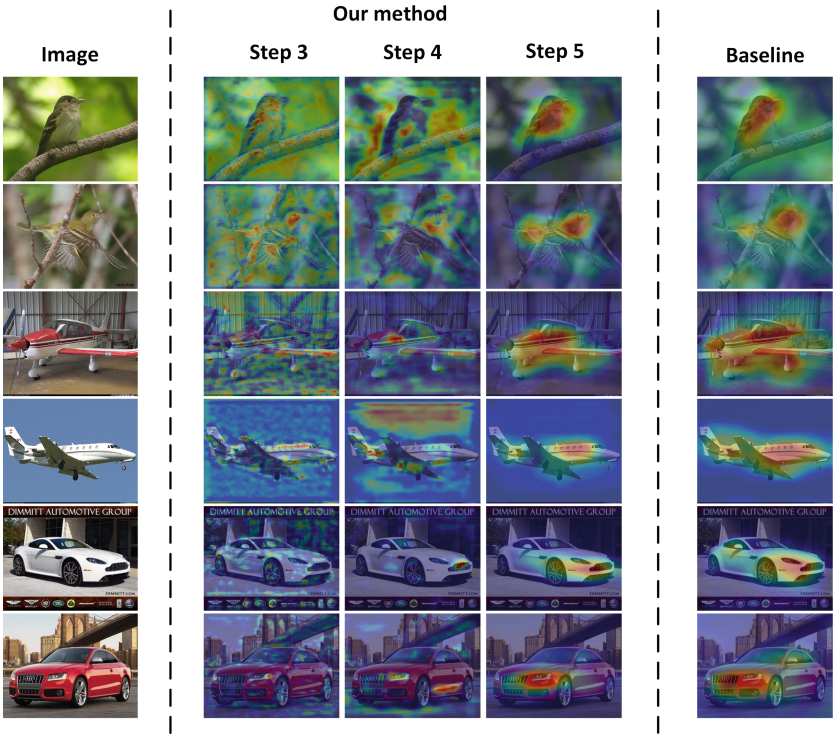

Fig. 5. Compared with the baseline, our method can better learn the discriminative local area, and can accurately distinguish the target and the background when the target is similar to the background.

5 Conclusion

In this study, we propose an approach for fine-grained classification tasks, aiming to improve classification performance by learning feature channel adaptive enhancement and information sufficient mining. First, the feature channel adaptive augmentation module aims to learn different discriminative part representations. Then use the information sufficient mining module to filter the learned important stimuli and focus on the less important stimuli. Our proposed method based on feature channel adaptive augmentation has achieved remarkable progress in meeting the challenges of fine-grained classification tasks. By enhancing useful global information, mining key information, and utilizing attention resources, our method is able to improve classification performance and adapt to different task demands.

Acknowledgements. This work is partially supported by the Guangxi Science and Technology Project (2021GXNSFBA220035, AD20159034), the Open Funds from Guilin University of Electronic Technology, Guangxi Key Laboratory of Image and

Graphic Intelligent Processing (GIIP2208) and the National Natural Science Foundation of China (61962014).

References

1. Bargal, S.A., et al.: Guided zoom: zooming into network evidence to refine fine-grained model decisions. IEEE Trans. Pattern Anal. Mach. Intell. **43**(11), 4196–4202 (2021)
2. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952 (2014)
3. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
4. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6599–6608 (2019)
5. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N.: Pairwise confusion for fine-grained visual classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 70–86 (2018)
6. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4438–4446 (2017)
7. Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.: Channel interaction networks for fine-grained image categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10818–10825 (2020)
8. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: a regularization method for convolutional networks. Adv. Neural Inf. Process. Syst. **31** (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
12. Hu, T., Qi, H., Huang, Q., Lu, Y.: See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891 (2019)
13. Hu, T., Xu, J., Huang, C., Qi, H., Huang, Q., Lu, Y.: Weakly supervised bilinear attention network for fine-grained visual classification. arXiv preprint arXiv:1808.02152 (2018)
14. Hu, Y., Yang, Y., Zhang, J., Cao, X., Zhen, X.: Attentional kernel encoding networks for fine-grained visual categorization. IEEE Trans. Circuits Syst. Video Technol. **31**(1), 301–314 (2020)
15. Huang, S., Wang, X., Tao, D.: Stochastic partial swap: enhanced model generalization and interpretability for fine-grained recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 620–629 (2021)

16. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked CNN for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1173–1182 (2016)
17. Imambi, S., Prakash, K.B., Kanagachidambaresan, G.: Pytorch. Programming with TensorFlow: Solution for Edge Computing Applications, pp. 87–104 (2021)
18. Jaderberg, M., et al.: Spatial transformer networks. Adv. Neural Inf. Process. Syst. **28** (2015)
19. Ji, R., et al.: Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10468–10477 (2020)
20. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: stanford dogs. In: Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), vol. 2. Citeseer (2011)
21. Kirkland, E.J., Kirkland, E.J.: Bilinear interpolation. In: Advanced Computing in Electron Microscopy, pp. 261–263 (2010)
22. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013)
23. Lin, D., Shen, X., Lu, C., Jia, J.: Deep lac: deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1666–1674 (2015)
24. Liu, C., Xie, H., Zha, Z.J., Ma, L., Yu, L., Zhang, Y.: Filtration and distillation: enhancing region attention for fine-grained visual categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11555–11562 (2020)
25. Loshchilov, I., Hutter, F.: Sgdr: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
26. Luo, W., et al.: Cross-x learning for fine-grained visual categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8242–8251 (2019)
27. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
28. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729. IEEE (2008)
29. Ouyang, W., et al.: Deepid-net: object detection with deformable part based convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(7), 1320–1334 (2016)
30. Song, J., Yang, R.: Feature boosting, suppression, and diversification for fine-grained visual classification. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021)
31. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
32. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 805–821 (2018)
33. Wang, J., Li, N., Luo, Z., Zhong, Z., Li, S.: High-order-interaction for weakly supervised fine-grained visual categorization. Neurocomputing **464**, 27–36 (2021)
34. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

35. Yang, S., Ramanan, D.: Multi-scale recognition with dag-CNNs. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1215–1223 (2015)
36. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 420–435 (2018)
37. Zhang, L., Huang, S., Liu, W., Tao, D.: Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8331–8340 (2019)
38. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 834–849. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_54
39. Zhang, T., Chang, D., Ma, Z., Guo, J.: Progressive co-attention network for fine-grained visual classification. In: 2021 International Conference on Visual Communications and Image Processing (VCIP), pp. 1–5. IEEE (2021)
40. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217 (2017)
41. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5012–5021 (2019)
42. Zhou, M., Bai, Y., Zhang, W., Zhao, T., Mei, T.: Look-into-object: self-supervised structure modeling for object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11774–11783 (2020)
43. Zhuang, P., Wang, Y., Qiao, Y.: Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13130–13137 (2020)