Vol. 9 (2025): Proceedings of the 2025 International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability
(IC-AIMEES 2025)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.202510

# Existing Results and Recent Advancements of Fine-grained Image Classification

**Weisheng Kong***

*School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, Guangdong, China*

**Corresponding author: Weisheng Kong**

## Abstract

Fine-grained image classification emphasizes the resolution of recognizing the subtle differences between subclass objects from the same category. Its classic examples of application include the classification of bird species, car brands and types of plants. During recent years, with the rapid popularization of machine learning and computer vision, there have been growing real circumstances concerning the issue of Fine-grained image classification, which greatly boosts the relevant research. However, there are few reviews of the existing results and recent advancements of Fine-grained image classification. This paper will systematically conclude the major research focuses of fine-grained image classification, highlighting two major aspects. The first aspect is fine-grained image classification based on feature fusion theory. The second aspect is fine-grained image classification based on an attention mechanism. Then, this paper introduces the collaborative application of feature fusion and attention mechanisms in fine-grained image classification. Next, the paper lists the mainstream data collection of fine-grained image classification. Lastly, this paper summarizes the main problems in current research and proposes future directions of this issue.

## Keywords

fine-grained image classification, feature fusion, attention mechanism, deep learning

## 1. Introduction

The essential purpose of fine-grained image classification is to distinguish the subclass objects that have high visual similarity to each other. The challenge of the research is that the difference between subclass objects is unclear and such a difference is mainly represented as local features. Furthermore, the variability of multiple factors, like background, posture, and illumination, escalates the difficulty of recognition. Complex scenes, such as background interference, blocked objects, and changeable posture, also demand higher standards for the accuracy and robustness of the image classification algorithm.

Conventional image classification emphasizes the extraction of global features. However, in fine-grained image classification tasks, local features are more effective for solving the issue. During recent years, researchers have devoted themselves to exploiting effective feature learning strategies, especially self-adaptive attention mechanisms focusing on the extraction of local features and key regions of the image. The development of deep learning technologies, especially the deep neural network, gives rise to the realization of two major research focuses of fine-grained image classification, namely, the one based on feature fusion theory

and the other based on the attention mechanism.

Nevertheless, fine-grained image classification still faces many challenges. Models now are supposed to be able to precisely extract key features from interfered information. Also, the ability to achieve a decent generalization performance with limited quantities of samples now plays a greater role in judging the effectiveness of a certain model. Hence, the resolution of dynamic self-adaptive features catering to different environments and scenes is now one of the most important areas of future research.

This paper makes a systematic conclusion of the existing results and recent advancements of fine-grained image classification and focuses on the technology innovation that aims at complex scenes. The author analyzes the two major research focuses of fine-grained image classification, namely the one based on feature fusion and the other based on attention mechanisms. The author also introduces the combined application of the two methods and their advantages, proposing potential future research directions and improvement measures. Throughout the deep discussion of fine-grained image classification, this paper aims to provide a theoretical basis and practical guidance for the progress of this technology, with the aim of facilitating the development and deployment of fine-grained image classification.

## 2. Fine-Grained Image Classification Based on Feature Fusion

Feature fusion is a technique that merges feature vectors of different scales, enhancing network performance by splicing feature vectors in the same order (Li et al., 2024a). This technique is designed to combine features of multiple sources for promoting the accuracy and robustness of the model. The existing application of the feature fusion technique in fine-grained image classification mainly contains two directions, namely cross-layer feature fusion and cross-sample feature fusion.

As for cross-layer feature fusion, Xin et al. combined a residual network with feature fusion techniques and successfully improved classification results (Xin and Dou, 2023). They concatenated features from different layers, because in a convolution neural network, higher level features contain rich semantic features while having fuzzy local features, but lower level features are more generalized and do not contain sufficient semantic information. Hence, fusing information from different layers can effectively contribute to the improvement of model performance. The technique of feature fusion initially comes from the feature pyramid network(FPN), and was later applied to larger numbers of networks such as VGG and Densenet.

As for cross-sample feature fusion, it shares and optimizes feature representations across multiple samples to enhance the generalization ability of the models. In fine-grained classification tasks, due to significant variations in the pose, illumination, occlusion, etc. of subclass objects within the same category, the feature representation of a single sample is often insufficient to comprehensively describe all the details of the object. Therefore, cross-sample feature fusion methods enhance the model's adaptability to diverse scenarios by sharing feature information among different samples. Cross-sample contrastive learning learns more discriminative feature vectors by contrasting the feature representations of different samples. Li et.al proposed a Multi-task Prior-reinforced and Cross-sample Network (MPCNet) to model triplet recognition when conducting surgical action triplet recognition, and their model successfully learned cross-sample semantic relationships through shared keys and values, demonstrating state-of-the-art results of experiments based on CholecT50 dataset (Li et al., 2024b). Zhao et.al equipped a semantic-aligned fusion transformer with a vertical fusion module with a vertical fusion module for cross-scale feature fusion and a horizontal fusion module for cross-sample feature fusion, which broadens the vision for each feature point from the support to a whole augmented feature pyramid (Zhao et al., 2022). In their later tests of one-shot object detection and image classification, the mechanism is proven to have the ability to facilitate semantic-aligned associations, which stands for the boost of accuracy and effectiveness of image classification. In practical applications, cross-level and cross-sample feature fusion methods can often be combined to fully improve the model's fine-grained classification capability in complex scenarios.

## 3. Fine-Grained Image Classification Based on Attention Mechanism

The fundamental principle of the attention mechanism comes from the logic by which humans view images. When viewing images, humans first capture the overall information, and then focus more attention on a specific

area (Li, 2024). Human brains tend to selectively process and memorize the important contents and ignore the irrelevant parts. In deep learning tasks, the attention mechanism imitates such behaviors of human eyes to filter crucial features that decide the type that the image falls into from large quantities of information. Nowadays, in image classification, the attention mechanism is widely applied to assign varied weights to features. The features with higher weights attached will be prioritized, while features with lower weights will be less likely to be preferred by the attention mechanism.

Attention mechanism in fine-grained image classification is often divided into channel attention mechanisms and spatial attention mechanisms. In most research, the two attention mechanisms are used together. Usually, in a convolutional neural network, when images are processed by a convolutional layer, it forms a feature map. The feature map is described as a three-dimensional tensor, known as C(channel)×H(height)×W(width).

The channel attention mechanism assesses the importance of each channel and assigns higher weights to the important ones. Different channel stands for different features, for instance, edge, texture, color, curve and so on. So essentially, the channel attention mechanism works to capture the important features that better classify the images among a large number of average features. Meanwhile, the spatial attention mechanism pays attention to the different importance of locations. It assigns different attention or importance to information in different locations or regions of the feature map. Its concept can be summarized as follows: it firstly conducts global maximum pooling and global average pooling on a feature map in the form of C×W×H, which generates two feature maps of size H×W×1. Next, the two feature maps generated are integrated to form a feature map size of H×W×2. Subsequently, a convolution operation will be executed on the integrated feature map to produce a feature map size of H×W×1. Ultimately, the spatial attention weight matrix is derived using the Sigmoid activation function (Li, 2024). The matrix is in the form of H×W×1.

To conclude, the attention mechanism allows the model to concentrate on more distinctive localized areas within the image via dynamic weight distribution, rather than uniformly analyzing the entire image. This makes the model particularly effective for detecting nuanced variations in images when processing fine-grained classification tasks. In many research channels, attention mechanisms and spatial attention mechanisms are combined to attain better effects for models. Zhao et.al used a channel-space hybrid attention module(CBAM) in their research to generate feature maps and successfully obtained enhanced feature maps (Zhao and Chen, 2024). In their experiment part, the model significantly enhances the network's ability to capture discriminative features due to CBAM's characteristic of emphasizing important features and suppressing irrelevant ones across both channel and spatial dimensions, which improves recognition accuracy to 89.1%. Xie et.al also proposed a self-attention mechanism for fine-grained classification problems by embedding the Squeeze-and-Excite module in the MobileNet architecture, which effectively improves the feature representation capability of convolutional neural networks (Xie and Luo, 2025). Compared with traditional CNN networks, there is a 3.92% increase in accuracy on the dataset. Many other studies also demonstrate the effectiveness of attention mechanisms' ability to improve feature representation capability of convolutional neural networks, contributing hugely to local feature extraction and differential expression amplification, which is essential in fine-grained image classification problems.

## 4. Combined Application of Feature Fusion And Attention Mechanism in Fine-Grained Image Classification

Feature fusion targets at integrating features from either different layers or different samples. The reason is that features from different layers vary greatly in semantic and local information, while features from multiple samples contain varied information of pose, illumination, or occlusion. Thus, integrated features can effectively improve the generalization ability of the model. As for the attention mechanism, it prioritizes the features that weigh more than others when classifying the images. Hence, images may vary greatly in many complex factors like pose, illumination, or occlusion, but the attention mechanism selects the essential features that truly take effect among a large number of average features, which greatly enhances the accuracy of fine-grained image classification models. To conclude, both techniques work under the same logic, which is enhancing the effectiveness of the features extracted as much as possible.

In recent research results, the author noticed that the combination of feature fusion and attention mechanisms in fine-grained image classification tasks has become prevalent. Li's research on shark fine-grained image classification proposed a network based on both feature fusion and attention mechanism (Li, 2024). He input images into two improved deep learning models for feature extraction. The first model, Resnet50, is adjusted and inserted with an attention mechanism to improve the model's ability to acquire intricate characteristics of sharks. The second model Resnet34 has fewer layers and is in charge of the extraction of low-level and mid-level features. Then, the features extracted by the two models are integrated to fully utilize the advantages of different models. The fused features are used to create the ultimate classification layer to classify the types of sharks. The model's classification accuracy reaches 90.5% and is obviously better than other models. Many other researchers also demonstrated the value of the combination of feature fusion and attention mechanism techniques.

## 5. Dataset Collection

Commonly used datasets of fine-grained image classification include the following：

Caltech-UCSD Birds-200-2011(CUB-200-2011): This dataset contains 11,788 images of 200 North American bird species. Each image is richly annotated with the species, bounding boxes for body parts, and attributes (Wah et al., 2011).

Stanford Cars: This dataset contains 16,185 images of 196 classes of cars. It provides labels for the car model, brand, and manufacturer (Krause et al., 2013). The images are diverse, covering various viewpoints and lighting conditions.

Stanford Dogs: This dataset focuses on fine-grained image classification of dog breeds. It comprises 20,580 images across 120 different dog breeds. Each image is annotated with the dog breed label and detailed breed information. The images are sourced from a variety of contexts, covering diverse poses, scenes, and shooting angles.

Oxford 102 Flower: This dataset contains 8,189 images of 102 flower categories. The images are rich in color and texture diversity, making them suitable for researching the role of visual characteristics in fine-grained image classification (Nilsback and Zisserman, 2008).

## 6. Limitations and Future Research Directions

When processing complex scenarios, relying solely on a single modality (e.g., RGB images) may be insufficient for effectively classifying fine-grained subclass objects. Future research could explore the fusion of multi-modal data, such as combining RGB images, depth information, infrared images, or other sensor data. This could help models better distinguish targets within complex backgrounds.

Fine-grained image classification tasks typically require large amounts of labeled data, but annotating such data is extremely time-consuming and expensive. Future research may increasingly investigate weakly-supervised learning methods to automatically learn effective fine-grained features with only limited labels or even in the absence of labels.

Future research of fine-grained image classification can be coupled with large pre-trained models(e.g., Vision Transformers). Future research can leverage pre-trained knowledge and these powerful models as a robust feature extractor or for full fine-tuning, transferring their broad visual knowledge to the specific fine-grained domain (Xu et al., 2024). Future research can also explore efficient adaptation techniques like prompt tuning or adapter layers to customize large models for fine-grained image classification without full retraining.

## 7. Conclusion

As a notable topic of computer vision and deep learning, fine-grained image classification has made significant strides and advancements in recent years. This paper concluded the results of its research and proposed its potential future research focuses. Categories in fine-grained imagery display subtle differentiations, contingent upon localized attributes and global structural dependencies. The reason for feature

fusion and attention mechanism techniques becoming major methods in fine-grained image classification is that both techniques can dynamically prioritize critical regions of images, which concurrently model the interdependencies between localized features and holistic structural configurations. The existing techniques of fine-grained image classification still face lots of challenges, like heavy reliance on extensive annotation, limited generalization and robustness, limited model interpretability and trustworthiness, and dependency on a single modality(RGB). Future research will continue to center on key issues such as computational complexity, generalization ability and data scarcity. Meanwhile, the current methods of fine-grained image classification usually require a large amount of data with labels, while the process of labeling data is time-consuming and expensive. Future research may increasingly explore weakly-supervised or self-supervised learning methods to automatically learn effective fine-grained features with a limited number of labels or even without labels. Future studies can also investigate multi-level structured classification models, which perform classification progressively from coarse to fine levels to enhance robustness and accuracy. Additionally, model performance can be improved through techniques such as adversarial training and noise injection. With the advent of the era of large models, it is possible to leverage these models for pre-training, harnessing their potential in the field of fine-grained image classification. With the optimization of the existing methods and the introduction of new learning strategies, fine-grained image classification will have broader application and brighter prospects.

## References

Krause, J., Stark, M., Deng, J. and Fei-Fei, L., (2013). Published. 3d object representations for fine-grained categorization. *Proceedings of the IEEE international conference on computer vision workshops*, Piscataway: NJ. IEEE, pp. 554-561.

Li, J., Yang, Y., Long, H. and Qiu, S., (2024a). Published. A fine-grained image classification method utilizing the Transformer's attention mechanism combined with feature fusion. *2024 12th International Conference on Information Systems and Computing Technology (ISCTech)*, Xi'an, China. IEEE, pp. 1-7.

Li, Y., (2024). Published. Shark Fine-Grained Image Classification Based on Feature Fusion and Attention Mechanism. *2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV)*, Zhenjiang, China. IEEE, pp. 171-175.

Li, Y., Zhao, Z. and Li, R., (2024b). Published. Surgical Action Triplet Recognition by Using A Multi-Task Prior-Reinforced and Cross-Sample Network. *2024 6th International Conference on Electronic Engineering and Informatics (EEI)*, Chongqing, China. IEEE, pp. 1283-1289.

Nilsback, M. E. and Zisserman, A., (2008). Published. Automated Flower Classification over a Large Number of Classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, Bhubaneswar, India. IEEE, pp. 722-729.

Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S., (2011). *The caltech-ucsd birds-200-2011 dataset,* Pasadena: California Institute of Technology.

Xie, B. and Luo, M., (2025). Published. Fine-grained Image Classification Algorithm Based on Channel Attention Enhancement. *2025 IEEE 34th Wireless and Optical Communications Conference (WOCC)*, Taipa, Macao. IEEE, pp. 179-183.

Xin, H. and Dou, R., (2023). Published. Research on Fine-Grained Image Classification with Residual Network and Feature Fusion. *2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS)*, Changchun, China. IEEE, pp. 43-47.

Xu, Y., Wu, S., Wang, B., Yang, M., Wu, Z., Yao, Y. and Wei, Z., (2024). Two-stage fine-grained image classification model based on multi-granularity feature fusion. *Pattern Recognition,* vol. 146, p. 110042.

Zhao, X. and Chen, C., (2024). Published. Multi-Scale Localization and Attention Mechanism for Fine-Grained Image Classification. *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, Shenzhen, China. IEEE, pp. 1173-1178.

Zhao, Y., Guo, X. and Lu, Y., (2022). Published. Semantic-aligned fusion transformer for one-shot object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA,. IEEE, pp. 7601-7611.

## Funding

This research received no external funding.

## Conflicts of Interest

The authors declare no conflict of interest.

## Acknowledgment

This paper is an output of the science project.

## Open Access