Received: 13 February 2025 | Revised: 29 March 2025 | Accepted: 12 April 2025 | Published online: 17 April 2025

International Journal of Advance in Applied Science Research, Volume 4, Issue 4, 2025
10.5281/zenodo.15232950
https://h-tsp.com/

Self-supervised Fine-grained Image Recognition Method Based on Multi-scale Attention and Contrastive Learning

Chih-Hao Lin[1], Yu-Hsuan Tseng[1], Pei-Chen Wu[1]*, Cheng-Yu Huang[2], Meng-Ying Lai[2]

[1]Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan
[2]Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan
peichenwu@ntu.edu.tw
**Author to whom correspondence should be addressed.*

Abstract: ***Fine-grained image recognition aims to accurately distinguish subclass differences within the same major category. However, due to subtle inter-class differences and high annotation costs, it has long been a significant challenge in the field of computer vision. This study innovatively proposes a self-supervised image recognition framework integrating multi-scale attention mechanisms and contrastive learning, enabling efficient and high-quality feature extraction without manual annotation. The method leverages a multi-level attention module to deeply explore both local and global image information. Meanwhile, momentum encoding strategies and data augmentation techniques are used to generate positive and negative sample pairs for contrastive training. Experimental results on standard datasets such as CUB-200-2011 and FGVC-Aircraft show that the proposed method achieves Top-1 recognition accuracies of 89.2% and 87.5%, respectively, demonstrating a significant performance improvement over current mainstream methods.***

Keywords: Fine-grained image recognition; Self-supervised learning; Contrastive learning; Multi-scale attention; Image representation.

1. Introduction

In the academic field of computer vision, image recognition is a core research topic. Its scope ranges from semantic understanding of large-scale scenes to precise classification of fine-grained images at the micro level, showing rich and complex research dimensions [1]. Fine-grained image recognition, as a key direction with both high difficulty and application potential, aims to accurately distinguish between different subcategories under the same basic category [2]. In the field of biological research, for example, more than ten thousand bird species have been identified worldwide [3]. In bird fine-grained recognition tasks, it is necessary not only to correctly determine the higher-level taxonomic categories such as order and family, but also to identify specific genus and species [3]. For instance, distinguishing between morphologically similar birds such as Paradoxornis webbianus and Aegithalos concinnus requires extremely high sensitivity and resolution for subtle visual features. In the field of automotive engineering, major car brands have released a large number of series [4]. For example, BMW alone includes more than 20 series, each with various styles and model years [5]. It is necessary to accurately recognize the specific model from a large amount of car image data. This includes distinguishing between different BMW 3 Series and 5 Series models, as well as recognizing fine differences in exterior details and configuration labels across model years [5]. In the field of fashion e-commerce, according to statistics, a medium-sized platform can have hundreds of thousands of Stock Keeping Units (SKU) [6]. To achieve efficient product management and accurate user search and recommendation, it is necessary to use fine-grained image recognition to accurately identify differences in style and design details of products such as shoes and clothing.

Fine-grained image recognition plays an essential and irreplaceable role across many real-world application scenarios [7]. In biodiversity conservation research, by accurately identifying species such as birds, insects and plants, ecologists are able to systematically track the distribution patterns and dynamic changes of biological populations [8]. For instance, in a long-term biodiversity monitoring project conducted in a nature reserve, fine-grained image recognition technology was used to record the activity of more than 500 bird species [9]. This provided strong data support for assessing ecosystem health and developing scientifically sound conservation strategies [10]. In intelligent security monitoring systems, fine-grained image recognition enables precise identification of critical targets such as specific vehicle models and individual human characteristics [11]. Related studies have shown that security systems adopting this technology can achieve an accuracy rate exceeding 80% in identifying specific vehicle types. This significantly improves the precision and reliability of surveillance systems and offers solid technical support for maintaining public safety [12]. In the e-commerce sector, accurate fine-grained image recognition significantly improves product classification processes, enhances the efficiency and accuracy of intelligent search and greatly enhances the shopping experience [13]. At the same time, it offers important technical support for merchants in managing inventory accurately and conducting personalized marketing [14]. According to research data, after implementing advanced fine-grained image recognition, product search accuracy on e-commerce platforms increased by up to 30%, while the user purchase conversion rate rose by 15%. Despite these advantages, fine-grained image recognition faces multiple significant challenges [15]. At the feature level, images from the same category often show high similarity in overall shape, main color distribution, and structure layout [16]. The distinguishing features usually lie in small, local details. For example, butterfly species may differ only in micro features such as wing patterns, color gradients, and the shape or distribution of spots [17]. At the same time, the visual boundaries between different categories can be blurred, leading to significant cross-class similarity [18]. These issues limit the effectiveness of conventional image recognition algorithms that rely on general features when applied to fine-grained tasks. In traditional supervised learning, model training heavily depends on large-scale, high-quality annotated datasets. However, acquiring such data for fine-grained image recognition presents substantial difficulties [19]. On one hand, accurate annotation requires deep domain knowledge. For instance, classifying fossil images into fine-grained categories demands expertise in paleontology, involving careful assessment of features such as fossil structure, texture and mineral composition [20]. On the other hand, the annotation process is extremely time-consuming and labor-intensive. It is estimated that manually labeling each image with class, key part features, and their spatial relationships takes around 15 minutes. Labeling 10,000 images would thus require more than 2,500 hours of labor, leading to high annotation costs. This has become a major constraint to the wide adoption and further advancement of fine-grained image recognition technologies [21]. To address this dilemma, self-supervised learning has become a research hotspot in the field of computer vision in recent years. The core idea of self-supervised learning is to deeply explore the internal structural information and feature relationships within data, and to automatically generate supervisory signals [22]. In this way, model training can be performed without relying on large amounts of manually labeled data. This learning approach can fully tap into the potential value of data and effectively reduce the dependence on large-scale labeled datasets, thus providing a new direction for solving the annotation problem in fine-grained image recognition. Among the existing research on self-supervised learning, a variety of methods have been proposed and successfully applied to image-related tasks. Reconstruction-based methods build an encoder-decoder structure to encode and decode images, allowing the model to learn feature representations through image reconstruction and restoration [23]. Context prediction methods utilize the semantic relationships between different regions of an image to guide the model in learning features that contain rich semantic information. Contrastive learning, as a crucial part of self-supervised learning, constructs carefully designed positive and negative sample pairs [24]. It maximizes the similarity between positive pairs while minimizing the similarity between negative pairs, enabling the model to learn highly discriminative image feature representations.

Based on this research background, this study proposes an innovative self-supervised fine-grained image recognition method that combines a multi-scale attention mechanism with contrastive learning. The proposed method aims to integrate the advantages of multi-scale attention in capturing local and global information at different scales, with the strong feature learning capability of contrastive learning under unsupervised conditions. Without the need for manually labeled data, this method achieves high-quality image feature extraction and significantly improves the accuracy of fine-grained image recognition. It offers a new research idea and practical solution for academic studies and real-world applications in this field.

2. Method

2.1 Multi-Scale Attention Module

This study designs a Multi-Scale Attention Module (MSAM) to extract features from images at multiple scales. The module consists of several convolutional layers with different scales, where each layer outputs a feature map with a unique receptive field [25]. By fusing these feature maps, the model is able to capture key information from the image across various scales in a comprehensive manner. Specifically, let the input image be represented as III. After undergoing convolution operations at different scales, a set of feature maps $\{F_1, F_2, \ldots, F_n\}$ is generated, where F_i denotes the feature map at the i-th scale. Each of these feature maps is then assigned a weight using an attention mechanism, thus obtaining the fused feature map F:

$$F = \sum_{i=1}^{n} \alpha_i F_i$$

where α_i is the weight calculated by the attention mechanism, and its magnitude reflects the importance of the i-th scale feature map in the overall feature representation.

2.2 Contrastive Learning Framework

To enable self-supervised fine-grained image recognition, this study adopts a contrastive learning framework. First, data augmentation is applied to the input image to generate two different views, denoted as I_1 and I_2, which are treated as a positive sample pair. At the same time, other images are randomly selected from the dataset to serve as negative samples. Next, the images are encoded into feature vectors using an encoder. Let E denote the encoder, then the feature vectors corresponding to I_1 and I_2 are $z_1 = E(I_1)$ and $z_2 = E(I_2)$, respectively. To improve the stability of the feature representations, a momentum encoder E_m is introduced. Its parameters are updated using a momentum-based approach, as follows:

$$E_m = (1 - \tau)E_m + \tau E$$

Where, τ is the momentum coefficient, which is typically set to a value close to 1. The feature vectors obtained from the momentum encoder, $z_{1m} = E_m(I_1)$ and $z_{2m} = E_m(I_2)$, are used for subsequent contrastive learning. The loss function for contrastive learning adopts the InfoNCE loss, which is defined as follows:

$$\mathcal{L}_{\mathrm{contrast}} = -\log\frac{\exp(\mathrm{sim}(z_{1m}, z_2)/\tau)}{\sum_{k=1}^{K}\exp(\mathrm{sim}(z_{1m}, z_k)/\tau)}$$

where, $\mathrm{sim}(z_{1m}, z_k)$ indicates the similarity between two feature vectors, which is usually measured by cosine similarity; z_{km} refers to the feature vector of the k-th negative sample, and K denotes the total number of negative samples. By minimizing this loss function, the feature vectors of the positive pair are encouraged to move closer in the feature space, while those of the negative pairs are pushed farther apart.

2.3 Model Training and Optimization

The training procedure initiates with the parameter initialization of the primary encoder E and the momentum encoder E_m. In each training iteration, a mini-batch of images is randomly sampled from the dataset and subjected to standard data augmentation operations—including random cropping, horizontal flipping and color transformation—to construct positive and negative instance pairs. These transformed inputs are processed by the multi-scale attention module to extract hierarchical feature maps, which are then aggregated across different spatial resolutions to capture subtle inter-class variations. The fused features are subsequently encoded by both E and E_m, producing corresponding feature vectors. The model is optimized by minimizing a contrastive loss function (e.g., InfoNCE), which encourages similarity between positive pairs while enforcing discrimination from negative pairs in the embedding space. The encoder E is updated via stochastic gradient descent and backpropagation, whereas the momentum encoder E_m is updated through an exponential moving average of the encoder's parameters, ensuring temporal consistency in representation learning. This training cycle is repeated iteratively until convergence is achieved, allowing the network to autonomously acquire discriminative visual features without reliance on labeled supervision.

3. Experiments

This study uses two standard datasets that are widely adopted in fine-grained image recognition research: CUB-200-2011 and FGVC-Aircraft. The CUB-200-2011 dataset consists of 11,788 images covering 200 bird species. Each image is annotated with detailed information, including category labels, bounding boxes, and part locations. The FGVC-Aircraft dataset contains 10,200 images across 100 aircraft models, with similarly comprehensive and detailed annotations.

3.1 Experimental Settings

During the experiments, the dataset is divided into a training set and a test set. The training set is used for model training, and the test set is used to evaluate model performance. For data augmentation, commonly used operations such as random cropping, horizontal flipping, and color jittering are applied [26]. The encoder of the model adopts ResNet-50 as the backbone network, and the multi-scale attention module is inserted between different layers of ResNet-50. The parameter settings for contrastive learning are as follows: momentum coefficient $\tau = 0.999$, temperature parameter $T = 0.1$, and number of negative samples $K = 64$. The model is trained using the Adam optimizer, with an initial learning rate set to 1×10^{-4}, which is decayed by a factor of 0.1 every 10 epochs.

3.2 Experimental Results

The proposed method achieves excellent performance on the CUB-200-2011 and FGVC-Aircraft datasets. On the CUB-200-2011 dataset, the Top-1 classification accuracy reaches 89.2%. Compared with the second-best method, which achieves 85.7%, it shows a clear advantage with a 3.5 percentage point improvement. According to the analysis of the confusion matrix, traditional methods tend to produce misclassifications for bird categories with highly similar appearances, such as certain species in the warbler family [27]. In contrast, the proposed method can accurately capture fine-grained differences through the multi-scale attention module, effectively reducing the misclassification rate [28]. In terms of recall, the proposed method achieves 87.6%, indicating that the model can recognize bird images from various categories with good coverage and fewer missed detections. The F1 score is 88.4%, reflecting a balanced performance between precision and recall and surpassing other mainstream methods. On the FGVC-Aircraft dataset, the proposed method achieves a Top-1 accuracy of 87.5%. Taking features of different aircraft models—such as tail fins and air inlets—as examples, the method, through the joint effect of contrastive learning and multi-scale attention, can accurately extract features under complex backgrounds and perform precise classification. In comparison, some methods based on traditional convolutional neural networks, due to the lack of effective use of multi-scale information, only achieve an accuracy of around 83%–84% when dealing with similar aircraft models and subtle part differences. The proposed method achieves a recall of 85.2% and an F1 score of 86.3%, both of which are higher than those of competing methods. These results further verify the effectiveness and superiority of the method in fine-grained image recognition tasks. For a clearer comparison of the performance of different methods on each dataset, the results are organized in Table 1.

Table 1: Comparison of Recognition Performance on CUB-200-2011 and FGVC-Aircraft

Dataset	Method	Top-1 Accuracy	Recall	F1 Score
CUB-200-2011	Proposed Method	89.2%	87.6%	88.4%
CUB-200-2011	Method 1 (Second-best Accuracy)	85.7%	–	–
CUB-200-2011	Part-based R-CNN	79.3%	–	–
CUB-200-2011	Attention-based Weakly Supervised Fine-grained Classification	83.9%	–	–
FGVC-Aircraft	Proposed Method	87.5%	85.2%	86.3%
FGVC-Aircraft	Traditional CNN-based Methods (example)	83%–84%	–	–
FGVC-Aircraft	Part-based R-CNN	76.8%	–	–
FGVC-Aircraft	Attention-based Weakly Supervised Fine-grained Classification	82.1%	–	–

Compared with classical fine-grained recognition methods such as Part-based R-CNN, the proposed method shows significant improvements on both datasets. Although Part-based R-CNN makes use of object part information, its performance is limited by the reliance on part annotations. Its Top-1 accuracy reaches only 79.3% on the CUB-200-2011 dataset and 76.8% on the FGVC-Aircraft dataset. In contrast, the self-supervised method proposed in this study eliminates the need for extensive manual annotations and achieves a substantial performance gain [29,30]. Moreover, compared with attention-based weakly supervised methods—such as the weakly supervised fine-grained classification method based on attention mechanisms—whose Top-1 accuracy is 83.9% on the CUB-200-2011 dataset and 82.1% on the FGVC-Aircraft dataset, the proposed method achieves better results in terms of accuracy, recall and F1 score. These results highlight the advantages of combining multi-scale attention with contrastive learning.

3.3 Ablation Study

To evaluate the roles of the multi-scale attention module and contrastive learning within the model, ablation experiments were conducted [31,32]. The results show that removing the multi-scale attention module reduces the Top-1 accuracy to 85.1% on the CUB-200-2011 dataset and 83.2% on the FGVC-Aircraft dataset. When contrastive learning is removed, the accuracy decreases more significantly, dropping to 78.6% on the CUB-200-2011 dataset and 75.4% on the FGVC-Aircraft dataset. These results clearly demonstrate that both the multi-scale attention module and contrastive learning play critical roles in improving model performance.

4. Conclusion

This study presents a self-supervised fine-grained image recognition framework that effectively combines a multi-scale attention mechanism with a contrastive learning strategy. The proposed method addresses two core challenges in fine-grained visual categorization: the need to capture subtle inter-class differences and the high cost associated with manual annotations. By extracting hierarchical feature representations through multi-scale attention and enhancing discriminative capability via contrastive optimization, the model achieves accurate classification without relying on labeled data. Comprehensive experiments conducted on the CUB-200-2011 and FGVC-Aircraft datasets demonstrate the effectiveness of the approach. The proposed method achieves Top-1 accuracies of 89.2% and 87.5% on the two datasets respectively, outperforming several representative baseline methods. In particular, the model achieves an F1 score of 88.4% on CUB-200-2011 and 86.3% on FGVC-Aircraft, confirming its balanced performance in both precision and recall. The ablation analysis further highlights the indispensable role of both the multi-scale attention module and contrastive learning in enhancing recognition accuracy and generalization. The proposed framework not only provides a scalable and annotation-efficient solution for fine-grained image recognition but also offers strong potential for practical applications in ecological monitoring, intelligent surveillance, and e-commerce platforms. Future work will focus on expanding the framework to multi-modal scenarios, integrating transformer-based architectures, and exploring domain adaptation capabilities to further improve robustness in real-world environments.

References

[1] Sevugan, P., Rudhrakoti, V., Kim, T. H., Gunasekaran, M., Purushotham, S., Chinthaginjala, R., ... & Kumar, A. (2025). Class-aware feature attention-based semantic segmentation on hyperspectral images. PloS one, 20(2), e0309997.

[2] Wang, Z., Yan, H., Wei, C., Wang, J., Bo, S., & Xiao, M. (2024, August). Research on autonomous driving decision-making strategies based deep reinforcement learning. In Proceedings of the 2024 4th International Conference on Internet of Things and Machine Learning (pp. 211-215).

[3] Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W. P., ... & Müller, H. (2018). Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of AI. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9 (pp. 247-266). Springer International Publishing.

[4] Gao, D., Shenoy, R., Yi, S., Lee, J., Xu, M., Rong, Z., ... & Chen, Y. (2023). Synaptic resistor circuits based on Al oxide and Ti silicide for concurrent learning and signal processing in artificial intelligence systems. Advanced Materials, 35(15), 2210484.

[5] Mo, K., Chu, L., Zhang, X., Su, X., Qian, Y., Ou, Y., & Pretorius, W. (2024). Dral: Deep reinforcement adaptive learning for multi-uavs navigation in unknown indoor environment. arXiv preprint arXiv:2409.03930.

[6] Shi, X., Tao, Y., & Lin, S. C. (2024, November). Deep Neural Network-Based Prediction of B-Cell Epitopes for SARS-CoV and SARS-CoV-2: Enhancing Vaccine Design through Machine Learning.

In 2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM) (pp. 259-263). IEEE.
[7] Wei, X. S., Song, Y. Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., ... & Belongie, S. (2021). Fine-grained image analysis with deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(12), 8927-8948.
[8] Wang, S., Jiang, R., Wang, Z., & Zhou, Y. (2024). Deep learning-based anomaly detection and log analysis for computer networks. arXiv preprint arXiv:2407.05639.
[9] Gong, C., Zhang, X., Lin, Y., Lu, H., Su, P. C., & Zhang, J. (2025). Federated Learning for Heterogeneous Data Integration and Privacy Protection.
[10] Shih, K., Han, Y., & Tan, L. (2025). Recommendation System in Advertising and Streaming Media: Unsupervised Data Enhancement Sequence Suggestions.
[11] Bao, Q., Chen, Y., & Ji, X. (2025). Research on evolution and early warning model of network public opinion based on online Latent Dirichlet distribution model and BP neural network. arXiv preprint arXiv:2503.03755.
[12] Zhu, J., Ortiz, J., & Sun, Y. (2024, November). Decoupled Deep Reinforcement Learning with Sensor Fusion and Imitation Learning for Autonomous Driving Optimization. In 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 306-310). IEEE.
[13] Liu, Z., Costa, C., & Wu, Y. (2024). Quantitative Assessment of Sustainable Supply Chain Practices Using Life Cycle and Economic Impact Analysis.
[14] Mugobo, V. V., & Baschiera, M. G. (2015). The impact of personalized engagement with customers and efficient stock management software systems on customer service at a clothing retailer in Cape Town, South Africa.
[15] Vepa, A., Yang, Z., Choi, A., Joo, J., Scalzo, F., & Sun, Y. (2024). Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation. Advances in Neural Information Processing Systems, 37, 71643-71671.
[16] Yang, Z., & Zhu, Z. (2024). Curiousllm: Elevating multi-document qa with reasoning-infused knowledge graph prompting. arXiv preprint arXiv:2404.09077.
[17] Li, Z., Ji, Q., Ling, X., & Liu, Q. (2025). A Comprehensive Review of Multi-Agent Reinforcement Learning in Video Games. Authorea Preprints.
[18] Feng, H. (2024, September). The research on machine-vision-based EMI source localization technology for DCDC converter circuit boards. In Sixth International Conference on Information Science, Electrical, and Automation Engineering (ISEAE 2024) (Vol. 13275, pp. 250-255). SPIE.
[19] Zhang, W., Li, Z., & Tian, Y. (2025). Research on Temperature Prediction Based on RF-LSTM Modeling. Authorea Preprints.
[20] Li, Z. (2024). Advances in Deep Reinforcement Learning for Computer Vision Applications. Journal of Industrial Engineering and Applied Science, 2(6), 16-26.
[21] Liu, Z., Costa, C., & Wu, Y. (2024). Leveraging Data-Driven Insights to Enhance Supplier Performance and Supply Chain Resilience.
[22] Liu, J., Li, K., Zhu, A., Hong, B., Zhao, P., Dai, S., ... & Su, H. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8(2), 243-260.
[23] Tang, X., Wang, Z., Cai, X., Su, H., & Wei, C. (2024, August). Research on heterogeneous computation resource allocation based on data-driven method. In 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS) (pp. 916-919). IEEE.
[24] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering, 35(1), 857-876.
[25] Feng, H. (2024). High-Efficiency Dual-Band 8-Port MIMO Antenna Array for Enhanced 5G Smartphone Communications. Journal of Artificial Intelligence and Information, 1, 71-78.

[26] Zhu, J., Wu, Y., Liu, Z., & Costa, C. (2025). Sustainable Optimization in Supply Chain Management Using Machine Learning. International Journal of Management Science Research, 8(1).

[27] Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., & Arriaga-Weiss, S. (2010). Classification in conservation biology: a comparison of five machine-learning methods. Ecological Informatics, 5(6), 441-450.

[28] Liu, Z., Costa, C., & Wu, Y. (2024). Data-Driven Optimization of Production Efficiency and Resilience in Global Supply Chains. Journal of Theory and Practice of Engineering Science, 4(08), 23-33.

[29] Sun, Y., Pai, N., Ramesh, V. V., Aldeer, M., & Ortiz, J. (2023). GeXSe (Generative Explanatory Sensor System): An Interpretable Deep Generative Model for Human Activity Recognition in Smart Spaces. arXiv preprint arXiv:2306.15857.

[30] Yang, J., Chen, T., Qin, F., Lam, M. S., & Landay, J. A. (2022, April). Hybridtrak: Adding full-body tracking to vr using an off-the-shelf webcam. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

[31] Wang, G., Qin, F., Liu, H., Tao, Y., Zhang, Y., Zhang, Y. J., & Yao, L. (2020). Morphing Circuit: An integrated design, simulation, and fabrication workflow for self-morphing electronics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1-26.

[32] Pissas, T., Ravasio, C. S., Cruz, L. D., & Bergeles, C. (2022, October). Multi-scale and cross-scale contrastive learning for semantic segmentation. In European Conference on Computer Vision (pp. 413-429). Cham: Springer Nature Switzerland.