Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9029

DBRS-Net: A Hybrid Framework for Food Image Classification

Dongmei Ma[1,2], Denghui Wang[1*]
[1]School of Physics and Electronic Engineering Northwest Normal University, Gansu Lanzhou, China
[2]Engineering Research Center of Gansu Province for Intelligent Information Technology and Application, Gansu Lanzhou, China
Email: 2216915210@qq.com

How to cite this paper: Ma, D. M., & Wang, D. H. (2025). DBRS-Net: A Hybrid Framework for Food Image Classification. *Journal of Computer Science and Frontier Technologies*, 2(1), 9–18. Print ISSN: 3104-4204; Online ISSN: 3104-4212.
https://doi.org/10.63313/JCSFT.9029
Published: 2025-12-08

Abstract

Food image classification holds significant importance in applications such as food retrieval, nutritional assessment, and dietary management. However, food images typically exhibit pronounced intra-class variations, high inter-class similarity, and complex shooting environments, limiting the classification performance of traditional deep learning approaches. To address this, this paper proposes a novel hybrid network architecture, the Dual-Branch ResNet-Swin Network (DBRS-Net). By integrating ResNet with the window-based attention mechanism of Swin Transformer, it leverages the complementary strengths of both architectures in local fine-grained feature extraction and global context modelling. Specifically, the ResNet branch captures local features such as texture and shape, while the Swin Transformer branch learns overall structure and long-range dependencies. A simple yet effective feature fusion strategy then synthesises comprehensive image representations. Experimental results demonstrate that even without introducing additional complex modules, DBRS-Net achieves stable classification performance on the Food11 dataset, attaining a top-1 accuracy of 94.83%. This provides a reliable foundation for research into food image recognition on larger-scale or more diverse datasets.

Keywords

Food image classification; Deep learning; ResNet; Swin Transformer; Hybrid net-work

1. Introduction

Food constitutes the fundamental basis of human survival, with its safety and nutritional quality directly impacting public health and quality of life. The rapid expansion of social networking platforms has fostered a growing tendency for individuals to document and share their daily dietary habits across various channels, thereby generating vast volumes of food image data. Against this backdrop, food image classification holds significant practical value for assessing dietary patterns, monitoring nutri-

tional intake, and supporting health management. Moreover, poor dietary habits have been established as significant risk factors for numerous chronic conditions, including cardiovascular disease and diabetes. Consequently, employing intelligent food recognition technology for dietary analysis holds considerable importance in promoting public health.

Despite the broad application prospects of food image classification, identifying food items from images remains a challenging task. The primary reasons lie in the inherently unstructured and non-rigid visual characteristics of food images. Their appearance patterns are complex and variable, lacking uniform shapes or geometric structures. Furthermore, at the category level, food images typically exhibit substantial intra-class variation and high inter-class similarity. For instance, the same food item may appear markedly different under varying photographic conditions, while distinct food categories may share striking similarities in colour or texture. These characteristics make it difficult for models to extract discriminative features, rendering food classification more akin to a fine-grained recognition task and further increasing modelling complexity.

Traditional food image recognition methods predominantly rely on artificially designed features,such as colour histograms, texture features, or local keypoints, combined with conventional machine learning classifiers like Support Vector Machines (SVM) or Random Forests for modelling. However, such approaches exhibit strong dependence on feature expressiveness and struggle to adapt to the complex, diverse, and irregular visual manifestations within food images, resulting in generally limited performance in practical scenarios. In contrast, deep learning models can automatically learn multi-level abstract representations through large-scale data, significantly enhancing the performance of food image classification. Currently, most relevant research centres on architectures based on convolutional neural networks (CNNs), such as typical models like ResNet and Inception. Although CNNs exhibit distinct advantages in extracting local textures and regional structures, their limited receptive fields within convolutional kernels constrain feature extraction to local neighbourhoods, thereby lacking efficient modelling capabilities for global contextual information[1]. This structural limitation means pure CNN models may fail to capture sufficient global semantic information when processing food images with high inter-class similarity and complex spatial relationships, ultimately compromising classification performance. In recent years, the rapid advancement of Transformers within computer vision has demonstrated immense potential for image classification. The Transformer architecture employing a Multi-Head Self-Attention (MHSA) mechanism, effectively captures global dependencies and establishes long-range information interactions, exhibiting distinct advantages when processing images with complex scene structures. However, compared to CNNs, Transformers exhibit relatively weaker modelling capabilities for local details and insufficient sensitivity to fine-grained textures. This conflicts with the inherent requirement of food image clas-

sification tasks, which heavily depend on local textures and details. Based on this analysis, we propose a dual-branch hybrid architecture that integrates CNNs and Transformers, fully leveraging their complementary strengths. The CNN branch extracts stable, intricate local features and texture information, while the Transformer branch captures global context and long-range dependencies. Finally, we effectively fuse these two types of features, enabling the model to possess both local recognition capabilities and holistic semantic understanding. This approach enhances the accuracy and generalisation performance of food image classification.

The paper is divided into five sections. Section 2 provides a literature review relevant to the entire study. Section 3 introduces the proposed new model and the methodology employed. Section 4 details the experimental setup and results. Section 5 contains the findings and discussion.

2. Related Work

In recent years, deep learning techniques have achieved significant progress in food image classification, particularly in the domains of CNNs and Transformers. Traditional image classification methods typically rely on manually extracting key features of food subjects and selecting machine learning classifiers. However, these approaches are often constrained by numerous factors, frequently struggling to accurately convey the meaning of food images reflected by these features, consequently leading to relatively low classification accuracy. With the advancement of deep learning, convolutional neural networks have gradually been applied to computer vision, yielding remarkable achievements in food image classification.

In the early stages, researchers applied existing CNN models such as AlexNet[2], VGG[3], GoogleNet[4], and DenseNet[5] to food image classification tasks. Kawano et al.[6] achieved a Top-1 accuracy of 72.26% on the UECFood100 dataset by combining image features extracted from CNNs with those derived from traditional machine learning methods. Chen et al. [7]constructed an InceptionV3 model using CNN architecture, achieving higher accuracy than traditional methods. Liu C et al. [8]refined the GoogLeNet model, attaining a Top-1 classification accuracy of 76.3% on UECFood100. Chen J et al. [9]refined the VGG16 model to generate the new Arch architecture, achieving an 82.12% Top-1 accuracy on UECFood100. David et al. [10]employed CNNs for food image classification on FOOD-101, ultimately attaining an accuracy of 86.97%. Martinel et al. [11]proposed the Wide-Inner-Structure-with-

Residual-Slicing (WISeR) network, leveraging the vertical hierarchical structure of food images. This approach elevated Top-1 accuracy to over 90% on public datasets including Food-101 and UECFood256, significantly outperforming traditional handcrafted features and general deep learning models. Mandal et al. [12]proposed a semi-supervised food recognition method based on deep convolutional generative adversarial networks (SSGAN). Even with only partially labelled data, it achieved recognition accuracy superior to existing fully supervised methods on the Food-101

and Indian Food Dataset. C.S. Won et al.[13] introduced a multi-scale CNN approach in 2020, employing controllable scaling and cropping strategies to simulate 'whole-part'perspectives. This achieved state-of-the-art accuracy in fine-grained food image recognition tasks lacking explicit component boundaries, validating the efficacy of 'scale complementarity'in enhancing classification performance. Deng et al. [14]introduced a context-domain alignment framework for mixed dish recognition in 2022. This approach increased mAP by 4.7% across multi-dish-per-plate datasets in cross-canteen detection, achieving high-precision localisation and classification of ambiguously overlapping dishes for the first time. Chen et al. [15]proposed a network architecture named Res-VMamba, achieving a Top-1 accuracy of 79.54% when evaluated on the novel fine-grained food dataset CNFOOD-241, thereby establishing a new benchmark for this dataset.

3. Methodology

3.1. Overall Framework

In this study, we propose a novel deep learning network termed DBRS-Net(Dual-Branch ResNet-Swin Network). It integrates ResNet50 and Transformer architectures for food image classification. The structure of DBRS-Net is illustrated in Figure 1. The network employs the VIT branch to model complex semantic relationships within images at a global level, whilst the CNN branch extracts rich local information. Ultimately, features extracted from both branches are fused for category prediction, with each branch constrained by its respective loss function. By combining the strengths of VIT and CNN, this approach enhances both global and local feature capture, thereby improving image understanding and the model's ability to perceive information across different scales.

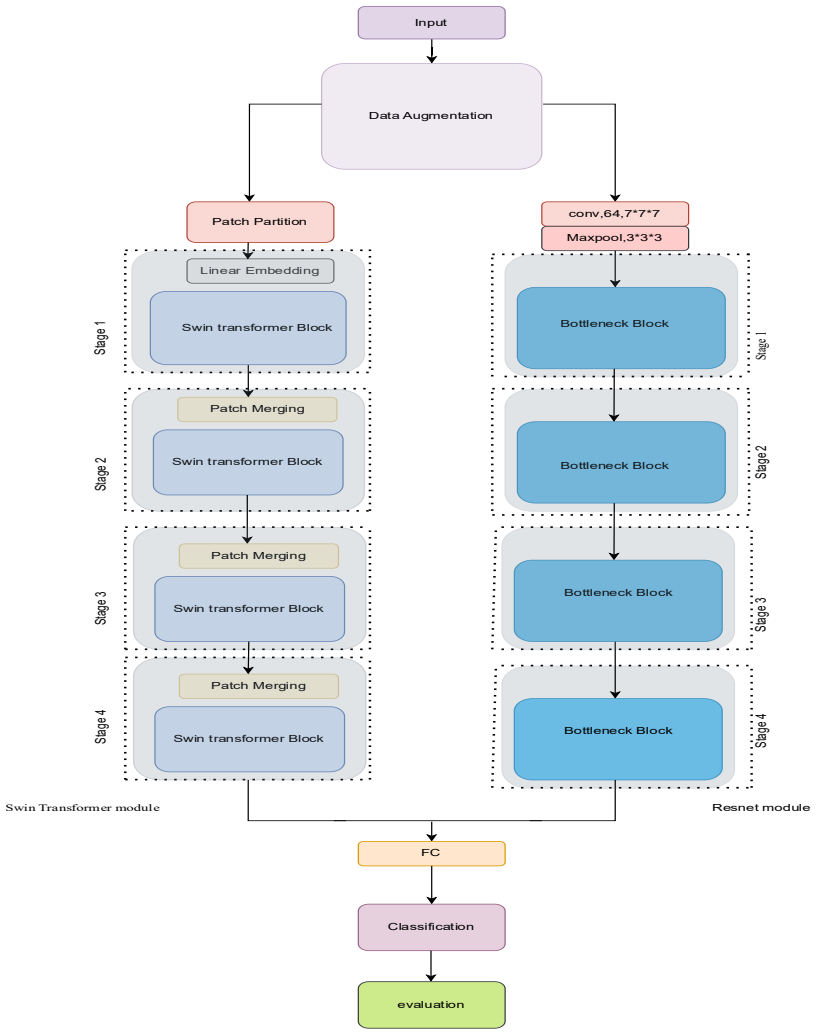

Fig 1. The structure of DBRS-Net

3.2. Transfer learning

Transfer learning involves applying the parameters of a general-purpose model, or one trained on another domain, to the target dataset as a whole. These parameters are then fine-tuned to suit the target dataset, thereby enhancing model performance and speed while conserving computational resources. Typically, we conduct pre-training on the ImageNet dataset. Given that both the target Food11 dataset and ImageNet comprise RGB images, transfer learning is feasible.

4. Experimental Result and Analysis

4.1. Dataset

We selected the Food-11 dataset as the subject of our experimental research. Sourced

from Kaggle, it comprises 16,643 images categorised into 11 primary groups: eggs, fried foods, bread, meat, noodles, desserts, rice, seafood, vegetables and fruit, soups, and dairy products. These foods represent commonplace varieties encountered in daily life, providing the diverse nutrients required for our physical well-being. The dataset is typically split in an 8:1:1 ratio. Figure 2 displays a selection of images from the dataset.

Fig 2. Sample images from the Food11 dataset

4.2. Experimental Configuration

The experiment utilised an NVIDIA GeForce RTX 4090 GPU, with Ubuntu 18.04.5 as the base operating system and an Intel Xeon Platinum 8470Q CPU. Training was conducted within the PyTorch 1.10.2 framework, employing Python as the programming language, with computational tasks supported by CUDA version 11.1. In this experiment, the dataset was proportionally divided into training, validation, and test sets. The training set comprised 9,866 samples, the validation set contained 3,430 samples, and the test set included 3,347 samples. In accordance with the requirements for training and testing the network, images underwent preprocessing to resize them to 224×224 pixels. To enhance training speed and stability, images within the dataset were standardised. During training, the cross-entropy loss function was employed. The Adamw optimiser was utilised with a batch size of 32 and an initial learning rate of 1×10^{-4}. Weight decay was set to 1×10^{-4}.

4.3. Evaluation Metrics

Evaluation metrics play a pivotal role in validating model performance. In the experiments conducted herein, we employed a suite of commonly used metrics to assess model behaviour, including Top-1 accuracy (Top-1Acc), Top-5 accuracy (Top-5Acc), F1 score(F1Score).The mathematical definitions for these metrics are presented in Equations (1) through (3).

$$Top-1Acc=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i=\hat{y}_i) \tag{1}$$

$$Top-5Acc=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i\in\hat{Y}_i^{(5)}) \tag{2}$$

$$F1Score=\frac{2TP}{2TP+FP+FN} \tag{3}$$

Among these, y_i is a genuine label, $\hat{y}_i$ is the predicted top-1 label, $\mathbb{I}(\square)$ is an indicator function, $\hat{Y}_i^{(5)}$ is the set of the first five predicted labels for sample I, N represents the number of categories in the dataset.TP denotes the number of samples where the true category is positive and correctly predicted as positive by the model; FP denotes the number of samples mewhere the true category is negative but erroneously predicted as positive by the model; FN denotes the number of samples with true positive labels incorrectly classified as negative by the model.

4.4. Ablation Experiment

To highlight the contribution of each branch to the final classification results, we conducted ablation experiments, with the outcomes presented in Table 1. To ensure fair comparison, all experiments employed identical data augmentation and training iterations. The top-1 accuracy for the CNN branch and Transformer branch stood at 93.31% and 92.38% respectively, while DBRS-Net achieved 94.83% – 1.52 percentage points higher than the standalone CNN branch and 2.45 percentage points higher than the standalone Transformer branch. Feature space visualisations for each branch are presented in Figure 3, revealing further differences in feature representation across architectures. The standalone CNN branch exhibits some intra-class clustering around local textures and edge information, yet its overall feature distribution remains relatively sparse, as shown in Figure 3(a). Overlap is evident between visually similar food categories, indicating that relying solely on local convolutional features struggles to capture cross-regional semantic relationships within food images. The standalone Transformer branch demonstrates advantages in global modelling, enhancing inter-class separability, as shown in Figure 3(b). However, its insufficient attention to local fine-grained textures results in unstable intra-class consistency, leading to sample distribution fragmentation across several categories. In contrast, DBRS-Net exhibits significantly more compact and discernible feature distributions in t-SNE space, as shown in Figure 3(c). On one hand, samples within the same category demonstrate higher aggregation, markedly reducing intra-class variation. This indicates the fusion architecture effectively integrates local convolutional features with global self-attention features, enabling the model to simultaneously attend to fine-grained textures and holistic semantic information. Conversely, boundaries between categories become more pronounced, with increased inter-class distances and virtually no discernible feature overlap regions. This distribution structure demonstrates DBRS-Net's superior robustness and expressive capability in discriminative feature learning, rendering it better suited to handling the complex and variable ap-

pearance features inherent in food imagery.

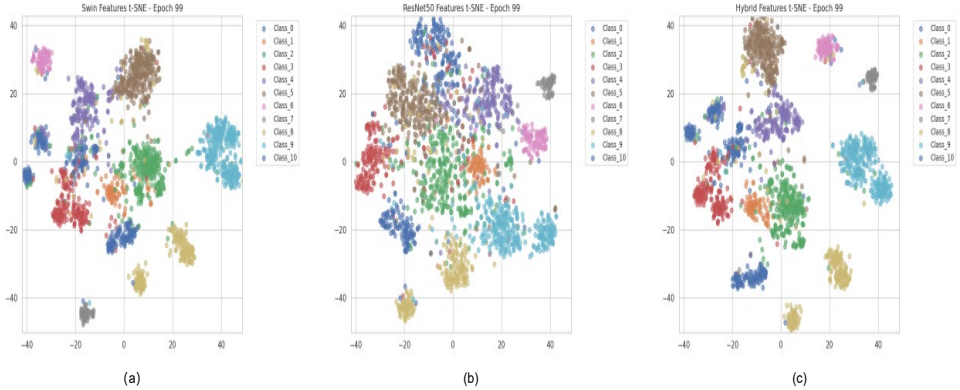

Fig 3. T-SNE plots for each branch

Table 1. Performance Comparison of Single Branch

Model	Top1-acc	Top5-acc	F1-score
Swin Transformer	92.38%	99.91%	92.44%
Resnet50	93.31%	99.70%	93.48%
DBRS-Net	94.83%	99.79%	94.98%

4.5. Comparative Experiment

To evaluate the efficacy of DBRS-Net in food image classification tasks, we conducted a series of comparative experiments aimed at benchmarking its performance against mainstream backbone neworks and various structural configurations,as shown in Table 2.

Table 2. Comparison with Other State-of-the-Art Methods

Methods	Backbone	Top1-acc
CMAL	Res2Next50	96.5%
Food-DCNN	Alexnet	86.9%
Inception-TL	Inception V3	92.9%
DBRS-Net	Resnet50+Swin transformer	94.98%

The table demonstrates that the proposed model DBRS-Net exhibits commendable performance on this task, achieving a Top-1 accuracy of 94.98% and ranking second among all comparison methods. Although slightly below the current state-of-the-art CMAL, the gap between the two is negligible, indicating that the proposed approach delivers robust overall results.

5. Conclusion

This study conducted a systematic analysis of food images, revealing that visual distinctions between food categories primarily reside in multi-scale features. These encompass fine-grained texture details, colour distribution, ingredient combinations, and coarse-grained overall structural characteristics. Such properties result in food image classification tasks exhibiting high inter-class similarity and pronounced intra-class variation, posing challenges to conventional feature extraction methods. To

address this, we propose a novel backbone architecture, DBRS-Net. This model combines the local feature extraction capabilities of ResNet50 with the global representation advantages of Swin Transformers, achieving deep integration of local and global features. Experimental results on the public Food11 dataset demonstrate that DBRS-Net outperforms traditional ResNet-based backbones and pure Transformer models in Top-1 classification accuracy, validating the effectiveness of this fusion architecture in capturing multi-scale discriminative features of food images. Further feature visualisation analysis reveals that DBRS-Net more accurately focuses on key regions and discriminative information within food images, reflecting its robust feature representation capabilities. Overall, DBRS-Net provides a robust and efficient backbone network for food-related computer vision tasks. Its outstanding ability to extract multi-scale discriminative features endows it with broad potential for downstream applications, including food image retrieval, ingredient recognition, and food attribute prediction. Future work may further explore multimodal information fusion, nutritional information modelling, and application to larger-scale datasets to enhance the model's generalisation capability and practical value.

References

[1] Y. Wu and M. Zhang, "Swin-CFNet: An Attempt at Fine-Grained Urban Green Space Classification Using Swin Transformer and Convolutional Neural Network," in IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024, Art no. 2503405, doi: 10.1109/LGRS.2024.3404393.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[3] A. R. Bushara, R. S. V. Kumar, and S. Kumar, "An ensemble method for the detection and classification of lung cancer using computed tomography images utilizing a capsule network with Visual Geometry Group," Biomed. Signal Process. Control, vol. 85, art. no. 104930, 2023.

[4] C. Szegedy, W. Liu, Y. Jia, et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 1–9.

[5] G. Huang, Z. Liu, K. Q. Weinberger and L. van der Maaten, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700 – 4708.

[6] Y. Kawano and K. Yanai, “Food image recognition with deep convolutional features,” in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp), New York, NY, USA, 2014, pp. 589–593.

[7] J. Chen and C. W. Ngo, “Deep-based ingredient recognition for cooking recipe retrieval,” in Proc. 24th ACM Int. Conf. Multimedia, 2016, pp. 32–41.

[8] C. Liu, Y. Cao, Y. Luo, et al., “DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment,” in Proc. IEEE Int. Conf. Smart Homes Health Telematics (ICOST), Cham, Switzerland, 2016, pp. 37–48.

[9] J. Chen and C. W. Ngo, “Deep-based ingredient recognition for cooking recipe retrieval,” in Proc. 24th ACM Int. Conf. Multimedia, New York, NY, USA, 2016, pp. 32–41.

[10] D. J. Attokaren, I. G. Fernandes, A. Sriram, Y. V. S. Murthy, and S. G. Koolagudi,“Food classification from images using convolutional neural networks,”in Proc. 2017 IEEE Region 10 Conference (TENCON), Penang, Malaysia, 2017, pp. 2801–2806.

[11] N. Martinel, G. L. Foresti, and C. Micheloni, “Wide-Slice residual networks for food recognition,” in Proc. 2018 IEEE Winter Conf. Appl. Comput. Vis. (WACV), Lake Tahoe,

NV, USA, Mar. 12–15, 2018, pp. 567–576.

[12] B. Mandal, N. B. Puhan, and A. Verma, "Deep convolutional generative adversarial network-based food recognition using partially labeled data," IEEE Sensors Letters, vol. 3, pp. 7000104, 2019. doi: 10.1109/LSENS.2019.2925538.

[13] C. S. Won, "Multi-scale CNN for fine-grained image recognition," IEEE Access, vol. 8, pp. 116663 – 116674, 2020. doi: 10.1109/ACCESS.2020.3001234.

[14] L. Deng et al., "Mixed Dish Recognition With Contextual Relation and Domain Alignment," in IEEE Transactions on Multimedia, vol. 24, pp. 2034-2045, 2022, doi: 10.1109/TMM.2021.3075037.

[15] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen, "Res-VMamba: Fine-grained food category visual classification using selective state space models with deep residual learning," arXiv:2402.15761 [cs.CV], 2024. [Online]. Available: https://arxiv.org/abs/2402.15761