Article

Food Classification for Dietary Support Using Fine-Grained Visual Recognition with the HERBS Network

Chi-Sheng Chen , Yu-Hsuan Yang , Guan-Ying Chen , Shao-Hsuan Chang *

Posted Date: 28 October 2024

doi: 10.20944/preprints202410.2080.v1

Keywords: Food dataset; HERBS; fine-grained food recognition; dietary assessment

Article

Food Classification for Dietary Support Using Fine-Grained Visual Recognition with the HERBS Network

Chi-Sheng Chen [1], Yu-Hsuan Yang [2], Guan-Ying Chen [1] and Shao-Hsuan Chang [2,*]

[1] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University
[2] Professional Master's Program of Biotechnology Management, National Taiwan University, Taipei, Taiwan
* Correspondence: changshao1311@gmail.com; Tel.: +886-223620502

Abstract: Background: Food properties can directly influence individual dietary intake. Therefore, computer vision-based food recognition could be used to estimate meal contents for patients with metabolic diseases. Food recognition based on deep learning can create opportunities for breakthroughs in dietary interventions for personal health management and have rapidly emerged as dietary assessment strategies. **Methods**: This study proposed a methodology for automatic food recognition based on fine-grained visual classification, the High-temperaturE Refinement and Background Suppression network (HERBS). A technical investigation was conducted involved in the HERBS model on CNFOOD241 benchmark to verify the model effectiveness. The visualization analysis of HERBS was compared with the VGG16 and RepViT on CNFOOD241. **Results**: The system achieved classification accuracy of 82.72 % and 97.19 % in Top-1 accuracy and Top-5 accuracy respectively. Data showed that HERBS structure widened the attention area and outperformed VGG16 and RepViT on feature maps. **Conclusions**: Our findings elucidated that the proposed methodology using HERBS model establishes a new benchmark for SOTA performance in food recognition on CNFOOD 241 dataset. This study proved the feasibility of the proposed approach in a challenging food dataset. We further described a vision-language model structure consisting of a multi-modal vision model, language encoder and a multilayer perceptron classifier for food composition recognition.

Keywords: Food dataset; HERBS; fine-grained food recognition; dietary assessment

Introduction

Currently, food composition compilation serves as a standard paradigm to estimate food nutrition information. Poor dietary and imbalanced food consumption have a multifaceted impact on patients with diabetes, prediabetes, and other metabolic disorders as important components of blood glucose management. Effective food recognition and tracking have always been critical but challenging aspects (Dalakleidi et al., 2022). With the rise of modern technology and significant changes in dietary habits, there is a growing focus on food computing (Min et al., 2019). Food image classification using convolutional neural networks (CNNs) has been developed for different purposes. CNNs show state-of-the-art (SOTA) performance in fixed-class image recognition tasks using closed datasets like ImageNet (Deng et al., 2009). A well-defined dataset significantly influences the development of possible research topics and the feature-learning capabilities of models. However, the biggest challenge for food classification is the low inter-category variation and high intra-category variation of a dataset. Even subtle differences in categories can result in visually similar but semantically different food types. In addition, the background information of an image also provides important information that can tell the model which features are unnecessary or even harmful for classification. As a result, models that focus primarily on subtle features may overlook global features and contextual information.

The algorithm for extracting features from discriminative regions is of importance for the fine-grained visual classification task. Recently, a high-temperature refinement module (High-temperaturE Refinement and Background Suppression, HERBS) has been proposed in order to enhance the learning of diverse features, including texture, shape, and appearance from various categories (Chou et al., 2023). The HERBS model is a novel network designed for fine-grained visual classification, consisting of the high-temperature refinement module and the background suppression module for extracting discriminative features and suppressing background noise. It addresses the challenges of classifying images that are very similar to each other by extracting subtle and discriminative features while suppressing irrelevant background noise.

The HERBS model integrates seamlessly with various neural network architectures, supporting end-to-end training. It was reported to achieve SOTA performance on bird classification datasets, the CUB-200-2011 and NABirds, surpassing 93% accuracy (Chou et al., 2023). However, this deep progressive region enhancement network has not been validated for food image classification, which is one practical use case of image recognition technology. Additionally, the background of food images could also present diverse features and provide unnecessary or harmful information for classification. The HERBS is expected to manage the feature scale effectively on food classification dataset, ensuring that the features are neither too broad to capture irrelevant details nor too narrow to miss contextual information. Therefore, this study evaluated current training strategies comparing with HERBS to obtain abundant and various food features from CNFOOD241 and other food datasets. This strategy can help to learn comprehensive and multiple fine-grained information as training progresses.

The current study evaluates these training approaches for food recognition and introduces HERBS strategy with significant impact on efficiency and accuracy. We clarified the research value of CNFOOD241 and partitioned the dataset into separate test and validation segments as a new fine-grained image classification benchmark. Additionally, we conducted extensive evaluation on CNFOOD241 to verify the effectiveness of HERBS, including popular deep networks, fine-grained recognition methods and existing food classification models.

Related Work

Food Datasets

The widely used food image databases include Food-101, UECFood-100, and UECFood-256 (Matsuda & Yanai, 2015). Food-101 consisting of 101 food categories with 1000 images per category is the commonly studied dataset in the field of food recognition (Bossard et al., 2014). Furthermore, ISIA Food-500 (Min et al., 2020) and Food2K (Min et al., 2023) as the expansive collections encompass nearly four hundred thousand and over a million images respectively. Although these datasets show comprehensive coverage of categories and larger quantity of images, the lack of uniformity in the size distribution of images across different categories can lead to substantial discrepancies in categories. The image size inconsistency poses challenges in maintaining the accuracy and reliability of computational models, especially those reliant on CNNs, ViTs and other image-processing architectures designed to extract detailed features from visual inputs. CNFOOD241 is a Chinese food dataset composing of 190,000 images with 241 diverse food categories (Fan et al., 2023). Unlike other food databases, CNFOOD241 preserves the aspect ratio of images and standardizes the size to 600×600 pixels. The disparity in image sizes may affect the performance of these models, leading to deviations in the extracted category-specific information and potentially impacting the overall effectiveness of the computational analysis. This preprocessing step prevents image deformation during data augmentation, which could potentially lead to models learning incorrect semantic features. CNFOOD-241 (Fan et al., 2023) with uniform image sizes renders it an exceptional resource for conducting detailed image analyses within the food computation domain, facilitating more accurate and reliable studies in food recognition, nutritional analysis, and other related areas.

Food Recognition

Although food recognition belongs to fine-grained analysis, it has unique characteristics. Many of these techniques rely on the recent success of deep learning for visual recognition, and use SOTA models to train a deep convolution network that can recognize a variety of food items. PARNet improves the performance of classification by mining discriminative food regions (Bossard et al., 2014). Other adapted models such as Adjusted AlexNet, DeepFood, Inception V3, ResNet, PRENet and WISeR have been evaluated with improved accuracy on large-scale food databases (Min et al., 2023). These approaches eliminate the need of manual feature extraction and allow models to learn hierarchical visual representations from large-scale labeled datasets. Recently, ViT has have gained popularity in food image analysis, which divides images into patches and applied self-attention to capture relationships between patches. Recently, models based on the State Space Model, such as VMamba (Liu et al., 2024), are regarded as outperforming ViTs on large image datasets like ImageNet (Deng et al., 2009). It retains the advantage of capturing both local and global information from input images as ViTs while also enhancing the model speed. Our previous work employed VMamba on food images and further designed ResVMamba model specifically for fine-grained datasets (Chen et al., 2024). This model incorporates a mechanism to share global and local feature states, aiming to enhance performance on detailed image classification tasks. ResVMamba model is well-suited for handling high-resolution and data-imbalanced scenarios, making it ideal for real-world applications in food recognition.

Fine-Grained Visual Classification

In the field of fine-grained visual classification, strategies have concentrated on extracting discriminative features from subtle areas between categories and enhancing the discriminative features. This area has seen the development and application of several SOTA models, each contributing to advancements in dataset-specific performance. Multi-attention CNN trains positioning and classification accuracy at the same time through clustering of feature maps into object parts (Zheng et al., 2017). The approach allows for simultaneous learning of discriminative features and positions. Each expert module processes feature maps from specific layers, delivering both a categorical prediction and an attention region. API-Net and PCA-Net calculate attention between feature maps to enhance discriminative representations (Zhuang et al., 2020). HERBS can effectively fuse features of varying scales, suppresses background noise, discriminative features at appropriate scales (Chou et al., 2023). The model includes two primary components:

1. High-temperature Refinement Module: This module enables the network to refine feature maps at various scales. It uses higher temperatures in initial layers to encourage the learning of global and contextual features, followed by lower temperatures to capture finer details. This approach helps in distinguishing between very similar categories by enhancing the learning of diverse features.

2. Background Suppression Module: This component works by first identifying the foreground (discriminative areas) and background regions in an image based on classification confidence scores. It then suppresses features in areas with low confidence while enhancing those in high-confidence, discriminative regions. This suppression helps in reducing noise and focusing on the important features of the target object. These advancements provide a rich tapestry of methodologies for addressing the nuanced challenges of food recognition.

Methods

Dataset Reconstruction

CNFOOD241 consists of 190,000 images with 241 diverse food categories, preserving the aspect ratio of images and standardizes the size to 600×600 pixels. The unveiling of the CNFOOD-241 dataset marks a significant advancement in fulfilling the essential demand for high-quality, uniform datasets within the domain of food computation. The dataset was re-established as a more equitable benchmark, and was partitioned into training, validation, and testing sets randomly. We further split the 'train600x600' folder in CNFOOD-241 into 2 groups at ratio of 7:3. Finally, the training set therefore contains 119,514 images and the validation set has 51,354 images in all 241 categories. The

doi:10.20944/preprints202410.2080.v1

testing set was the original 'val600x600' folder which included 20,943 images as shown in **Figure 1**. The train-validation split list can be available on the github: https://github.com/ChiShengChen/ResVMamba.

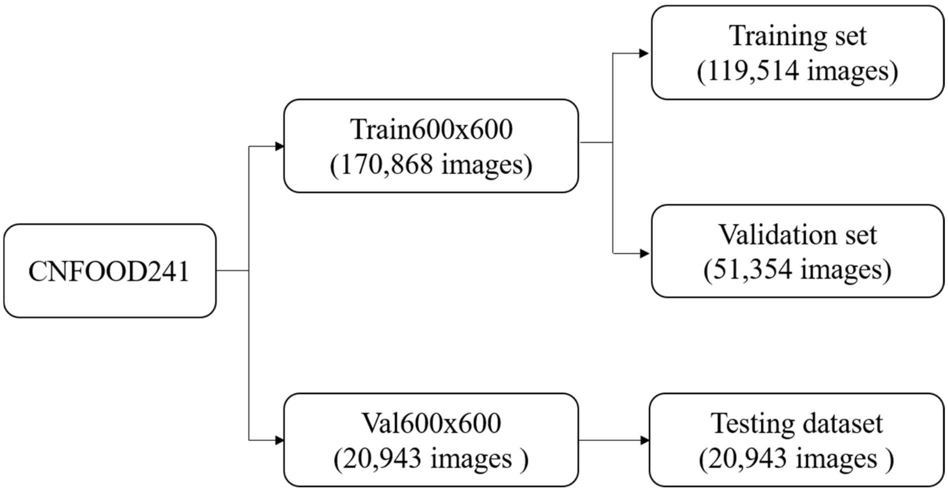

Figure 1. CNFOOD-241 data split flow.

Unbalanced Number of Images Cross Categories on CNFOOD-241

To assess the imbalance of image counts across categories in the dataset, we calculated the normalized entropy. The normalization is achieved by dividing the entropy H by $\log_2(n)$, where n represents the total number of categories. This normalized entropy, referred to as H_{norm}, is computed as follows:

$$H_{norm} = \frac{H}{log_2(n)} \qquad (1)$$

Given the original definition of entropy:

$$H = -\sum_{i=1}^{n} pi \log_2(p_i) \qquad (2)$$

The normalized entropy H_{norm} ranges from 0 to 1. A value of $H_{norm}=1$ signifies a perfectly balanced dataset, where each category has an equal representation. Conversely, $H_{norm}=0$ indicates complete imbalance, with all instances belonging to a single category. This normalization facilitates easier comparisons of entropy values across datasets with varying numbers of categories, offering a standardized measure of category balance. Hence, the existing food recognition benchmarks together with CNFOOD241 including the size of dataset, the quantity of images for each category, and entropy results was demonstrated in **Table 1 and Figure 2**.

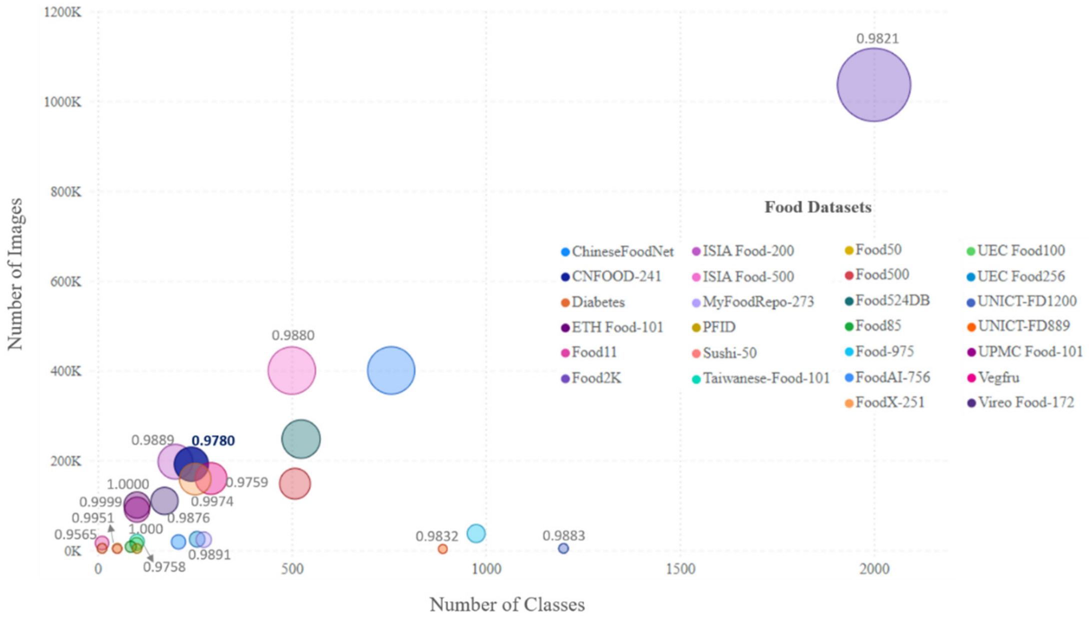

Figure 2. Comparison of CNFOOD241 with predecessor food databases. CNFOOD-241 (in dark blue) exhibited a greater degree of class imbalance compared to other datasets with entropy of 0.9780.

Table 1. Comparison of current food recognition datasets.

Dataset	Year	Classes/Images	Category Entropy	Type	Public
PFID	2009	101/4,545	-	Western	-
Food50	2010	50/5,000	-	Misc.	-
Food85	2010	85/8,500	-	Misc.	-
UEC Food100	2012	100/14,361	0.9758	Japanese	V
UEC Food256	2014	256/25,088	0.9891	Japanese	V
ETH Food-101	2014	101/101,000	1.0000	Western	V
Diabetes	2014	11/4,868	-	Misc.	-
UPMC Food-101	2015	101/90,840	0.9999	Western	V
UNICT-FD889	2015	889/3,583	0.9832	Misc.	V
Vireo Food-172	2016	172/110,241	0.9876	Chinese	V
Food-975	2016	975/37,785	-	Misc.	-
Food500	2016	508/148,408	-	Misc.	-
Food11	2016	11/16,643	0.9565	Misc.	V
UNICT-FD1200	2016	1,200/4,754	0.9883	Misc.	V
Food524DB	2017	524/247,636	-	Misc.	-
ChineseFoodNet	2017	208/192,000	-	Chinese	-
Vegfru	2017	292/160,000	0.9759	Misc.	V
Sushi-50	2019	50/3,963	0.9951	Japanese	V
FoodX-251	2019	251/158,846	0.9974	Misc.	V
ISIA Food-200	2019	200/197,323	0.9889	Misc.	V
FoodAI-756	2019	756/400,000	-	Misc.	-
Taiwanese-Food-101	2020	101/20,200	1.0000	Chinese	V
ISIA Food-500	2020	500/399,726	0.9880	Misc.	V
Food2K	2021	2000/1,036,564	0.9821	Misc.	V
MyFoodRepo-273	2022	273/24,119	-	Misc.	-
CNFOOD-241	2022	241/191,811	0.9780	Chinese	V

The comparison of entropy values across datasets with different numbers of categories provided a standardized measure of category balance. CNFOOD-241 exhibited a greater degree of class imbalance compared to other datasets, which consequently increases the challenging nature of this dataset. Considering the above factors, we selected CNFOOD241 for experiments on fine-grained food classification.

Comparison with SOTA Methods

We compare various deep networks including VGG16 (Simonyan & Zisserman, 2014), ViT-B (Dosovitskiy et al., 2021), ResNet101 (Zahisham et al., 2020), DenseNet121 (Huang et al., 2017), InceptionV4 (Szegedy et al., 2016), PRENet (Min et al., 2023), SENet 154 (Hu et al., 2018), RepViT (Wang et al., 2023), ConvNeXT-B (Liu et al., 2022), EfficientNet-B6 (Le, 2019), CMAL-Net (Liu et al., 2023), VMamba-S (Chen et al., 2024), and HERBS (Chou et al., 2023). PRENet was previous SOTA method on Food2K for food recognition (Min et al., 2023). Previously, we demonstrated the updated SOTA method, VMamba-S, on the CNFOOD-241 benchmark (Chen et al., 2024). The Top-1 (Top-1 Acc.) and Top-5 classification accuracy (Top-5 Acc.) were adopted as evaluation metrics. All the neural network models are implemented using the PyTorch framework. Experiments were completed on a single Nvidia GeForce RTX 3090, and the Pytorch toolbox is used as the main implementation substrate. For the deep networks, we trained HERBS network with Swin-Transformer as the backbone network, which pretrained weights was from ImageNet. HERBS was trained with a learning rate of 10^{-2}, and divided by 10 after 30 epoches. All the networks were optimized using the stochastic gradient descent with a momentum of 0.9, and weight decay of 10^{-4}. Training and testing sets are performed with an image size of 600 x 600. All the hyperparameters were followed its default setting.

Visualization Analysis

We employed Grad-CAM (Gradient-weighted Class Activation Mapping), a deep neural network visualization technique by gradient-based localization, as our model interpretation method (Selvaraju et al., 2017). This technique highlights the regions of an image that are most significant for the decisions made by a convolutional neural network (CNN). It produces a coarse localization map that highlights the parts of the image relevant for predicting a particular class. The Grad-CAM applied to HERBS uses its original code obtained from GitHub, while others utilize the Captum package.

Results

Recognition Performance on Other Datasets

Results showed that HERBS method surpassed the latest SOTA methods with model pre-trained on ImageNet. The HERBS module can be utilized not only with transformer structures but also with convolution-based methods, which outperforms typical fine-grained methods with Swin-Transformer as the backbone. **Table 2** described the performance comparison of different SOTA methods on CNFOOD-241. The recognition performance showed consistent improvement when adopting with more advanced networks. Furthermore, most fine-grained methods performed better than deep networks. In the bottom row of Table 2, HERBS reached 82.72 % and 97.19 % in Top-1 accuracy and Top-5 accuracy respectively, which was 1.02% and 0.34% higher than the previous SOTA approach (ResVMamba). The proposed HERBS network can effectively filter out background noise and extract appropriately sized discriminative features, enabling the identification of fine-grained food categories accurately.

Table 2. Comparison of Top-1 and Top-5 Accuracy (%) on CNFOOD241 with our split dataset.

Model	Top-1 Val. Acc.	Top-5 Val. Acc.	Top-1 Test Acc.	Top-5 Test Acc.
VGG16	66.98	90.10	65.06	89.60
ViT-B	73.14	92.06	71.58	91.62
ResNet101	74.42	93.62	72.59	93.16
DenseNet121	76.46	94.57	74.77	94.29
InceptionV4	77.30	94.28	75.70	93.89

SEnet154	77.47	94.86	76.02	94.61
PRENet	77.28	95.16	76.28	94.85
RepViT	78.08	95.41	76.86	95.02
ConvNeXT-B	78.30	94.36	76.76	93.90
EfficientNet-B6	80.10	94.64	78.48	94.22
CMAL-Net	80.16	95.99	78.56	95.40
VMamba-S	82.15	96.91	80.58	96.71
ResVMamba	79.54[#]	95.72[#]	81.70	96.83
HERBS	83.56	97.31	82.72	97.19

Val., Acc. denote validation and accuracy respectively. [#]ResVMamba without pretrained weight on CNFOOD-241.

However, the performance trend on CNFOOD-241 is not consistent with existing fine-grained datasets, such as birds and cars datasets. As these fine-grained methods may focus more on the same and common semantic parts (ex: bird's mouth), or to capture the interlayer part feature relations and extracts the co-occurring features to integrate a unified representation for classification. Nevertheless, many food categories have non-rigid structures, and they do not have fixed semantic information. Furthermore, it was reported that these existing fine-grained models paid more attention to a large amount of background area, and they are not designed as a food-oriented network (Chou et al., 2023). Therefore, HERBS is expected to achieve optimal performance for food recognition as it filters out background noise and focuses on discriminative features while maintaining a proper attention area scale. Our results showed that HERBS significantly improves accuracy and outperforms SOTA methods on CNFOOD-241, indicating its better model generalization ability and higher tolerance for misclassification.

Visualization Analysis

To gain insight into HERBS method, we further conducted visualization analysis via Grad-CAM. The output feature maps were visualized and compared HERBS method with the VGG16 and RepViT. The corresponding heat map was presented in **Figure 3**. The attentional regions expand as the model evolves, incorporating more discriminative and detailed elements. Take "Steamed bun" for example (the first row in Figure 4), VGG16 and RepViT captured only limited information, while HERBS maintained detail with a notable improvement on accuracy. The overall characteristics of the "Steamed bun" can be specified using HERBS, emphasizing the use of background suppression and high-temperature refinement techniques for recognition.

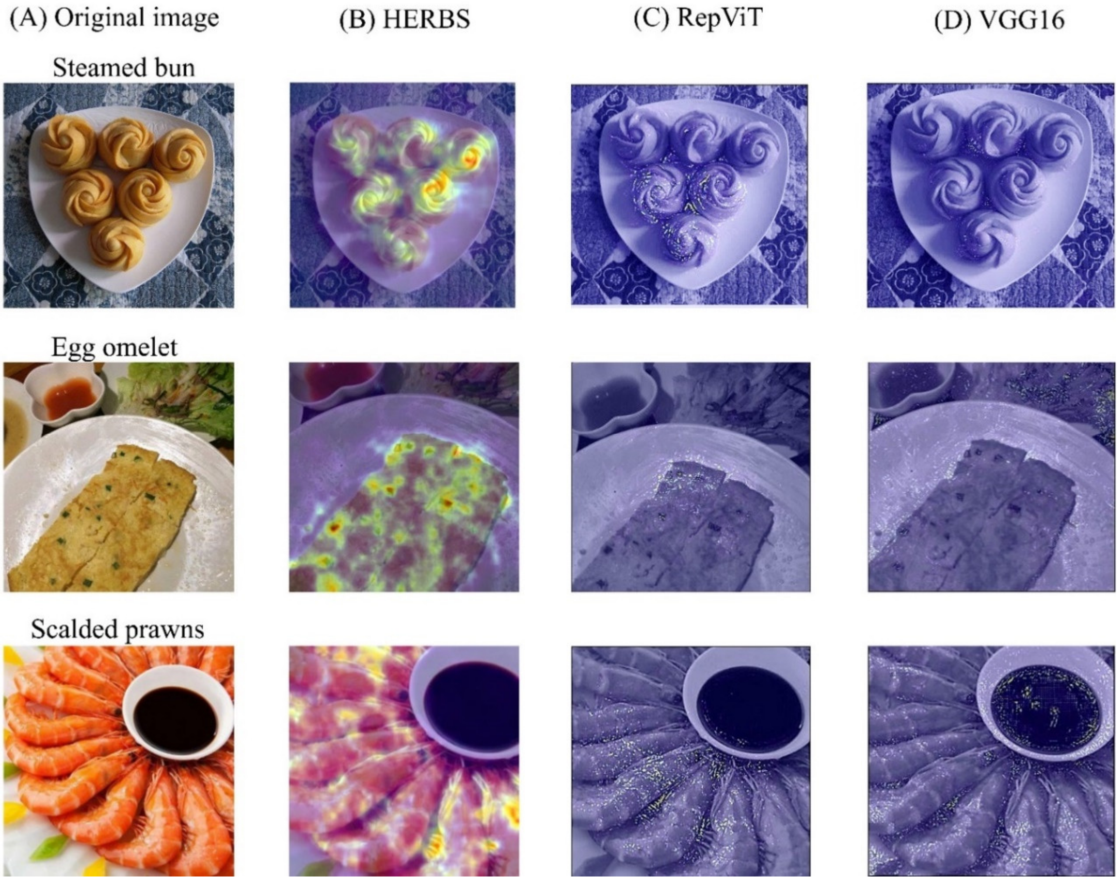

Figure 3. Visualization of heat maps on food samples from CNFOOD241 generated from different models. (A) original color image, (B) HERBS, (C) RepViT, and (D) VGG16.

Food Image Generation

In this work, we employed CNFOOD241 database which exhibited a greater degree of class imbalance compared to other datasets. Considering the challenging nature of food dataset consisting of images from given ingredients and cooking instructions, we proposed a vision-language model structure incorporating a multi-modal vision model, language encoder and a multilayer perceptron classifier for food composition recognition as shown in **Figure 4**. The fast R-CNN and HERBS take food images as inputs and generate image embeddings to feed into the following multilayer perceptron classifier. The two-stage transformers process food recipes to initiate food ingredient statements, and hence generates text embeddings as classifier inputs. We proposed a practical solution for personalizing classifiers for each food image registration. By integrating personalized diets, individuals will be better supported in adhering to healthy dietary recommendations.

doi:10.20944/preprints202410.2080.v1

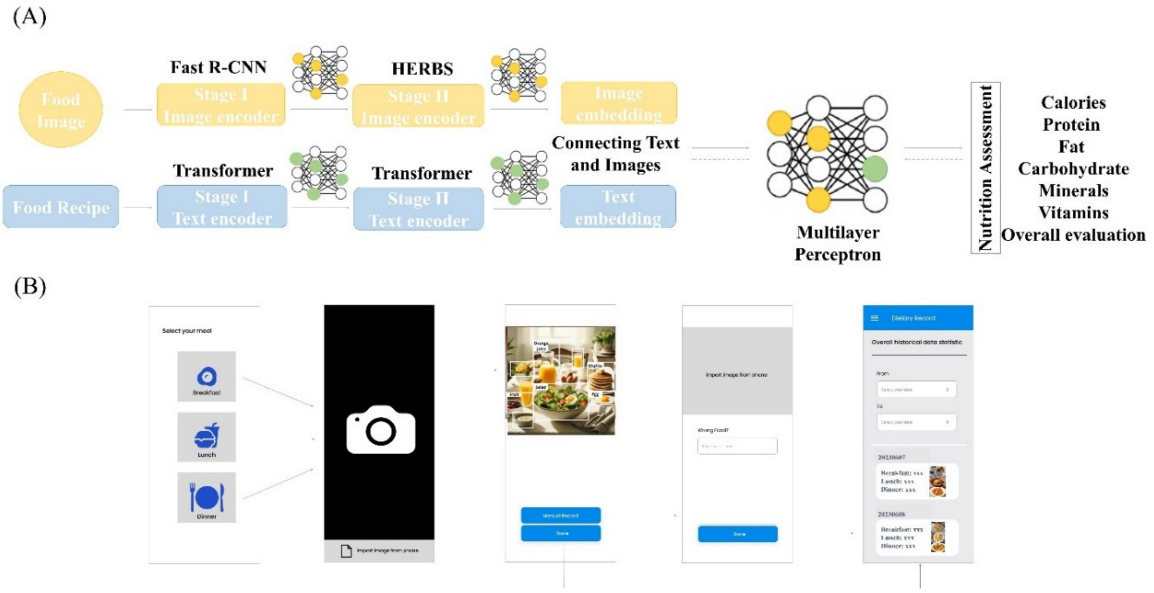

Figure 4. The illustration of the proposed vision-language model structure for food composition recognition. (A) The proposed structure consists of a multi-modal vision model, language encoder and a multilayer perceptron classifier. (B) Flow of food logging process and food composition profile estimation through image recognition.

Discussion

Fine-grained visual classification is a challenging task in the field of computer vision that focuses on distinguishing between similar object or image categories, often requiring a deep analysis of subtle features. This area has seen the development and application of several state-of-the-art (SOTA) models, each contributing to advancements in dataset-specific performance. By comparing the CNFOOD-241 dataset with other food databases, we highlighted the characteristics of CNFOOD241 as a high-resolution, data-imbalanced, and therefore a challenging dataset. For the datasets of Food2K and ETH Food-101, PRENet achieved top-1 accuracies of 83.75% and 91.13%, respectively (Min et al., 2023). However, its top-1 accuracy on CNFOOD-241 reached only 76.2%, demonstrating the considerable difficulty of CNFOOD-241. The entropy values from our results also highlighted CNFOOD241 as a complex dataset tests the limits of current models. To address the challenge, we previously introduce ResVMamba model, incorporating a residual deep learning structure to improve the performance in processing complex food dataset with the accuracy of 81.70%. In this work, we further employed HERBS with ImageNet pretrained weight reached the SOTA on CNFOOD-241 with top-1 accuracy to 82.72%, resulting in a notable improvement of 1.02% in top-1 accuracy. Our findings elucidated that the proposed methodology using HERBS model establishes a new benchmark for SOTA performance in food recognition on CNFOOD 241 dataset. This system effectively extracted discriminative features and suppresses background noise for food images classification.

To further compare the capabilities of different model structures, we conducted visualization analysis with the VGG16 and RepViT on CNFOOD241. Results showed that HERBS structure widened the attention area and outperformed VGG16 and RepViT on feature maps. VGG16 is often used as a benchmark for many subsequent models (Simonyan & Zisserman, 2014). However, VGG16 as a CNN model does not incorporate specific background suppression techniques. It relies on deep layers of convolutions to extract features, which can sometimes lead to sensitivity to background noise, while the ViTs incorporate a self-attention module to capture long-range dependencies, emphasizing the learning of global features. Wang et al. introduced the Squeeze-and-Excitation (SE) block into vision transformers (ViTs), which contributed to the development of RepViT (Wang et al., 2023). Thus, RepViT aims to enhance the learning of local features while maintaining the ability to extract global features, highlighting the combination of local and global features. It handles background information but does not specifically focus on background suppression. HERBES proposed by Chou et al. focuses on suppressing background noise and enhances target features through high-temperature refinement techniques, thereby improving classification accuracy, particularly in fine-grained classification tasks (Chou et al., 2023). HERBS achieves higher accuracy in fine-grained classification tasks on CNFOOD241, while VGG16 and RepViT predicted the image incorrectly and had a relatively narrow focus. This demonstrates that fine-grained visual classification requires detailed features rather than features that are too narrow. In addition, it is crucial to separate backgrounds by feature values rather than class probability. These characteristics enable HERBS to outperform VGG16 and RepViT in processing complex food dataset.

Recently, advancements in computer vision techniques have greatly enhanced research and visual features on food recognition (Zhou et al., 2019). We proposed a practical solution for personalizing classifiers for each food image registration. Moving forward, precision health management will focus on a more detailed characterization of food components, which is crucial for effective nutritional management. However, personalized dietary recall poses significant challenges due to its inherent complexity, including time-variant characteristics and ongoing advancements in nutrition science that emphasize more tailored dietary recommendations (de Toro-Martin et al., 2017). The HERBS fine-grained food classification can play a vital role in nutrition by enabling accurate nutritional analysis through precise identification of various food types, which is essential for understanding their nutritional value. Different foods have distinct nutrient profiles, and this accurate classification helps users comprehend their dietary composition. Additionally, the HERBS model facilitates precise food recognition, allowing for the tracking of dietary habits. Identifying these components can improve disease management and reduce cardiovascular risks for adults with metabolic diseases. As a result, precision nutrition has emerged as a vital strategy for disease prevention and clinical intervention, contributing to the alleviation of illness burdens.

Conclusion

This study assessed the HERBS network against the CNFOOD241 benchmark to evaluate its effectiveness on food classification. The HERBS model demonstrated its superior performance compared to VGG16 and RepViT by widening the attention area in feature maps. The findings found that HERBS achieved the highest accuracy in food recognition on CNFOOD241 dataset, validating its applicability in challenging food contexts. Additionally, we introduced a vision-language model that incorporates a multi-modal vision model, a language encoder and a multilayer perceptron classifier for the potential of computer vision-based food recognition on individual dietary intake.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no competing interests.

Reference

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101 – mining discriminative components with random forests. *Computer Vision– ECCV 2014*, 446-461.

Chen, C. S., Chen, G. Y., Zhou, D., Jiang, D., & Chen, D. (2024). Res-VMamba: Fine-Grained Food Category Visual Classification Using Selective State Space Models with Deep Residual Learning. *arXiv:2402.15761*. https://doi.org/10.48550/arXiv.2402.15761

Chou, P. Y., Kao, Y. Y., & Lin, C. H. (2023). Fine-grained Visual Classification with High-temperature Refinement and Background Suppression. *arXiv:2303.06442*. https://doi.org/10.48550/arXiv.2303.06442

Dalakleidi, K. V., Papadelli, M., Kapolos, I., & Papadimitriou, K. (2022). Applying Image-Based Food-Recognition Systems on Dietary Assessment: A Systematic Review. *Adv Nutr*, *13*(6), 2590-2619. https://doi.org/10.1093/advances/nmac078

de Toro-Martin, J., Arsenault, B. J., Despres, J. P., & Vohl, M. C. (2017). Precision Nutrition: A Review of Personalized Nutritional Approaches for the Prevention and Management of Metabolic Syndrome. *Nutrients*, *9*(8). https://doi.org/10.3390/nu9080913

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F. (2009). Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248-255.

Dosovitskiy, A., Beyer, L., Kolesnikov, K., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy.

Fan, B., Li, W., Dong, L., Li, J., & Nie, Z. (2023). Automatic Chinese Food recognition based on a stacking fusion model. *Annu Int Conf IEEE Eng Med Biol Soc*, *2023*, 1-4. https://doi.org/10.1109/EMBC40787.2023.10340620

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Le, Q., Tan, M. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *Technical report*.

Liu, D., Zhao, L., Wang, Y., & Kato, J. (2023). Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. . *Pattern Recognition*, *140*. https://doi.org/10.1016/j.patcog.2023.109550

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). Vmamba: Visual state space model. *Technical report*.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & S., X. (2022). A convnet for the 2020s. *Technical report*.

Matsuda, Y., & Yanai, K. (2015). Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. *Computer Vision -ECCV 2014 Workshops*, 3-17. https://doi.org/10.1007/978-3-319-16199-0_1

Min, W., Jiang, S., Liu, L., Rui, Y., & Jain, R. (2019). A survey on food computing. *ACM Comput. Surv.*, *52*(5).

Min, W., Liu, L., Wang, Z., Luo, Z., Wei, X., Wei, X., & Jiang, S. (2020). Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. *In Proceedings of the 28th ACM International Conference on Multimedia*.

Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., & Jiang, S. (2023). Large Scale Visual Food Recognition. *IEEE Trans Pattern Anal Mach Intell*, *45*(8), 9932-9949. https://doi.org/10.1109/TPAMI.2023.3237871

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *arXiv:1610.02391* https://doi.org/10.48550/arXiv.1610.02391

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Loffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. *Technical report*.

Wang, A., Chen, H., Lin, Z., Han, J., & G., D. (2023). Repvit: Revisiting mobile cnn from vit perspective. *Technical report*.

Zahisham, Z., Lee, C. P., & Lim, K. M. (2020). Food recognition with resnet-50. *In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 1-5. https://doi.org/10.1109/IICAIET49801.2020.9257825

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. *IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2017.557

Zhou, L., Zhang, C., Liu, F., Qiu, Z., & He, Y. (2019). Application of Deep Learning in Food: A Review. *Compr Rev Food Sci Food Saf*, *18*(6), 1793-1811. https://doi.org/10.1111/1541-4337.12492

Zhuang, P., Wang, Y., & Qiao, Y. (2020). Learning attentive pairwise interaction for fine-grained classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(7), 13130–13137.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.