A Fine-Grained Long-Tail Distribution Food Image Classification Model with Attention Mechanism

Yanhao Bao

{c2062174@ncl.ac.uk}

School of Computing, Newcastle University, Newcastle upon Tyne, United Kingdom

Abstract. The fine-grained food image classification is a challenging task due to the presence of long-tail distributions in food categories, which results in imbalanced datasets. This imbalance will lead to biased models that underperform on rare classes, thereby affecting the model's robustness and accuracy. This paper proposes FL-FoodNet, a classification model designed to enhance fine-grained recognition in food image classification. FL-FoodNet integrates both channel attention and spatial attention mechanisms, which help the model focus on intrinsic features and precise locations within food images. Copping with the challenges posed by the long-tail distribution of categories, Focal Loss is added, allowing the model to focus on underrepresented classes and improving generalization across diverse food types. Experimental results demonstrate that FL-FoodNet achieves superior performance on the Food-101 dataset and UEC-Food 256 dataset, with Top 1 accuracy of 90.75% and 85.51%, respectively, and Top 5 accuracy of 98.95% and 96.13%, outperforming existing fine-grained image classification models.

Keywords: Long-tail distribution, feature augmentation, focal loss, attention mechanism.

1 Introduction

For humans, food is indispensable, but not all foods are beneficial to our health. By categorizing foods, we can identify which are harmful and which are beneficial, as well as assess the safety of food processing.

Traditional deep learning-based CNN models can classify food, but the similar appearance of different ingredients after processing makes it challenging for these models, which are designed for coarse-grained classification, to capture fine-grained distinctions. Studies have shown that fine-grained food images exhibit weaker spatial and channel structure features, making fine-grained food image classification more challenging than conventional fine-grained image classification [1]. Moreover, food image datasets often present a long-tail distribution, where a few popular food categories dominate while rare items are significantly underrepresented. This distribution imbalance can lead to biased models that underperform on rare classes, as highlighted in prior research [2]. For instance, the study by He et al.[3] introduced the Food101-LT and VFN-LT datasets to specifically address the challenges of long-tailed food classification. They proposed a two-stage approach using visual-enhanced CutMix and knowledge distillation to improve classification performance on tail categories. However, this method primarily focuses on enhancing data representation without directly addressing feature extraction improvements, limiting its ability to generalize across diverse food categories in fine-grained

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354626

settings. As a result, under-representation of certain features during training still affects the model's ability of distinguish between different food items.

Aiming to solve the long-tail distribution challenges, we propose a fine-grained long-tail distribution food image classification model with attention mechanisms, named FL-FoodNet. To fully extract the fine-grained features of food images, we integrate channel attention and spatial attention mechanisms to capture food characteristics from different dimensions, allowing the model to focus more on the fine details of the food. Therefore, the model can better focus on the intrinsic features and locations within the images. We also use focus loss to address class imbalance in food datasets by adjusting the weights of positive and negative samples to reduce the weights of samples that are already well predicted by the model, forcing the model to focus more on hard-to-predict samples, thus improving the model's generalisation ability. Our contributions include:

1. We propose FL-FoodNet, a model specifically designed for fine-grained food image classification that integrates both channel and spatial attention mechanisms. These mechanisms enhance the model's ability to extract intrinsic features from various dimensions, focusing on both the characteristics and precise locations of food items.

2. We experimentally validate the performance of FL-FoodNet in classifying food images with long-tailed distributions and compare it with a variety of existing fine-grained image classification models. The experimental results show that FL-FoodNet outperforms other models in classification accuracy on both Food-101 and UEC-Food 256 datasets, effectively solving the category imbalance problem in the dataset.

2 Related work

2.1 Food image recognition

Food image classification can be categorized into methods based on hand-crafted features and methods based on deep features. Traditional hand-crafted feature recognition methods have been largely replaced by deep learning models due to the superior performance of deep models in image processing tasks.

Earlier works, such as those by Kagaya et al [4] optimized the parameters of CNNs and applied them to food recognition tasks, but their studies relies heavily on parameter tuning, which may limit its generalizability to diverse datasets. Other researches, such as ResNet model [5] was utilized to analysis and classification of food images. However, their research primarily focused on accuracy improvements without addressing long-tail distribution issues in food categories. Random Forest-based methods [6] mainly targeted broader classifications without catering to the specificity required for fine-grained tasks. Therefore, Hassannejad et al [7] employed Deep Convolutional Neural Networks (DCNN) as it can automatically learn more complex patterns suitable for image tasks.

2.2 Approaches for long-tail distribution

Approaches for addressing long-tail distribution in food classification have evolved alongside advances in deep learning. Early methods primarily focused on adjusting class weights or

sampling strategies to balance data distributions. For example, The study on Long-Tailed Food Classification [3] introduced two datasets, Food101-LT and VFN-LT, specifically designed for tackling long-tailed food classification by incorporating power-law distributions. Despite their effectiveness in simulating class imbalances, these datasets focus mainly on data distribution and do not directly address improvements in feature extraction techniques for tail classes. To improve this, some recent studies have introduced techniques such as feature space enhancement and mixed data enhancement to improve the model's performance for rare classes under long-tailed distributions.

However, these techniques often fall short in fine-grained food image tasks, where rare categories may have distinctive features that general balancing techniques cannot capture. The LOng-tailed FIne-Grained Network for Food Recognition (LOFI) [8] proposed a feature-space augmentation strategy to enhance tail class performance in a continual learning context. Although this method showed promise in mitigating catastrophic forgetting, it primarily targets incremental learning scenarios, which may limit its applicability to static datasets. The Dynamic Mixup approaches [9], on the other hand, deals with the multi-label long-tail problem by mixing data enhancement but has limited feature differentiation ability when dealing with visually similar classes encountered in fine-grained classification tasks.

By contrast, FL-FoodNet proposed in this paper combines the channel and spatial attention mechanisms and Focal Loss to compensate for the shortcomings of traditional methods. Specifically, FL-FoodNet not only improves the sensitivity of the model to a few classes at the feature extraction level through the channel and spatial attention mechanism, but also adjusts the sample weights through Focal Loss, which enables the model to have higher generalization ability in fine-grained tasks under long-tailed distributions. This combined approach overcomes the limitations of traditional methods in feature extraction and class differentiation. FL-FoodNet can accurately focus on rare classes on unbalanced datasets, thus providing more stable classification performance in scenarios with large data diversity.

3 Method

3.1 Backbone network

This paper is based on the classic fine-grained classification model WSDAN, which is improved upon. In WSDAN, multiple sets of attention maps will be obtained by performing convolution operations on feature maps, and each set of attention maps represents the features of a particular part of the image. Bilinear attention pooling is then applied to obtain the final multi-part feature matrix. After obtaining the attention maps, two methods, crop and drop, are used to optimize the training data. Once the training is complete, the average of the attention maps is taken to find the exact position of the object in the image, which, together with the input image, is used for forward inference. The result is obtained by averaging the outcomes. However, WSDAN has limitations in feature augmentation and its attention generation is restricted to a preliminary feature representation, which leads to a higher sensitivity to noise and complex backgrounds. The model is more affected by noise when extracting key features, making it challenging in fine-grained image classification tasks.FL-FoodNet further adds CBAM (Convolutional Block Attention Module) after WSDAN has computed the attention image to refine the features generated by WSDAN and enhance the model's ability to capture important features. CBAM

has a channel attention mechanism and a spatial attention mechanism, which enable the model to effectively focus on feature recognition and suppress background noise, which is crucial for fine-grained classification tasks.

This combination brings significant benefits. Firstly, through channel attention, FL-FoodNet can more accurately capture features that are highly relevant to the classification task, thus improving classification accuracy. Second, the spatial attention mechanism helps the model to maintain focus on key features in complex backgrounds, significantly reducing the effect of noise. The integration of these two attention mechanisms enables FL-FoodNet to optimise the feature extraction process in different dimensions, which in turn demonstrates higher robustness and accuracy in the fine-grained food images classification task.

3.2 Spatial attention mechanism and channel attention mechanism

In FL-FoodNet, the attention mechanisms are seamlessly integrated with the WSDAN architecture to refine its focus on crucial details within food images. Based on Figure 1, FL-FoodNet leverages both channel and spatial attention mechanisms. The channel attention component, which aggregates features using max pooling and average pooling, emphasizes essential features by generating a channel attention map.

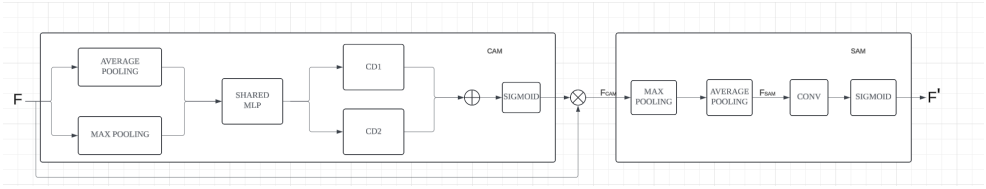

Fig. 1. The spatial attention mechanism and channel attention mechanism (Photo/Picture credit: Original).

Initially, the input feature map F, obtained from the backbone network, undergoes max pooling and average pooling then get the results spatial descriptors Fm and Fa , which are then merge into a shared multi-layer perceptron (MLP). The calculation formula for channel attention is as follows:

$$CA(x) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{1}$$

In the formula, the symbol σ represents the sigmoid function and $MLP(\cdot)$ denotes the multi-layer perceptron. The generated channel attention map multiply with F to obtain the CAM feature map FCAM, which is defined as:

$$\mathrm{F}_{CAM} = \mathrm{CA(F)} \otimes \mathrm{F} \tag{2}$$

After obtainingFCAM, then maximum pooling and average pooling are used again to calculate the space attention graph FSAM, and the output is fed into a 7x7 convolution layer, followed by a sigmoid function to get F', the final spatial attention map. The calculation formula is as follows:

$$SA(F_{CAM}) = \sigma(Conv(F_{SAM})) \tag{3}$$

$$F' = SA(F_{CAM}) \otimes F_{CAM} \tag{4}$$

3.3 Focal loss

Focal loss is a loss function designed to solve the problem of class imbalance. Compared to traditional loss functions like cross-entropy, Focal Loss offers advantages for handling class imbalance. While cross-entropy focuses equally on all samples, Focal Loss reduces the contribution of easy-to-classify samples and emphasizes harder ones, thanks to its modulation factor. It is an improvement upon the traditional cross-entropy loss function, which is represented as:

$$CE(p_t) = -\log(p_t) \tag{5}$$

Where p_t is the predicted probability of the model for the actual class. Focal Loss introduces a modulation factor, and a scaling factor based on the cross-entropy loss function, the function is shown as:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{6}$$

Where α_t represents the scaling factor used to balance the weights of positive and negative samples. It is a coefficient between 0 and 1, adjusted or set according to the proportion of each class to address class imbalance. $(1 - p_t)^{\gamma}$ represents the modulation factor, with γ being a hyperparameter greater than or equal to 0, used to reduce the contribution of easily classified samples to the loss. When γ equals 0, Focal Loss degenerates to the ordinary cross-entropy loss.

4 Experimental results

4.1 Dataset

Experiments used two food image datasets: Food-101[6] and UEC-Food 256[10]. The Food-101 dataset was published by Stanford University in 2014, consists of 101 food categories used primarily for image classification, with a total of 101 food classes and 101,000 food item images. For each category, there are 250 test images and 750 training images. The training images are uncleaned, and all images are resized to a maximum dimension of 512 pixels. The UEC-Food 256 dataset, created by Waseda University in Japan, contains food images captured from various angles and lighting conditions, with many samples per category. This dataset features diversity and complexity and includes not only food images but also detailed annotation information.

4.2 Experimental setup

The experiments were conducted on Ubuntu 18.04 operating system using an Nvidia GeForce RTX 3090 GPU. The baseline model was set to ResNet-50 in a PyTorch environment, with the SGD optimizer, the learning rate set to 0.001, batch size set to 16, and the number of epochs set to 160. During feature augmentation, all images were resized to 448x448 as input data, then randomly cropped to 224x224. Random horizontal flipping, colour jittering, and random sharpness adjustments were applied. The normalized images were then fed into the model, with the normalization mean set to [0.4569, 0.4693, 0.3175] and variance set to [0.2200, 0.2149, 0.2064].

4.3 Comparison method

In this experiment, the proposed FL-FoodNet and 5 fine-grained image classification models were tested on Food-101 dataset and UEC-Food 256 dataset. Liu et al [11] create the Deepfood model, which addresses the challenge of accurately recognizing diverse food items for computer-aided dietary assessment. However, it tends to perform poorly on underrepresented classes in datasets with long-tailed distributions due to its lack of focus on balancing class representation. The paper [12] introduced a progressive multi-granularity training method. By progressively incorporating global to local information, the model improves the recognition accuracy of subtle visual differences among similar categories. This method is effective in processing highly similar images, but it is more sensitive to background noise and may ignore global information.

Zhang et al [13] proposed RAEF, a model that combines ingredient information with visual features, improving recognition of visually similar food categories. However, this method relies on the availability of ingredient data, which limits its applicability in purely visual classification tasks. In contrast, FL-FoodNet solely relies on image data while still achieving high accuracy by enhancing feature extraction through attention mechanisms. WSRN [14] leverages weakly supervised attention learning and feature augmentation to enhance feature extraction and localization, yet it lacks robustness in noisy environments and struggles with complex backgrounds. FL-FoodNet improves upon WSDAN by incorporating CBAM, which enables more effective suppression of irrelevant features and better focus on key image details, thus improving performance in challenging visual conditions. The paper [15] presented VTNet+, a model that combines handcrafted features with deep learning methods to address food cuisine classification tasks. However, relying on manually crafted features can limit its adaptability to complex and diverse datasets. In contrast, FL-FoodNet uses a fully automated feature extraction process, integrating attention mechanisms and Focal Loss to dynamically adjust its focus on challenging and underrepresented samples.

4.4 Comparison with state-of-the-art

During the experiments, we compared the FL-FoodNet with DeepFood, WSDAN, PMG and RAEF models respectively on Food-101 dataset, and the result has shown in table1.

Table 1. Comparison of FL-FoodNet and other models on the Food-101 dataset.

Model	Top1(%)	Top5(%)
DeepFood [11]	77.40	93.70
PMG[12]	86.93	97.21
RAEF[13]	90.03	98.87
WSDAN	90.16	98.42
FL-FoodNet (ours)	**90.75**	**98.95**

As can be seen from Table 1, FL-FoodNet performs best, with a Top1 accuracy of 90.75%, slightly higher than the WSDAN and RAEF model, and significantly better than PMG's 86.93%

and DeepFood's 77.40%. In terms of Top5 accuracy, FL-FoodNet also leads with 98.95%, which shows that FL-FoodNet has a high accuracy in food fine-grained image classification.

At the same time, we conducted comparative experiments with WSDAN, DeepFood, WISeR and VTnet+ models on the UEC-Food 256 dataset. The results can be seen from Table 2.

Table 2. Comparison of FL-FoodNet and other models on the UEC-Food 256 dataset.

Model	Top1(%)	Top5(%)
DeepFood [11]	63.80	87.20
WISeR[14]	83.15	95.45
VTnet+[15]	83.43	91.94
WSDAN	83.95	93.73
FL-FoodNet (ours)	**85.51**	**96.13**

As can be seen from Table 2, FL-FoodNet ranks first in Top1 accuracy with 85.51%, which is higher than WSDAN's 83.95%, VTnet+'s 83.43% and WIseR's 83.15%, while DeepFood has the lowest accuracy of only 63.80%. In Top5 accuracy, Fl-FoodNet also leads with 96.13%, which shows that FL-FoodNet can not only predict possible categories most accurately in food fine-grained image classification, but also has the highest probability of including the correct answer in the Top5 predictions.

FL-FoodNet achieves superior Top1 and Top5 accuracy due to the enhancements introduced in FL-FoodNet for handling fine-grained food image classification with long-tail distributions. First, the model integrates channel and spatial attention mechanisms, which allow the model to extract intrinsic features from multiple dimensions, focusing on fine details and significant regions within food images. Additionally, it employs attention-guided feature augmentation, which helps the model prioritize crucial features during training. Lastly, by incorporating Focal Loss, FL-FoodNet effectively addresses class imbalance, adjusting weights to focus more on minority classes, thereby enhancing overall generalization. These contributions allow FL-FoodNet to outperform existing fine-grained classification models on the Food-101 and UEC-Food 256 datasets.

We plotted FL-FoodNet's Top1 accuracy and loss during training and testing, respectively. Figure 2 shows the accuracy and loss of FL-FoodNet's train and loss on the Food-101 dataset, and Figure 3 is about UEC-Food 256 dataset.

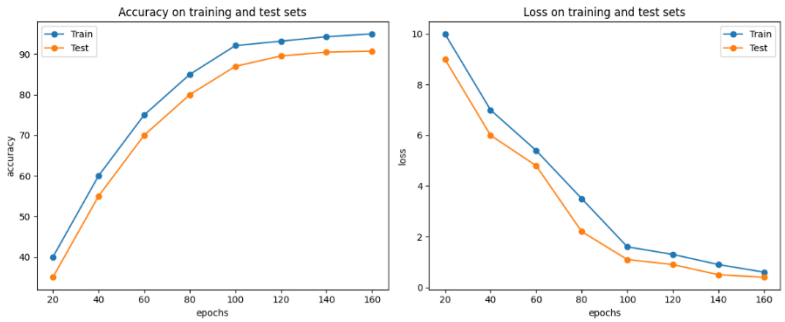

Fig. 2. Accuracy and loss on training and test sets of FL-FoodNet on the Food-101 dataset (Photo/Picture credit : Original).

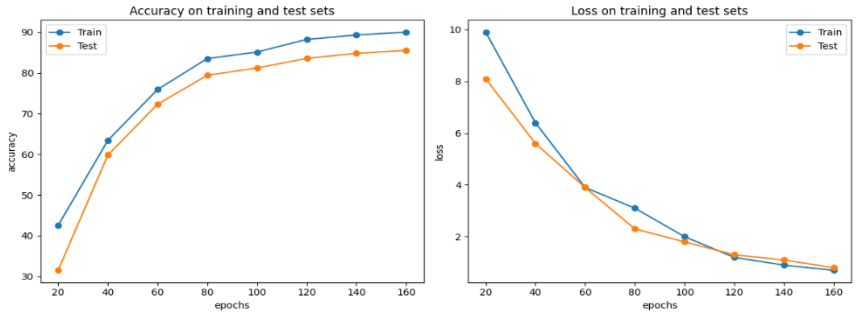

Fig. 3. Accuracy and loss on training and test sets of FL-FoodNet on the UEC-Food 256 dataset (Photo/Picture credit : Original).

As shown on Figures 2 and Figure 3, FL-FoodNet can be fully trained on both datasets. The accuracy begins to converge at the 140th iteration, and the loss gradually decreases and tends to be stable, which shows that after training, FL-FoodNet can obtain accurate food fine-grained image classification results on the test set.

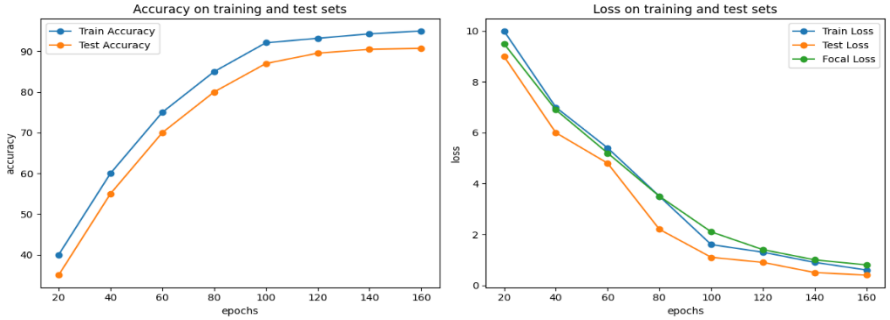

Fig. 4. Accuracy and loss on training and test sets of WSDAN on the Food-101 dataset (Photo/Picture credit : Original).

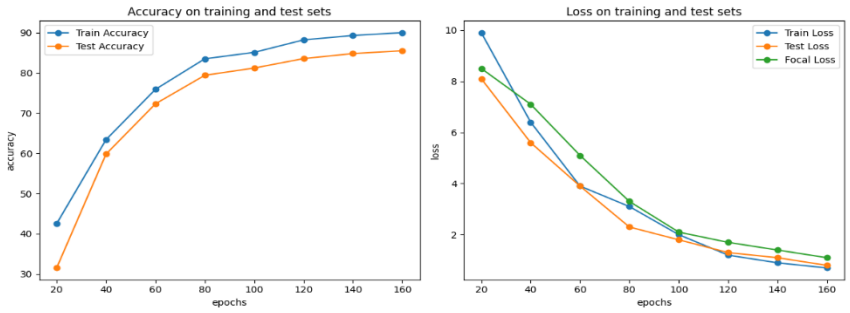

Fig. 5. Accuracy and loss on training and test sets of WSDAN on the UEC-Food 256 dataset (Photo/Picture credit : Original).

In Figures 4 and Figure5, the accuracy and loss curves of WSDAN on the Food-101 and UEC-Food 256 datasets can be seen. The diagrams clearly show that Focal Loss makes the training process of FL-FoodNet more stable compared to the traditional loss function and shows better convergence characteristics by rapidly decreasing the loss value when dealing with samples with long-tailed distributions. This effect indicates that Focal Loss effectively reducing the impact of misclassified samples, enabling the model to achieve higher classification accuracy in a shorter time.

Specifically comparing Figures 2 and Figure 4, on the Food-101 dataset, the accuracy convergence speed and stability of FL-FoodNet is significantly better than that of WSDAN (Figure. 4.) FL-FoodNet's accuracy tends to stabilise at about 140 iterations, while WSDAN takes longer. In addition, FL-FoodNet also performs better in terms of loss drop, with the loss value dropping sharply and stabilising at the beginning of training, while WSDAN's loss curve decreases slower, suggesting that FL-FoodNet's Focal Loss is more effective in coping with unbalanced datasets.

Similarly, when comparing Figures 3 and 5, FL-FoodNet still outperforms WSDAN on the UEC-Food 256 dataset. The diagrams shows that the accuracy curve of FL-FoodNet rises rapidly and stabilises, with a higher accuracy in the end, whereas WSDAN's accuracy still fluctuates in the later stages of training. In addition, FL-FoodNet's loss decreases more smoothly and stably, indicating that it reaches convergence more quickly during training and reduces the impact on hard-to-classify samples.

The comparison of these two datasets can be concluded that the application of Focal Loss in FL-FoodNet not only improves the accuracy, but also makes the loss converge faster, which indeed improves the classification performance and training efficiency of the model on the long-tailed distribution dataset.

5 Conclusion

During the research, we proposed FL-FoodNet, which is a fine-grained food image classification model specifically designed to address the challenges posed by long-tail distributions in food datasets. By integrating channel and spatial attention mechanisms, FL-FoodNet can effectively

focus on intrinsic features within food images, which improves its ability to find the difference between similar classes. Besides, the use of Focal Loss enables the model to decrease the impact of easy-to-classify samples and places more emphasis on harder, underrepresented samples, thus improving overall classification performance and generalization on imbalanced datasets. The experimental results on the Food-101 and UEC-Food 256 datasets proves that FL-FoodNet is better than other models, such as WSDAN. FL-FoodNet not only achieved higher Top-1 and Top-5 accuracies but also displayed more stable and rapid convergence in terms of loss reduction. Compared to WSDAN, FL-FoodNet's improved feature extraction and attention mechanisms provide greater robustness and adaptability to complex food images, proving its superiority in both accuracy and training efficiency. These advancements highlight the potential of attention mechanisms and tailored loss functions in enhancing model performance for fine-grained classification tasks.

Further research could study more on the integration of multi-modal data, like text and image fusion, to further improve classification accuracy. Additionally, future studies might investigate the applicability of FL-FoodNet to other domains where long-tail distributions and fine-grained classification are critical, potentially extending its benefits to broader and more diverse datasets.

References

[1] Mezgec, S., & Seljak, B. K. (2019). Using deep learning for food and beverage image recognition. 2019 IEEE International Conference on Big Data (Big Data), 5149-5151.

[2] Rao, Y., Chen, G., Lu, J., & Zhou, J. (2021). Counterfactual attention learning for fine-grained visual categorization and re-identification. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1025-1034.

[3] He, J., Lin, L., Eicher-Miller, H. A., & Zhu, F. (2023). Long-tailed food classification. Nutrients, 15(12), 2751.

[4] Kagaya, H., Aizawa, K., & Ogawa, M. (2014). Food detection and recognition using convolutional neural network. Proceedings of the 22nd ACM International Conference on Multimedia, 1085-1088.

[5] Ming, Z.-Y., Chen, J., Cao, Y., Forde, C., Ngo, C.-W., & Chua, T. S. (2018). Food photo recognition for dietary tracking: System and experiment. MultiMedia Modeling: 24th International Conference, 2018, 129-141.

[6] Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–mining discriminative components with random forests. Computer vision–ECCV 2014: 13th European conference, 446-461.

[7] Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., & Cagnoni, S. (2016). Food image recognition using very deep convolutional networks. Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, 41-49.

[8] Rodríguez-De-Vera, J. M., Estepa, I. G., Bolaños, M., Nagarajan, B., & Radeva, P. (2024). LOFI: LOng-tailed FIne-Grained Network for Food Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3750-3760.

[9] Gao, J., Chen, J., Fu, H., & Jiang, Y.-G. (2022). Dynamic mixup for multi-label long-tailed food ingredient recognition. IEEE Trans. Multimed., 25, 4764-4773.

[10] Kawano, Y., & Yanai, K. (2015). Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13. Springer, 3-17.

[11] Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., & Ma, Y. (2016). Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. Inclusive Smart Cities and Digital Health: 14th International Conference on Smart Homes and Health Telematics, ICOST 2016, Wuhan, China, May 25-27, 2016. Proceedings 14. Springer, 37-48.

[12] Du, R., et al. (2020). Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. European Conference on Computer Vision. Springer, 153-168.

[13] Zhang, R., Ouyang, D., He, L., Kuang, L., & Bai, H. (2024). Recognize after early fusion: The Chinese food recognition based on the alignment of image and ingredients. Multimed. Syst., 30(2), 93.

[14] Martinel, N., Foresti, G. L., & Micheloni, C. (2018). Wide-slice residual networks for food recognition. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 567-576.

[15] Nijhawan, R., Sinha, G., Batra, A., Kumar, M., & Sharma, H. (2024). VTnet+ Handcrafted based approach for food cuisines classification. Multimed. Tools Appl., 83(4), 10695-10715.