Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

# A fine-grained recognition technique for identifying Chinese food images

Shuo Feng, Yangang Wang [*], Jianhong Gong, Xiang Li, Shangxuan Li

*School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, 264209, China*

ARTICLE INFO

*Keywords:*
Food image processing
Automatic recognition
Fine-grained recognition
Laplacian pyramid
Bilinear pooling

ABSTRACT

As a crucial area of research in the field of computer vision, food recognition technology has become a core technology in many food-related fields, such as unmanned restaurants and food nutrition analysis, which are closely related to our healthy lives. Obtaining accurate classification results is the most important task in food recognition. Food classification is a fine-grained recognition process, which involves extracting features from a group of objects with similar appearances and accurately classifying them into different categories. In a such usage environment, the network is required to not only overview the overall image, but also capture the subtle details within it. In addition, since Chinese food images have unique texture features, the model needs to extract texture information from the image. However, existing CNN methods have not focused on and processed this information. To classify food as accurately as possible, this paper introduces the Laplace pyramid into the convolution layer and proposes a bilinear network that can perceive image texture features and multi-scale features (LMB-Net). The proposed model was evaluated on a public dataset, and the results demonstrate that LMB-Net achieves state-of-the-art classification performance.

## 1. Introduction

Texture feature extraction is a crucial aspect of research into computer vision and traditional machine learning algorithms [1–3]. With the continuous growth of image data, the application scenarios of image recognition technology are becoming increasingly widespread, including security, medical care, and autonomous driving. Its application prospects are extremely broad. Food recognition is a highly promising research direction at present [4], and its research results can affect all aspects of human health and life, such as cooking automation [5], food calorie prediction [6], terminal self-service settlement [7], nutritional assessment [8], and so on.

Traditional machine learning algorithms were used in previous food recognition systems. Firstly, it is necessary to perform image feature selection, such as color, contour, texture, etc. Then, these features need to be extracted and encoded, and finally, the encoded features are passed to the classifier for classification. Joutou et al. [9] proposed a method for food image classification by extracting color histogram features, SIFT features, and Gabor texture features, and introduced a support vector machine (SVM)-based multi-kernel learning classification approach. They adaptively adjusted the weights of these three features and combined them. While this method can accomplish basic classification tasks, it exhibits suboptimal performance when faced with images that have similar features, and the classification accuracy is also influenced by the imaging quality of food images. To enhance the robustness of the

[*] Corresponding author.
*E-mail addresses:* 202137617@mail.sdu.edu.cn (S. Feng), wangyg@sdu.edu.cn (Y. Wang), gongjh@sdu.edu.cn (J. Gong), xx_0801@163.com (X. Li), 2792708940@qq.com (S. Li).

https://doi.org/10.1016/j.heliyon.2023.e21565
Received 13 July 2023; Received in revised form 19 October 2023; Accepted 24 October 2023
Available online 31 October 2023
2405-8440/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

classification method, Kawano et al. [10] divided food images into 32 regions and encoded all RootHoG blocks and color blocks as Fisher vectors for feature representation. This block-wise encoding method increased the dimensionality of the image features, effectively improving the robustness of the classification method. On the other hand, Giovany et al. [11] proposed a method for extracting food features using the SIFT operator. Considering the high computational cost of SIFT feature processing, they employed the K-means algorithm to cluster the features, thereby reducing the computational cost. Finally, they utilized backpropagation neural networks for food image classification. Through experimental research, Fahira et al. [12] concluded that the combination of color histogram and Gabor features is more suitable for food image recognition tasks. They also tested five machine learning classification methods for food image classification, and the results showed that the random forest model had the shortest computation time while achieving the highest accuracy. The manual feature extraction methods have certain limitations as they require significant human involvement in the feature selection and design process. Additionally, this method excessively relies on parameters and image quality, resulting in a significant impact on recognition accuracy when objective conditions such as lighting environment and shooting angles of the images change. Since the emergence of AlexNet [13], food recognition has entered the era of deep learning [14,15] due to its significant assistance in image recognition tasks [16,17]. In recent years, with the advent of classic structures such as ResNet [18], GoogLeNet [19], DenseNet [20], etc., there have been a hundred schools of thought in the design of food recognition networks.

Food recognition differs from general object recognition tasks in that many foods are non-rigid objects that lack unique spatial arrangements and key points. Particularly for Chinese cuisine, many ingredients are processed into small chunks or thin strips, and many food colors are very similar. Images of the same food category may exhibit significant differences, while those of different categories may have similarities [21]. In this case, spatial and color features of images are difficult to be learned by conventional classification models, which results in poor performance of Chinese food recognition tasks in regular object classification models. Food recognition falls under the umbrella of fine-grained classification, which first discovers fixed semantic parts of certain subjects [22]. However, the vast majority of food images do not have common semantic components, making it difficult to perceive the unique semantic information in food images through existing fine-grained recognition networks.

In the field of food image processing, researchers have proposed various methods to improve the segmentation and recognition of food images. Wang et al. [23] introduced a pyramid network based on the Swin Transformer that adaptively combines local and global features, thereby enhancing the representation of global information and enriching the boundary details. Subsequently, Zhou et al. [24] proposed a semantic-guided window attention fusion framework for food recognition. This method measures the importance of each pixel and adaptively crops discriminative regions as input for the next stage. To further enhance the accuracy of food recognition, researchers conducted experimental studies on model ensemble and voting-based decision-making. Fakhrou et al. [25] integrated two models, InceptionV3 and DenseNet201, and analyzed the results using an average voting decision approach. Additionally, Tasci [26] investigated the classification performance of different combinations of CNN architectures, and the experimental results demonstrated that weighted voting operations provide more flexible and effective results for voting-based classification processes. For the task of distinguishing visually similar food items, Raju et al. [27] utilized a thermal imaging sensor to acquire thermal data and combined it with RGB images as a basis for classification, thereby adding a new dimension of data to food image features. However, a drawback of this method is that temperature information is susceptible to instability and environmental influences.

Chinese food images are characterized by rich colors, abundant details, and diverse shapes. It is difficult to comprehensively summarize these features by simply combining multiple networks or adding attention mechanisms. To address this issue, texture features need to be incorporated into the network to describe the characteristics of the images. Most of the methods currently used by researchers do not design feature extraction methods specifically for food images. So as to tackle the issues mentioned above, we propose a multiscale feature fusion network with texture feature extraction capability, called LMB-Net. Specifically, the network is divided into two main parts. To address the problem of texture feature extraction, Laplacian pyramids are fused into the first part of the network. To extract features from multiple receptive fields, features extracted by different convolution kernels are combined. Finally, the Laplacian pyramid convolution layer and multiscale feature fusion layer are combined to form the first part of the network. The second part consists of a homologous bilinear pooling layer and classification layer, which amplifies the impact of features on model judgment, aims to improve feature expression ability, and enhance the stability of the model. The contributions of this article are summarized below:

1. We have created a texture feature extraction module that can be integrated into the convolutional layers. This module uses a Gaussian pyramid to process the image and obtain images of different sampling layers, and then constructs a Laplacian pyramid by differentiating the downsampled images between layers. This module is suitable for use in end-to-end neural networks and greatly enhances the model's ability to learn texture features.
2. We cascade our self-designed multiscale extraction module based on food recognition with the texture feature module, and combine it with a bilinear pooling layer to enrich the model's learning methods for features and improve its robustness while reducing overfitting.

## 2. Related work

### 2.1. Extraction of texture features

Texture feature extraction is a crucial aspect of research into computer vision and traditional machine learning algorithms, involving methods such as gray level co-occurrence matrix (GLCM), texture filters, local binary pattern (LBP), wavelet transform, and others. These algorithms allow for the capture of information about different texture aspects through multi-scale and multi-orientation analysis of images, which can then be converted into digital feature vectors for subsequent processing and classification. Yuita Arum Sari et al. [28] used LBP and GLCM to extract texture features of food, finding that texture features were an essential part of describing food image characteristics. Puteri Khatya Fahira et al. [29] used Haralick texture features to retrieve texture features in images and incorporated this information into a random forest classifier, achieving good classification results. Claudio Cusano et al. [30] tested the classification effect of handcrafted features and CNN on texture descriptors in a food dataset, finding that the CNN method could not effectively classify image textures and proposing a mean product voting strategy to fuse prediction results from multiple methods. Pooja Sharma et al. [31] used multiple texture filters to sharpen and enhance the edges of food images to make their features more easily extractable. Samuel Verdú et al. [32] used laser scattering imaging technology to simulate the texture of fruit and inferred the storage time of the fruit using CNN technology.

Over the years, many research teams have focused on how to effectively extract texture features manually and provide them to classifiers for judgment, and have tried many methods of extraction, including the recently popular deep learning extraction method, even highlighting texture information in image imaging in order to predict more accurately. Indeed, the research work of these scholars has proved the unshakable position of texture features in food image classification, and by enhancing the extraction of food texture, more powerful classification models have been obtained. However, so far, the extraction of texture features has been done manually, and there currently exists no end-to-end model capable of seamlessly integrating texture features extracted from food images into the sensory layer of convolutional neural networks for classification. The implementation of such a solution would undoubtedly augment the learning capacity of CNNs., making the classification results more accurate. Therefore, this paper proposes a method of using the Laplacian pyramid to express texture features of food images and writes it into a module, which is integrated into CNN to improve the model's detection capability.

### 2.2. *Model design of multiple receptive fields*

Chinese cuisine features rich colors and diverse shapes. Due to the characteristics of the images, models need to consider both the overall information and exact particulars of the food images, requiring a multi-receptive field approach to evaluate the food images. Niki Martinel et al. [33] submitted a sliced convolutional structure based on ResNet and combined it with deep residual blocks to capture the vertical structural information of the dishes. Liang et al. [34] input segmented pre- and post-segmentation food images into four different branch networks that complement each other's global and local features. Min et al. [35] adopted an improved progressive training approach to learn complementary local features, and used an attention mechanism to integrate contextual and

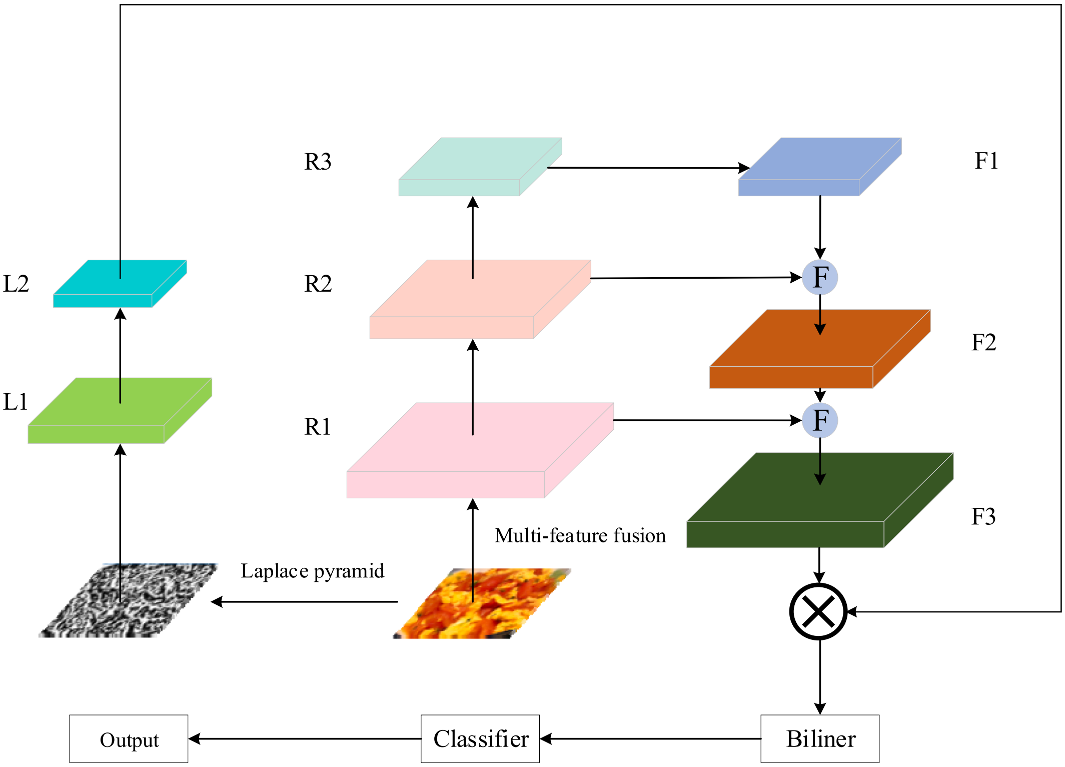


**Fig. 1.** The basic structure of this network. L1 and L2 denote two feature maps obtained by extracting from the Laplacian pyramid. R1-R3 represent the feature maps extracted from three different receptive fields. F1–F3 indicate the fused feature map layers resulting from the fusion operation. 'F' signifies the fusion operation.

multi-scale information into local features. S. Khan et al. [36] used multiple classifiers, combined all classification results, and strengthened the model's performance through a fusion strategy with learning weights. To avoid feature repetition, Xiao et al. [37] proposed a new convolution method that divides the sliding step of the convolution into odd and even steps to retrieve information, and complements the information to achieve result determination.

The achievements of the above researchers demonstrate the complexity of the elements in food images. To capture as much information as possible from Chinese food images without causing overfitting due to excessively complex extraction structures, this article uses three receptive fields combined with the texture feature extraction module mentioned earlier to form the detection part of the network.

## 3. Proposed methods

In this section, we propose a multi-scale bilinear pooling network that can extract texture features from food images. First, we use Gaussian filtering to process the image, retaining its details while reducing noise. Then, we construct a four-layer Gaussian pyramid [38], in which each layer is computed by subtracting the up-sampled and convolved prediction image from its previous layer. The texture information of a series of difference images is merged and incorporated into the perception layer of the model, along with the image information extracted by the multi-receptive field. Finally, we use the homogenous bilinear pooling layer to combine the feature maps extracted by different convolution kernels, obtaining more comprehensive feature expressions and improving the robustness of the model without increasing its complexity. Fig. 1 below illustrates the general framework of the model.

### 3.1. *Laplace pyramid module*

The Laplacian pyramid is commonly utilized in the fields of image enhancement [39], image fusion [40], feature matching [41], etc., but it is rarely involved in deep learning classification tasks. Its sensitivity to high-frequency information and texture information is exactly the starting point for the application of this method. In this paper, the image is first subjected to Gaussian filtering, with the function of the Gaussian filter shown in Eq. (1):

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

where x and y are the coordinates of pixels, σ is the standard deviation of the Gaussian function, G (x, y) is the value of the Gaussian function.

For an image, the operation of a Gaussian filter can be formulated as Eq. (2):

$$I^{'}(x,y)=\frac{1}{K}\sum_{i=-\mathrm{k}}^{\mathrm{k}}\sum_{\mathrm{j}=-\mathrm{k}}^{\mathrm{k}}I(x+i,y+j)G(i,j) \tag{2}$$

where I is the source image, I′ is the filtered image, K is a normalization factor, k is the radius of the filter, and G (i, j) is the value of the Gaussian function at (i, j). This paper chooses the smallest possible value of k to retain as much detail information in the image as possible for subsequent work.

After the Gaussian blur process, this paper designed a four-layer Laplacian pyramid to learn the texture features of the image. The Laplacian pyramid predicts the residual in image processing by upsampling and reconstructing unsampled upper-level images, resulting in an image that is larger than the original image. The upsampling process is as follows:

(1) Double the image in each direction and fill the added rows and columns with zeros.
(2) Convolve the enlarged image using the previously used kernel to obtain an approximate value for the "newly added pixels".

Let the ith layer of the Gaussian pyramid be Gi, the i+1th layer of the Gaussian pyramid be Gi+1, and the calculation formula for the i-th layer of the Laplacian pyramid Li be expressed as Eq. (3):

$$L_i=G_i-Up(G_{i+1})\otimes\kappa_{3\times3} \tag{3}$$

*Up* represents upsampling, where the pixel at position (x, y) in the source image is mapped to position (2x+1, 2y+1) in the target image.

$\otimes$ is used to represent convolution.

$\kappa_{3\times3}$ represents a kernel of size 3 × 3.

The texture information after processing with the Laplacian pyramid is shown in Fig. 2.

From Fig. 2, it can be seen that the texture information is clearly outlined, and the Laplacian pyramid used in this paper can effectively learn the texture characteristics of food images.

### 3.2. *Multi-scale feature extraction model*

Chinese food images contain diverse information and usually require the use of different sizes of convolution kernels or pooling layers to capture information of different scales in the image. Specifically, assuming the input image is $X \in R^{H \times W \times C}$, the variables H, W, and C represent the height, width, and channel dimensions of the image, correspondingly. Let the output of layer l be $F_l \in R^{Hl \times Wl \times Dl}$, where Dl denotes the output feature dimension produced by the layer. In order to perform multi-scale feature extraction, we can use multiple convolutional kernels or pooling layers in the lth layer to generate output feature maps of different scales. For example, using two convolutional kernels with sizes of 3 × 3 and 5 × 5, we can obtain two output feature maps Fl,1 and Fl,2 with a size of $H_l \times \ W_l \times D_l$. If we also add a pooling layer with a size of 2 × 2, we can obtain an output feature map Fl,3 with a size of $H_l / 2 \times \ W_l / 2 \times \ D_l$. In this way, we obtain three feature maps of different scales, each capturing different details of the input image. In convolutional and pooling layers, the parameters between multiple kernels or pooling layers are shared. Specifically, assuming that the lth layer has K convolutional kernels or pooling layers, as illustrated in Eq. (4):

$$H_l^{(1)} \times W_l^{(1)} \times C \times D_l, H_l^{(2)} \times W_l^{(2)} \times C \times D_l, \dots, H_l^{(K)} \times W_l^{(K)} \times C \times D_l \tag{4}$$

The output $F_l \in R^{H_l \times W_l \times D_l}$ of the lth layer can be represented as Eq. (5):

$$F_l = \sum\nolimits_{k=1}^{K} F_{l-1} * W_l^{(k)} + b_l^{(k)} \tag{5}$$

where Fl−1 represents the output feature map of the (l−1)th layer, * denotes the convolution operation, $W_l^{(k)}$ and $b_l^{(k)}$ represent the weight and bias of the k-th convolutional kernel or pooling layer of the l-th layer. The summation symbol ($\sum$) represents the sum of all convolutional kernels or pooling layers' outputs. This article designs three scales of convolutional receptive fields, each with two convolutional kernels. Fig. 3 demonstrates the multi-dimensional feature extraction scheme proposed in our study.

To visually showcase the efficacy of this approach in feature extraction from food images, A depiction of the extracted features of the image after being processed by the proposed extraction structure was illustrated in this paper, as depicted in Fig. 4.

As shown in Fig. 4, it can be seen that with the change of the convolution kernel, features of different scales are acquired by the model. This approach extracts both the global features and the high-level semantic information of food images.

### 3.3. Homologous bilinear pooling

In traditional Convolutional Neural Networks, pooling layers are commonly employed to decrease the dimensions of feature maps, resulting in a reduction of the model's parameters. However, in fine-grained image classification tasks, pooling operations have the potential to cause the loss of crucial details in the information, resulting in decreased classification performance. The bilinear pooling method has a unique advantage in food classification tasks due to its powerful semantic representation ability [42]. It is a linear model that combines two input features and maps them to a shared feature space to achieve the fusion and extraction of multimodal data. Let f1 and f2 be two input feature maps, and $W \in \mathbb{R}^{d \times d}$ be a learnable weight matrix. For position (i,j), the output of the bilinear pooling operation over the same source is expressed as Eq. (6):

$$Y_{i,j} = \sum_{\mathrm{m}=1}^{H} \sum_{\mathrm{n}=1}^{W} f_1(m,i) w_{(m,n)} f_2(n,j) \tag{6}$$

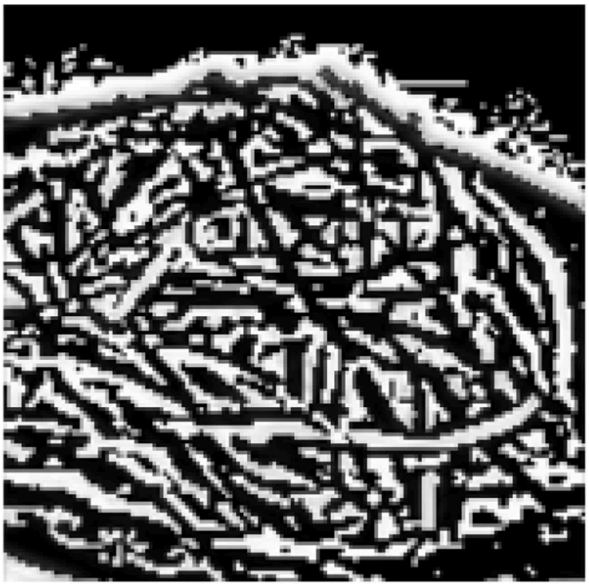


**Fig. 2.** Texture feature map after Laplacian module processing.

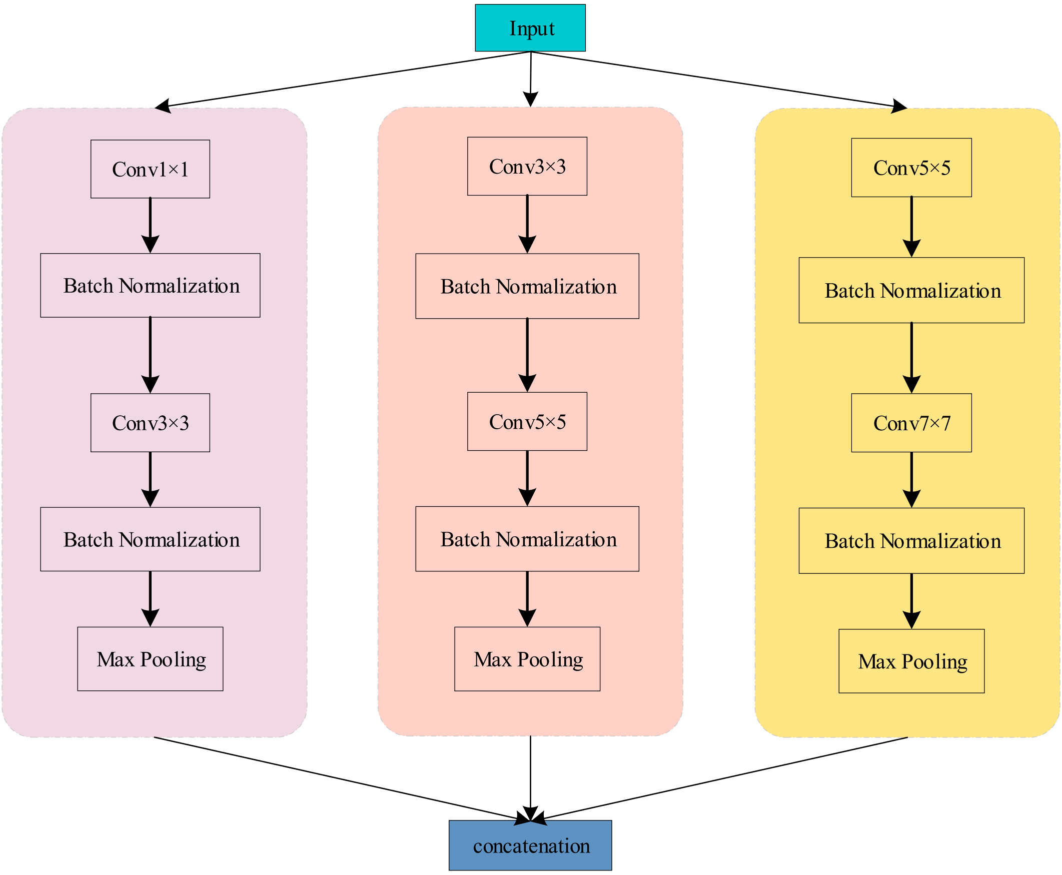


**Fig. 3.** We employed three distinct receptive fields for feature extraction. During this process, convolutional operations were performed using different kernel sizes. The specific convolutional kernel combinations consisted of 1 × 1 and 3 × 3 kernels, 3 × 3 and 5 × 5 kernels, and 5 × 5 and 7 × 7 kernels. Following the convolutional operations, we uniformly normalized the outputs through batch normalization and further compressed them through pooling.

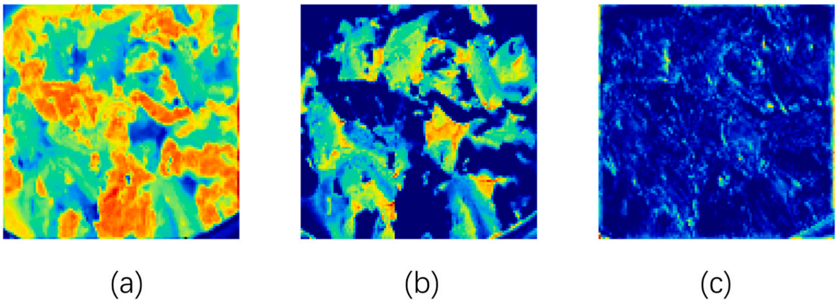


**Fig. 4.** The extracted visual feature maps. The features extracted from (a), (b), and (c) represent a gradual transition from smaller to larger convolutional kernels. The extracted image information progresses from shallow layers, encompassing contours and colors, towards deeper layers that capture abstract information. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

where H and W symbolize the height and width of feature maps f1 and f2 respectively, $w_{(m,n)}$ represents the element's value at the m-th row and n-th column of weight matrix W. Homogeneous bilinear pooling obtains a two-dimensional similarity matrix between the two feature maps f1 and f2 by performing bilinear interpolation. Then, the matrix is added to the convolution result of feature maps f1 and f2. The final matrix is normalized to obtain the pooling result Y, which can be used as the feature representation value and passed to the classification layer. The structure of the homomorphic bilinear model constructed in this study is illustrated in Fig. 5.

From Fig. 5, it can be seen that the method of bilinear pooling can capture the interaction information between different local regions while preserving local information. Furthermore, bilinear pooling, which is integrated into end-to-end classification models, can adaptively adjust the interaction weights between different regions by learning parameters, thus better adapting to different fine-grained classification tasks.

## 4. Experiments

### 4.1. Experimental setting

The testing environment is implemented using the Pytorch framework, with the Adam optimizer utilized for the model. The learning rate and weight decay are set to 0.0001 for optimization. The learning rate is updated based on the maximum validation accuracy during the training process. After every 24 epochs, if no change has been observed, the current learning rate will be decreased to one-tenth of its original value. The cross-entropy loss function is adopted for the calculation of the loss. The iteration number for training the network is 1000. Before training, all images are uniformly standardized in size, randomly affine transformed, and their grayscale values normalized.

#### 4.1.1. Dataset

In this study, we conducted tests using two datasets: ChineseFoodNet [43] and VieroFood 172 [44]. From each dataset, we randomly selected seven distinct Chinese cuisine food items, which were subsequently labeled in sequential order from 1 to 7. The dataset was then divided into three subsets: the training set, testing set, and validation set. The training set encompassed 60% of the total dataset, while the validation set and testing set comprised 20% each. To mitigate any potential bias and ensure the accuracy and reliability of the data, we implemented a random shuffling process to disrupt the image ordering within the dataset.

The food categories extracted from ChineseFoodNet dataset are as follows: Fried Potato, Green Pepper & Eggplant, Hot and Sour Potato Silk, Sautéed Diced Chicken with Chili and Pepper, Sautéed Sliced Pork, Eggs and Black Fungus, Sautéed Spicy Pork, Scrambled Egg with Tomato, and Yu-Shiang Shredded Pork. The image quantity corresponding to each partition is demonstrated in Fig. 6.

The food categories extracted from VieroFood 172 dataset are as follows: Spicy-style tofu dish、Stir-fried eggplant with green peppers、Garlic sautéed oil-seared greens、Deep-fried lotus root patties、Cold-tossed cucumber salad、Steamed spare ribs with rice flour、Stir-fried pork with bamboo shoots slices. The image quantity corresponding to each partition is demonstrated in Fig. 7.

A portion of the dataset samples is depicted in Fig. 8.

#### 4.1.2. Evaluation metric

This article will comprehensively compare the performance of models based on accuracy, recall, precision, and F1 score [45].

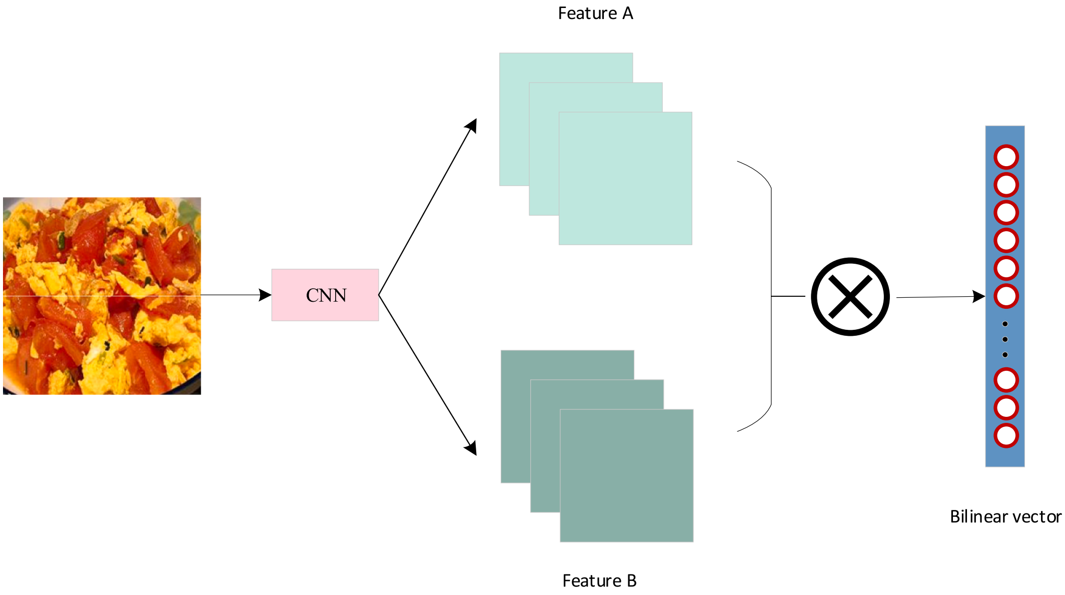


**Fig. 5.** The homogenous bilinear pooling method used in this article. In a neural network, Feature A and Feature B represent two feature map layers with identical spatial dimensions. Once subjected to bilinear interpolation, a novel feature vector called the Bilinear vector is generated. This process entails reducing the dimensionality of Feature A and Feature B, performing an outer product operation to construct a high-dimensional matrix, and subsequently conducting interpolation to acquire the Bilinear vector.

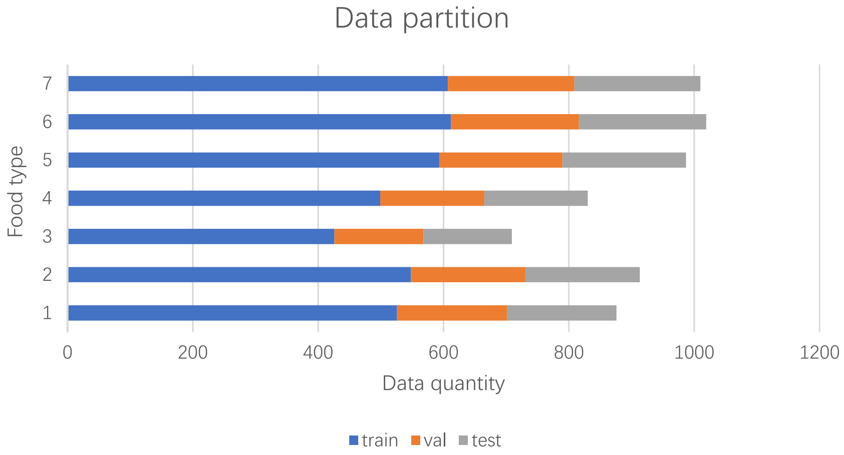


**Fig. 6.** The data structure composition of the seven food items extracted from the ChineseFoodNet dataset is presented, consisting of a total of 6344 images.

### 4.2. *Comparison of classification performance*

Within the context of this section, we compared our method with several advanced architectures and classic CNN models, including residual networks, deep networks, dense networks, multi-scale detection networks, and two networks with attention mechanisms. Furthermore, an assessment of the models' performance within the context of the test dataset was carried out and the confusion matrix of LMB-Net was generated. The classification evaluation of these methods is depicted in Tables 1 and 2.

As indicated by Tables 1 and 2, our proposed approach outperformed the compared methods in terms of accuracy, recall, precision, and F1-score. This method combined multi-feature fusion, Laplacian pyramid structure, and bilinear pooling method, which considered the texture features of food images and better mimicked the context in which humans recognize food. The confusion matrix of our proposed approach in the classification task is displayed in Fig. 9.

### 4.3. *Ablation study*

In an effort to validate the effectiveness of individual modules in the network and measure their degree of efficacy, ablation experiments were conducted on them in this article. All experiments were based on the experimental parameters mentioned in the previous section and ChineseFoodNet dataset was used for the test. The classification results of individual modules and their combinations are presented in Table 3.

According to Tables 3, it can be inferred that each module we used is an essential part of the entire model. The combination of multi-scale receptive field modules for texture feature detection, along with the information amplification function of the bilinear pooling module, greatly enhances the effect of detecting Chinese food image features in the model. As shown in Table 3, the overall accuracy of the model composed of these modules also achieved the highest performance.

## 5. Conclusion

This article analyzes the characteristics of Chinese food images and proposes a texture feature perception method based on the Laplacian pyramid. This method is cascaded with a multi-scale feature extraction module to create an end-to-end approach for texture information extraction. After integrating these two modules, we added a bilinear pooling layer that preserves local image details while also learning the interaction between the local areas, raising the accuracy level of the classification task. Experimental evidence supports the notion that, under unchanged experimental conditions, our proposed model achieves higher detection accuracy than some current advanced models and methods.

However, the current model still exhibits some limitations that require optimization in future work. Firstly, the model has a large number of parameters, necessitating a reduction in redundant parameterization by minimizing feature duplication. This will lower the overall parameter count, aligning it with the computational capacity required for future practical deployment. Secondly, the model's classification accuracy for certain food categories falls behind that of others. This discrepancy may arise from limitations in the model's feature extraction module regarding comprehension of image structure and representation. Therefore, in future research, it is crucial to explore additional features in alternative representations to enhance the model's classification performance.

## Data availability statement

Data will be made available on request.

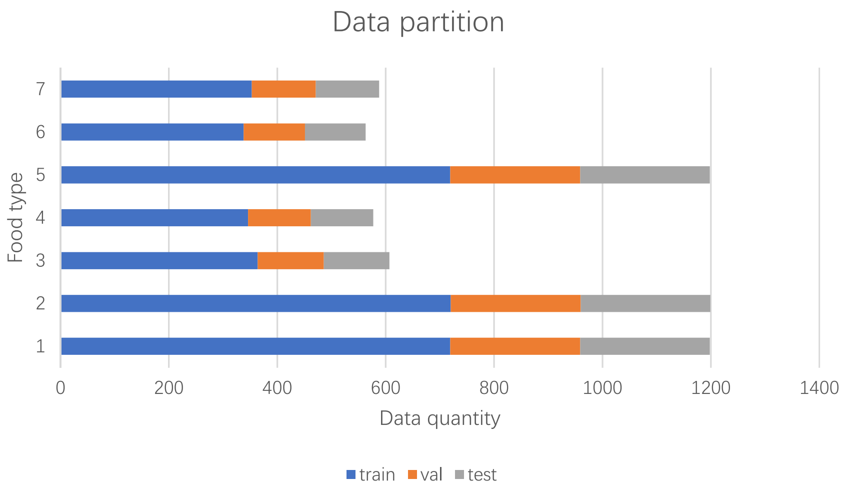


**Fig. 7.** The data structure composition of the seven food items extracted from the VieroFood 172 dataset is presented, consisting of a total of 5930 images.

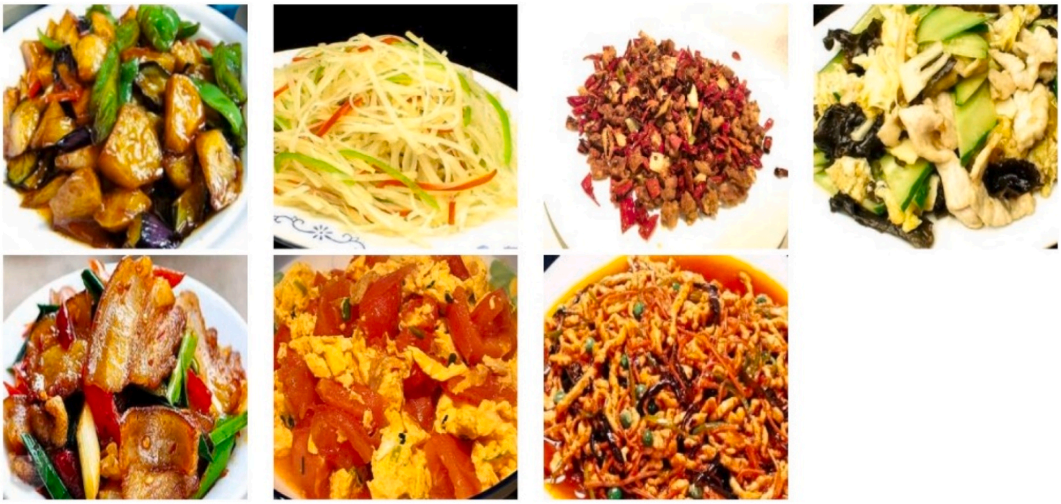


**Fig. 8.** Samples in the dataset.

**Table 1**
The classification results of all networks on the ChineseFoodNet dataset.

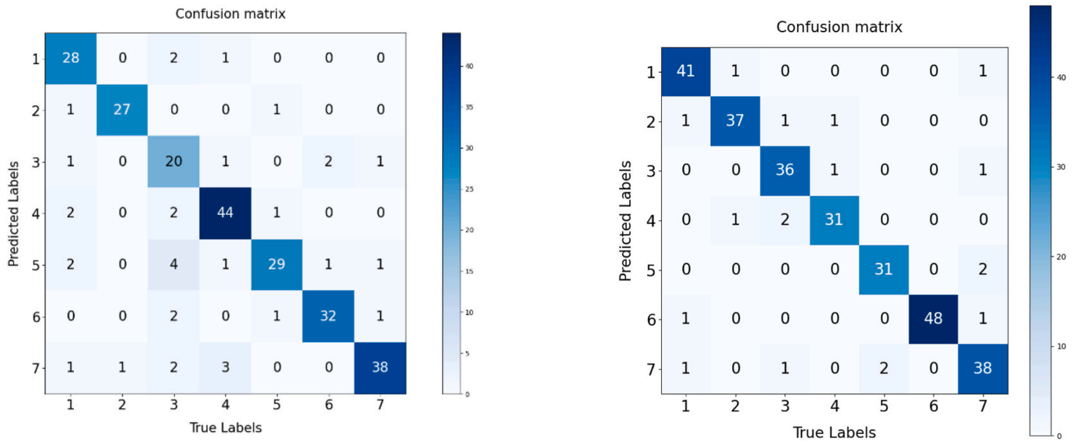
| Methods | Params(M) | ACC | Recall | Precision | F1-score |
|---|---|---|---|---|---|
| Resnet18+Bilinear [12] | 13.0 | 0.636 | 0.633 | 0.645 | 0.639 |
| Vgg19+Bilinear [39] | 287.6 | 0.735 | 0.726 | 0.756 | 0.741 |
| Densenet-121 [14] | 7.0 | 0.708 | 0.702 | 0.697 | 0.699 |
| Inception-v4 [40] | 42.6 | 0.735 | 0.732 | 0.746 | 0.739 |
| multiscale + CBAM + Bilinear [41] | 10.3 | 0.779 | 0.768 | 0.782 | 0.775 |
| multiscale + ECA + Bilinear [42] | 10.3 | 0.794 | 0.789 | 0.799 | 0.794 |
| DNPM [46] | 89.0 | 0.803 | 0.784 | 0.859 | 0.819 |
| SCG-WAFM [47] | 24.1 | 0.834 | 0.817 | 0.836 | 0.826 |
| Ours | 67.2 | 0.854 | 0.851 | 0.856 | 0.853 |

## CRediT authorship contribution statement

**Shuo Feng:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yangang Wang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Jianhong Gong:**

**Table 2**
The classification results of all networks on the VieroFood 172 dataset.

| Methods | Params(M) | ACC | Recall | Precision | F1-score |
|---|---|---|---|---|---|
| Resnet18+Bilinear [12] | 13.0 | 0.687 | 0.684 | 0.672 | 0.678 |
| Vgg19+Bilinear [39] | 287.6 | 0.765 | 0.784 | 0.743 | 0.763 |
| Densenet-121 [14] | 7.0 | 0.748 | 0.742 | 0.737 | 0.739 |
| Inception-v4 [40] | 42.6 | 0.798 | 0.791 | 0.783 | 0.787 |
| multiscale + CBAM + Bilinear [41] | 10.3 | 0.856 | 0.849 | 0.855 | 0.852 |
| multiscale + ECA + Bilinear [42] | 10.3 | 0.835 | 0.848 | 0.861 | 0.854 |
| MVANet101 [48] | 39.1 | 0.851 | 0.877 | 0.843 | 0.859 |
| SCG-WAFM [47] | 24.1 | 0.908 | 0.894 | 0.919 | 0.906 |
| Ours | 67.2 | 0.935 | 0.942 | 0.937 | 0.939 |



(a) ChineseFoodNet (b) VieroFood 172

**Fig. 9.** The confusion matrix obtained after applying the proposed method on the test sets of the two datasets. (a) and (b) represent the confusion matrix obtained from the data sets ChineseFoodNet and VieroFood 172, respectively.

**Table 3**
The ablation results.

| Multiscale | Laplacian pyramid | Bilinear pooling | ACC | Recall | Precision | F1-score |
|---|---|---|---|---|---|---|
| ✓ | – | – | 0.767 | 0.763 | 0.770 | 0.766 |
| – | ✓ | – | 0.743 | 0.738 | 0.748 | 0.743 |
| ✓ | – | ✓ | 0.771 | 0.763 | 0.764 | 0.763 |
| – | ✓ | ✓ | 0.818 | 0.811 | 0.813 | 0.812 |
| ✓ | ✓ | ✓ | 0.854 | 0.851 | 0.856 | 0.853 |

Formal analysis, Methodology, Project administration, Resources, Validation, Writing – review & editing. **Xiang Li:** Data curation, Investigation, Resources, Validation, Writing – review & editing. **Shangxuan Li:** Data curation, Investigation, Methodology, Validation, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Key Research and Development Program of Shandong Province (Grant No. 2019GSF109059).

## References

[1] T. Liu, J. Wang, B. Yang, X. Wang, NGDNet: nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom, Neurocomputing 436 (2021) 210–220, https://doi.org/10.1016/j.neucom.2020.12.090.

[2] H. Liu, H. Nie, Z. Zhang, Y.-F. Li, Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction, Neurocomputing 433 (2021) 310–322, https://doi.org/10.1016/j.neucom.2020.09.068.

[3] H. Liu, D. Li, X. Wang, L. Liu, Z. Zhang, S. Subramanian, Precise head pose estimation on HPD5A database for attention recognition based on convolutional neural network in human-computer interaction, Infrared Phys. Technol. 116 (2021), 103740, https://doi.org/10.1016/j.infrared.2021.103740.

[4] L.M. Amugongo, A. Kriebitz, A. Boch, C. Lütge, Mobile computer vision-based applications for food recognition and volume and calorific estimation: a systematic review, Healthcare 11 (2023), https://doi.org/10.3390/healthcare11010059.

[5] D.P. Papadopoulos, E. Mora, N. Chepurko, K.W. Huang, F. Ofli, A. Torralba, Learning Program representations for food images and cooking recipes, 2022-June, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (2022) 16538–16548, https://doi.org/10.1109/CVPR52688.2022.01606.

[6] S. Sathish, S. Ashwin, M. Abdul Quadir, L.K. Pavithra, Analysis of Convolutional Neural Networks on Indian food detection and estimation of calories, Mater. Today Proc. 62 (2022) 4665–4670, https://doi.org/10.1016/j.matpr.2022.03.122.

[7] S. Lan, C. Wan, L. Chen, M. Jin, S. Yu, Deep learning-based recognition of Chinese dishes in a waiterless restaurant, in: Int. Conf. Signal Process. Proceedings, ICSP. 2022-Octob, 2022, pp. 390–394, https://doi.org/10.1109/ICSP56322.2022.9965207.

[8] P. Ma, C.P. Lau, N. Yu, A. Li, J. Sheng, Application of deep learning for image-based Chinese market food nutrients estimation, Food Chem. 373 (2022), https://doi.org/10.1016/j.foodchem.2021.130994.

[9] T. Joutou, K. Yanai, A food image recognition system with multiple kernel learning, Icip 2009 (2009) 285–288.

[10] Y. Kawano, K. Yanai, Food image recognition with deep convolutional features, UbiComp 2014 - adjun, in: Proc. 2014 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput., 2014, pp. 589–593, https://doi.org/10.1145/2638728.2641339.

[11] S. Giovany, A. Putra, A.S. Hariawan, L.A. Wulandhari, Machine learning and SIFT approach for Indonesian food image recognition, Procedia Comput. Sci. 116 (2017) 612–620, https://doi.org/10.1016/j.procs.2017.10.020.

[12] P.K. Fahira, A. Wibisono, H.A. Wisesa, Z.P. Rahmadhani, P. Mursanto, A. Nurhadiyatna, Sumatra traditional food image classification using classical machine learning, in: ICICOS 2019 - 3rd Int. Conf. Informatics Comput. Sci. Accel. Informatics Comput. Res. Smarter Soc. Era Ind. 4.0, Proc., 2019, pp. 3–7, https://doi.org/10.1109/ICICoS48119.2019.8982447.

[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90, https://doi.org/10.1145/3065386.

[14] H. Liu, T. Liu, Y. Chen, Z. Zhang, Y.-F. Li, EHPE: skeleton cues-based Gaussian coordinate encoding for efficient human pose estimation, IEEE Trans. Multimed. (2022) 1–12, https://doi.org/10.1109/TMM.2022.3197364.

[15] T. Liu, J. Wang, B. Yang, X. Wang, Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom, Infrared Phys. Technol. 112 (2021), 103594, https://doi.org/10.1016/j.infrared.2020.103594.

[16] T. Liu, H. Liu, B. Yang, Z. Zhang, LDCNet: limb direction cues-aware network for flexible human pose estimation in industrial behavioral biometrics systems, IEEE Trans. Ind. Inf. (2023) 1–11, https://doi.org/10.1109/TII.2023.3266366.

[17] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, Z. Zhang, Y.-F. Li, TransIFC: invariant cues-aware feature concentration learning for efficient fine-grained bird image classification, IEEE Trans. Multimed. (2023) 1–14, https://doi.org/10.1109/TMM.2023.3238548.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, 07-12-June-2015, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (2015) 1–9, https://doi.org/10.1109/CVPR.2015.7298594.

[20] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua, 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

[21] H. Ran, G. Weidong, M. Jihang, Z. Zhenwei, Fine-Grained Recognition of Chinese Food Image Based on DenseNet with Attention Mechanism, 2021, p. 105, https://doi.org/10.1117/12.2589449.

[22] J. He, B. Yang, Research on Fine-grained Image Recognition, 7, 2023, pp. 58–64, https://doi.org/10.23977/acss.2023.070208.

[23] Q. Wang, X. Dong, R. Wang, H. Sun, Swin transformer based pyramid pooling network for food segmentation, 2022 2nd, in: IEEE Int. Conf. Softw. Eng. Artif. Intell. SEAI 2022, 2022, pp. 64–68, https://doi.org/10.1109/SEAI55746.2022.9832133.

[24] Y. Zhou, J. Chen, X. Zhang, W. Kang, Z. Ming, in: S. Yu, Z. Zhang, P.C. Yuen, J. Han, T. Tan, Y. Guo, J. Lai, J. Zhang (Eds.), Semantic Center Guided Windows Attention Fusion Framework for Food Recognition BT - Pattern Recognition and Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 626–638.

[25] A. Fakhrou, J. Kunhoth, S. Al Maadeed, Smartphone-based food recognition system using multiple deep CNN models, Multimed. Tool. Appl. 80 (2021) 33011–33032, https://doi.org/10.1007/s11042-021-11329-6.

[26] E. Tasci, Voting combinations-based ensemble of fine-tuned convolutional neural networks for food image recognition, Multimed. Tool. Appl. 79 (2020) 30397–30418, https://doi.org/10.1007/s11042-020-09486-1.

[27] V.B. Raju, M.H. Imtiaz, E. Sazonov, Food image segmentation using multi-modal imaging sensors with color and thermal data, Sensors 23 (2023) 1–15, https://doi.org/10.3390/s23020560.

[28] Y.A. Sari, F. Utaminingrum, S. Adinugroho, R.K. Dewi, P.P. Adikara, R.C. Wihandika, S. Mutrofin, A. Izzah, Indonesian traditional food image identification using random forest classifier based on color and texture features, in: Proc. 2019 4th Int. Conf. Sustain. Inf. Eng. Technol. SIET 2019, 2019, pp. 206–211, https://doi.org/10.1109/SIET48054.2019.8986058.

[29] P.K. Fahira, Z.P. Rahmadhani, P. Mursanto, A. Wibisono, H.A. Wisesa, Classical machine learning classification for Javanese traditional food image, in: ICICoS 2020 - Proceeding 4th Int. Conf. Informatics Comput. Sci., 2020, https://doi.org/10.1109/ICICoS51170.2020.9299039.

[30] C. Cusano, P. Napoletano, R. Schettini, Combining multiple features for color texture classification, J. Electron. Imag. 25 (2016), 061410, https://doi.org/10.1117/1.jei.25.6.061410.

[31] P. Sharma, S. Jadon, A. Sharma, Hybrid approach for food recognition using various filters, Int. J. Adv. Comput. Technol. 11 (2022) 1–5.

[32] S. Verdú, J.M. Barat, R. Grau, Laser scattering imaging combined with CNNs to model the textural variability in a vegetable food tissue, J. Food Eng. 336 (2023), https://doi.org/10.1016/j.jfoodeng.2022.111199.

[33] N. Martinel, G.L. Foresti, C. Micheloni, Wide-slice residual networks for food recognition, 2018, in: Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018, 2018, pp. 567–576, https://doi.org/10.1109/WACV.2018.00068.

[34] Y. Liang, J. Li, Q. Zhao, W. Rao, C. Zhang, C. Wang, Image segmentation and recognition for multi-class Chinese food school of software engineering, 1, Tongji University, Department of Endocrinology & Metabolism, Shanghai Fourth People ' s Hospital, School of Medicine, Tongji University, 2022, pp. 3938–3942.

[35] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, S. Jiang, Large scale visual food recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2023) 1–18, https://doi.org/10.1109/tpami.2023.3237871.

[36] S. Khan, K. Ahmad, T. Ahmad, N. Ahmad, Food items detection and recognition via multiple deep models, J. Electron. Imag. 28 (2019) 1, https://doi.org/10.1117/1.jei.28.1.013020.

[37] L. Xiao, T. Lan, D. Xu, W. Gao, C. Li, A simplified CNNs visual perception learning network algorithm for foods recognition, Comput. Electr. Eng. 92 (2021) 1–9, https://doi.org/10.1016/j.compeleceng.2021.107152.

[38] P.J. Burt, E.H. Adelson, in: M.A. Fischler, R. O.B.T, C.V. Firschein (Eds.), The Laplacian Pyramid as a Compact Image Code, Morgan Kaufmann, San Francisco (CA), 1987, pp. 671–679, https://doi.org/10.1016/B978-0-08-051581-6.50065-9.

[39] S. Takao, Zero-Shot Image Enhancement with Renovated Laplacian Pyramid, Springer Nature Switzerland, 2023, https://doi.org/10.1007/978-3-031-25069-9_46.

[40] G. Choudhary, D. Sethi, From Conventional Approach to Machine Learning and Deep Learning Approach: an Experimental and Comprehensive Review of Image Fusion Techniques, Springer Netherlands, 2023, https://doi.org/10.1007/s11831-022-09833-5.

[41] S. Chen, J. Zhao, Y. Zhou, H. Wang, R. Yao, L. Zhang, Y. Xue, Info-Fpn, An Informative Feature Pyramid Network for object detection in remote sensing images, Expert Syst. Appl. 214 (2023), 119132, https://doi.org/10.1016/j.eswa.2022.119132.

[42] M. Lin, Q. Chen, S. Yan, Network in network, in: 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc., 2014, pp. 1–10.

[43] X. Chen, Y. Zhu, H. Zhou, L. Diao, D. Wang, ChineseFoodNet: A large-scale Image Dataset for Chinese Food Recognition, 2017, pp. 1–8. http://arxiv.org/abs/1705.02743.

[44] J. Chen, C.W. Ngo, Deep-based ingredient recognition for cooking recipe retrieval, in: MM 2016 - Proc. 2016 ACM Multimed. Conf., 2016, pp. 32–41, https://doi.org/10.1145/2964284.2964315.

[45] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (2012) 78–87, https://doi.org/10.1145/2347736.2347755.

[46] H. Li, G. Yang, Dietary nutritional information autonomous perception method based on machine vision in smart homes, Entropy 24 (2022), https://doi.org/10.3390/e24070868.

[47] Y. Zhou, J. Chen, X. Zhang, W. Kang, Z. Ming, in: S. Yu, Z. Zhang, P.C. Yuen, J. Han, T. Tan, Y. Guo, J. Lai, J. Zhang (Eds.), Semantic Center Guided Windows Attention Fusion Framework for Food Recognition BT - Pattern Recognition and Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 626–638.

[48] H. Liang, G. Wen, Y. Hu, M. Luo, P. Yang, Y. Xu, MVANet: multi-task guided multi-view attention network for Chinese food recognition, IEEE Trans. Multimed. 23 (2021) 3551–3561, https://doi.org/10.1109/TMM.2020.3028478.