Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052174

ARTICLE

Attention Guided Food Recognition via Multi-Stage Local Feature Fusion

Gonghui Deng, Dunzhi Wu and Weizhen Chen*

School of Electrical and Electronic Engineering, Wuhan Polytechnic University, Wuhan, 430048, China

*Corresponding Author: Weizhen Chen. Email: chenwz_2005@126.com

Received: 25 March 2024 Accepted: 18 June 2024 Published: 15 August 2024

ABSTRACT

The task of food image recognition, a nuanced subset of fine-grained image recognition, grapples with substantial intra-class variation and minimal inter-class differences. These challenges are compounded by the irregular and multi-scale nature of food images. Addressing these complexities, our study introduces an advanced model that leverages multiple attention mechanisms and multi-stage local fusion, grounded in the ConvNeXt architecture. Our model employs hybrid attention (HA) mechanisms to pinpoint critical discriminative regions within images, substantially mitigating the influence of background noise. Furthermore, it introduces a multi-stage local fusion (MSLF) module, fostering long-distance dependencies between feature maps at varying stages. This approach facilitates the assimilation of complementary features across scales, significantly bolstering the model's capacity for feature extraction. Furthermore, we constructed a dataset named Roushi60, which consists of 60 different categories of common meat dishes. Empirical evaluation of the ETH Food-101, ChineseFoodNet, and Roushi60 datasets reveals that our model achieves recognition accuracies of 91.12%, 82.86%, and 92.50%, respectively. These figures not only mark an improvement of 1.04%, 3.42%, and 1.36% over the foundational ConvNeXt network but also surpass the performance of most contemporary food image recognition methods. Such advancements underscore the efficacy of our proposed model in navigating the intricate landscape of food image recognition, setting a new benchmark for the field.

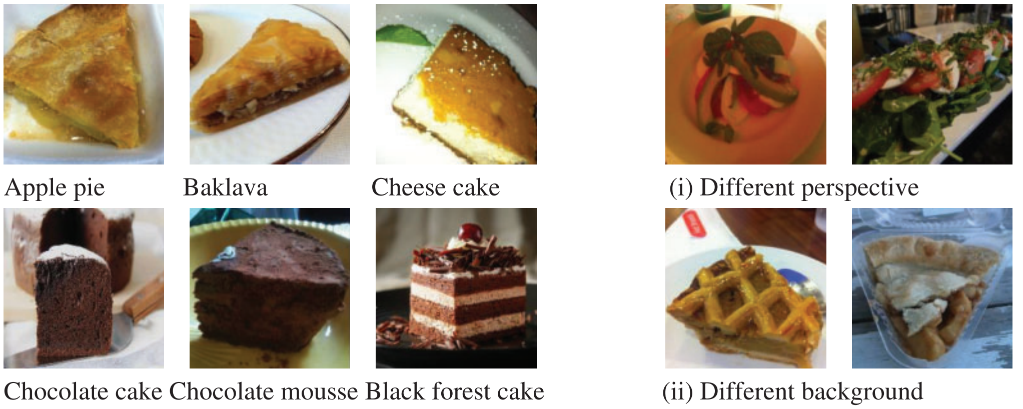
KEYWORDS

Fine-grained image recognition; food image recognition; attention mechanism; local feature fusion

1 Introduction

With the rapid development of society, food has begun to evolve in diverse directions. While many people pay attention to the taste and color of food, they often overlook the potential health risks posed by excessive caloric intake, which can lead to various latent diseases and endanger health. According to statistics, unhealthy lifestyles are one of the factors leading to human mortality, accounting for 6% of global deaths. Additionally, excessive intake of food nutrients and the lack of nutrients can cause discomfort in people's bodies, and more seriously, lead to various diseases [1]. Therefore, it is very important to have a reasonable intake of food nutrients and a regular lifestyle. Furthermore, food not only meets people's basic needs but also promotes cultural exchanges and economic development through emerging food cultures around the world. Based on this, in recent years, an increasing number of researchers have begun to devote time and energy to studies related to food, including food perception [2], food consumption [3], and food choices [4]. Due to the vastness of the research field and various existing issues, there is a lack of a systematic overview. In 2019, Min et al. [5] first proposed a systematic framework for food computing, which covers aspects such as food retrieval, food perception, food recommendation, and identification. Food identification and classification are the most fundamental and core steps, and solving this task greatly enhances subsequent tasks. Food image recognition, a subtask of fine-grained image recognition [6,7], is distinct from other fine-grained recognition tasks (such as birds, planes, and cars) due to the unique characteristics of food images. Most current food image recognition methods rely on convolutional neural networks [8–10] to extract visual features for recognition without considering the unique features of food images. Food image recognition possesses its own set of distinctive visual characteristics.

Firstly, many types of food do not have a fixed spatial structure. Food images are often non-rigid, containing significant noise and extraneous background information, which can interfere with network training, thus affecting the extraction of truly discriminative regional features. This interference tends to further reduce already small inter-class differences, ultimately leading to poor recognition performance. Various food images exhibit small inter-class differences due to similarities in ingredients and cooking methods (as shown in Fig. 1a, where each row represents different types of similar foods); moreover, food is affected by factors such as viewing angle and background, resulting in large intra-class differences (as illustrated in Fig. 1b, where each row pertains to the same food category). To address the distinct visual patterns and unfixed spatial structures of different channels in food image feature maps, this paper innovates upon the convolutional block attention module (CBAM) foundation [11]. It incorporates multiple hybrid attention mechanisms into the backbone network, extracting discriminative regional features from both spatial and channel dimensions. This enhancement boosts the network’s capability to extract key area features and reduces noise interference.

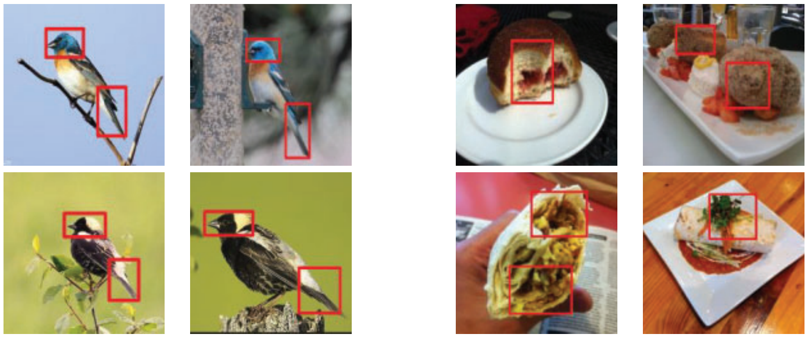


(a) Inter-class similarity of food images

(b) Intra-class differences of food images

Figure 1: Examples of inter-class similarity and intra-class difference of food images

Secondly, food images lack fixed feature patterns, and their fine-grained features are characterized by irregularities and multi-scales. Traditional fine-grained images (such as birds) possess fixed semantic patterns (for example, the tail feathers and heads of birds marked by red boxes in Fig. 2a), which can enhance recognition performance through these fixed features. Since food images do not contain fixed semantic information (as shown in Fig. 2b), it is difficult to extract common semantic features across multiple categories of food images. Moreover, fine-grained features of different food categories vary significantly in scale and shape. To address this, this paper introduces a multi-stage non-local module into the food image recognition network. By computing the feature relationships between the verification and response stages, the model establishes spatial dependencies among feature maps at different stages. This approach enables the model to learn features with multi-scale receptive fields, extracting multi-scale discriminative regional features, thereby better adapting to the irregular and multi-scale characteristics of food images. This paper's main contributions are as follows:



(a) CUB200-2011dataset (b) ETH Food-101dataset

Figure 2: Comparison of semantic information in selected images from CUB200-2011 [12] bird dataset and ETH Food-101 [13] food dataset

1. Leveraging ConvNeXt as the foundational network, we proposed a multi-attention and multi-stage local feature fusion model (MAMS-net) for food image recognition. This model employs multiple hybrid attention mechanisms to simultaneously capture discriminative and complementary features across both spatial and channel dimensions of feature maps, thereby enhancing the backbone network's feature extraction capability.

2. We introduced a multi-stage local fusion module tailored to the characteristics of food images. This module is designed to establish spatial relationships among feature maps at different stages, accessing multi-scale local features of the image. This further mines the image's deep features and improves the model's generalization ability.

3. This paper constructed a new food image dataset, Roushi60, which not only includes coarse-grained images of multiple major categories but also encompasses images of several subcategories within a major category, placing higher demands on the model's recognition capabilities.

4. Experiments conducted on three datasets have achieved excellent recognition performance, competitive with advanced models, proving the effectiveness of the methods proposed in this paper.

The basic framework of this paper is as follows. Section 2 discusses the relationship between fine-grained image recognition and food image recognition, and the research methods of both. Section 3 describes the overall framework of the model proposed in this paper, as well as detailed information about the modules proposed. Section 4 first introduces the details of the new dataset created for this study and the datasets used in the experiments. It then describes the basic configuration details of the experiments and the evaluation metrics. Section 5 validates the superiority and effectiveness of the model and modules proposed in this paper through a series of comparative experiments and ablation studies. Section 6 provides a comprehensive summary and plans for future research.

2 Related Work

2.1 Fine-Grained Image Recognition

Fine-grained image recognition refers to the identification of different sub-categories under a superclass, such as the recognition of bird species, flowers, etc. These images are characterized by large intra-class differences and small inter-class differences, especially the confusing nature of intra-class variations, making recognition challenging. Traditional fine-grained recognition methods are mainly divided into two types: strongly supervised and weakly supervised methods. For strongly supervised methods, it is essential to fully utilize various annotations and bounding boxes in the dataset to enable the network to extract effective features from target regions. Zhang et al. [14] designed modules for part-level and object-level image feature extraction under the local R-CNN framework, using dataset bounding boxes and part annotations, and ultimately fused these features for recognition. Lam et al. [15] generated part proposal boxes for images based on existing image bounding boxes, allowing the network to extract fine-grained features. Although these strongly supervised methods are somewhat effective, they are time-consuming and inefficient due to the need for extensive manual annotation of image datasets, leading researchers to shift their focus to weakly supervised learning. With the advancement of deep learning and research, weakly supervised fine-grained recognition methods that do not require additional information have emerged. Chen et al. [16] designed the destruction and construction learning (DCL) model to extract local discriminative information from food images, pruning and reconstructing the original images to improve recognition accuracy. Wang et al. [17], under the existing Transform structure, designed the mutual attention weight selection (MAWS) module, which can efficiently select discriminative image blocks without introducing additional parameters and computational burden, compensating for information captured in the network's shallower layers. Yang et al. [18] proposed the NTS-net network for locating and recognizing local features through reinforcement learning theories. Wang et al. [19] designed a graph propagation relevance learning model to mine distinctive regions, avoiding the network's neglect of the interrelations between regions. The model consists of a cross-graph propagation sub-network and a relevant feature enhancement sub-network. It captures the internal connections of discriminative feature vectors by guiding the propagation of discriminative information and suppressing meaningless vectors. However, this method suffers from prolonged model training and inference times, failing to meet practical demands. Zhang et al. [20] designed a progressive joint attention network to eliminate the prominent areas of channel enhancement, mining complementary information of target images, and forcing the network to pay attention to other discriminative areas. To drive the model to learn fine-grained representations of discriminative features, Dubey et al. [21] specifically designed a particular loss function that can effectively characterize fine-grained features and enhance the overall performance of the model. Additionally, self-supervised learning strategies have shown excellent performance in various recognition tasks, with some researchers applying these strategies to fine-grained image recognition tasks. The CAST [22] utilized GradCAM to focus on salient regions of images, which are detected by an adapted saliency detector, thereby extracting features. Cross-View Saliency Alignment (CVSA) [23] introduced a cross-view saliency alignment framework, adopting a cross-view alignment loss to motivate the model to relearn features from foreground tasks. Shu et al. [24] combined self-supervised learning with GradCAM, proposing an additional screening mechanism to identify common discriminative local features among examples and categories, achieving optimal performance in unsupervised learning tasks for fine-grained recognition. As things evolved, researchers found that the aforementioned methods did not achieve significant effects in the field of food image recognition, with some methods even degrading the performance of food recognition. For example, Min et al. [25] employed multimodal deep Boltzmann machines to fuse image information for classification. Ciocca et al. [26] used convolutional neural networks to recognize different states of food. Subsequently, many researchers devoted a considerable amount of time to studying food recognition and gradually made progress in this field.

2.2 Food Image Recognition

Currently, food image recognition mainly includes the recognition of fruits and vegetables, packaged foods, and dishes, with its application technologies touching various aspects of life, such as restaurant billing services, health management, and calorie estimation. This paper focuses on the recognition of dishes, using datasets that cover a variety of cuisines from both Eastern and Western cultures to verify the generalization ability of the proposed model. The feature extraction methods for food image recognition can be divided into manual feature methods and deep feature methods. Early research was mostly based on manually marked features, i.e., strongly supervised learning methods. With the development of science and technology and deep learning, various convolutional neural networks have been applied to food recognition. Ming et al. [9] were among the first to use the ResNet [27] network for food recognition, achieving notable success. Kagaya et al. [28] applied the AlexNet network to food recognition. Although various convolutional neural networks have been applied to food image recognition, none were specifically designed based on the characteristics of food images themselves, and their recognition effects have not reached the ideal state. Martinel et al. [29] addressed this challenge by designing the Wide Residual Network (WRN) and Slice Convolution Network (SCN) branches on top of the existing network infrastructure. The former primarily extracts global features of food images, while the latter extracts the vertical structure of images. By fusing features from both branches, they achieved the final feature representation and named this model MR, achieving state-of-the-art recognition results at the time. However, this method is only suitable for specific food recognition scenarios and lacks strong generalization capability. Nguyen et al. [30–32] have proposed a multi-task network for food image recognition to understand food recognition, counting and segmentation at the instance level and pixel level respectively. Food categories are coded on the basis of pixels, and then prior knowledge is provided for extracting features from examples. Qiu et al. [33] designed a PAR-net network by mining discriminative features of food, which consists of a backbone network and an auxiliary network. The backbone network performs basic classification, while the auxiliary network classifies the mined discriminative regions, and finally, the global and local features are fused to complete the classification. Min et al. [34] designed a cascaded multi-attention network (IG-CMAN) for food recognition by combining LSTM and spatial transformer networks, learning attention areas of different granularities from coarse to fine subnetworks. Although this network has high recognition accuracy, its structure is complex, making it difficult to train. Jiang et al. [35] fused high-level semantic features, mid-level attribute features, and deep visual features into a unified feature representation, aiming to mine granular features at different levels to complete food image recognition. Min et al. [36], in order to develop advanced food image recognition algorithms, built an ISIA Food-500 dataset with nearly 400,000 food images and designed a Stacked Global and Local Attention network (SGLAnet) that aggregates multi-scale, multi-layer discriminative features into a global representation. Liu et al. [37] used a jigsaw puzzle reconstruction module to disrupt the original image for extracting local features, also integrating a pyramid module to capture discriminative feature information. However, this method has a long model training time, which is not conducive to the subsequent plans. Ma et al. [38] introduced the Laplacian pyramid into the network based on the texture features of Chinese food images, proposed a bilinear network (LMB-Net) that perceives multi-scale features and image texture features, and achieved good recognition results. Min et al. [39] established new benchmarks for exploring food representation learning models by creating a food image dataset containing over one million images and designing a deep progressive regional enhancement network (PRENet) composed of progressive local feature learning and regional feature enhancement. Feng et al. [40] emphasized the relationship between food and ingredient information and the problem of ingredient occlusion, proposing a multitask structured network, FoodNet [41], consisting of a Multiscale Relationship Learning Module (MSRL) and a Label Dependency Learning Module (LDL). However, due to the complex composition of the models, the training time is relatively long.

The task of food image recognition has been addressed to some extent using deep learning technologies, and it has propelled the research progress in fine-grained image recognition tasks. Due to the unique texture features of food images, most of the research methods mentioned above are designed to explore the common global features of food images and do not consider the local texture features. The multiple hybrid attention mechanism we designed enables the network not only to automatically focus on the common global features but also to some extent eliminate the interference of background noise information. On this basis, we propose a multi-stage local feature fusion method to extract and synthesize the local texture information of images into more comprehensive contextual information. For food image recognition, most people use backbone networks like ResNet [27], DenseNet [42], and Transformer along with their variants [43]. The backbone network used in this paper is the ConvNeXt network, designed by Liu et al. [44] combining the frameworks of ResNet and Swin-Transformer, which merges the advantages of both, offering faster inference speeds. On this basis, we embed a multi-head hybrid attention mechanism and multi-stage local fusion method into the network to validate the effectiveness of the proposed method in the field of food recognition.

3 The Proposed Method

3.1 MAMS-Net Architecture

The overall structure of the MAMS-net model proposed in this paper is shown in Fig. 3. The entire model explores strategies of hybrid attention and local fusion. This paper first subjects the food images to data preprocessing, such as random cropping and horizontal flipping, then inputs the preprocessed images into the ConvNeXt network. Through a multi-head hybrid attention mechanism, the network mines areas of "greatest interest", namely, those that are discriminative and complementary. The images are further processed through attention modules at stage 2 and stage 3, and the outputs serve as response features for the multi-stage local fusion module. These are combined with the output of the stage 4 attention module as verification features, forming two types of input features for the multi-stage local fusion module. The MSLF (multi-stage local fusion) module outputs multi-scale feature maps, which are then passed through average pooling and fully connected layers to produce the final output. Considering that features extracted at different stages can process food images from various perspectives, such as the irregular structures and scale information of food, multi-stage local feature fusion can capture the characteristic information of food images to the greatest extent. Such a unified representation not only covers the features needed to distinguish image regions but also makes the network more robust and generalizable.

3.2 Hybrid Attention

The attention mechanism essentially mimics the human eye, which usually focuses on the most prominent objects or certain noticeable areas. Currently, the types of attention used in image tasks can be divided into soft attention, hard attention, and self-attention, among others. This paper embeds multiple hybrid attention (HA) in the ConvNeXt network to learn image features collaboratively, aiming to capture the complementary features of food images across multiple dimensions, allowing the network to learn the best hierarchical features to the greatest extent. As shown in Fig. 4, the hybrid attention module is a conjunction of spatial attention and channel attention. Hybrid attention is different from the traditional CBAM attention and shows better recognition performance in this task. In the CBAM framework, a set of feature maps is first generated through the channel attention mechanism, followed by the generation of corresponding spatial feature maps. Although this method can capture certain feature information, many critical features might be lost during this process. However, the HA module in this paper calculates and merges the attention features of both space and channel simultaneously, effectively reducing the loss of feature information and making the operation of the HA module more efficient and concise.

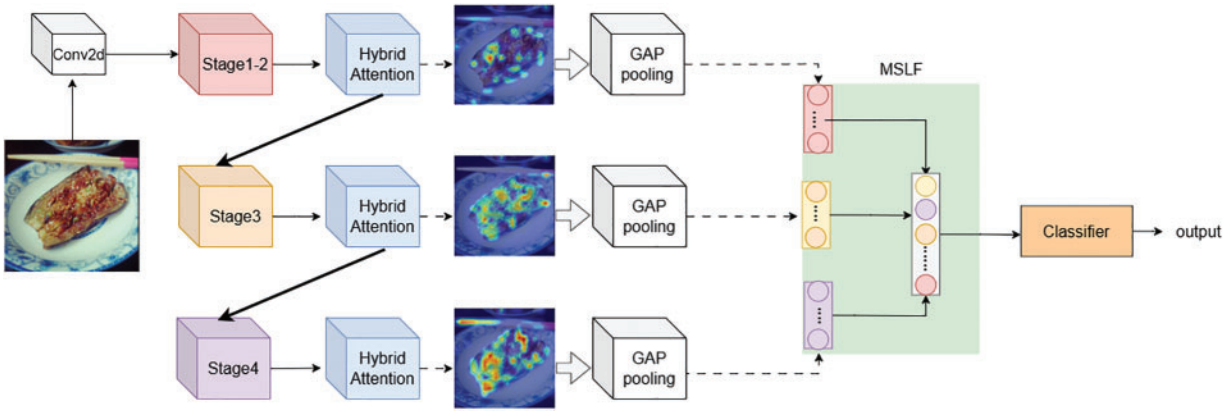


Figure 3: Overall Structure of the MAMS-net model. The ConvNeXt consists of Conv2d and four stages, each stage comprising a downsampling layer and ConvNeXt blocks. The structure of the ConvNeXt block is similar to that of the ResNet block; GAP pooling refers to global average pooling

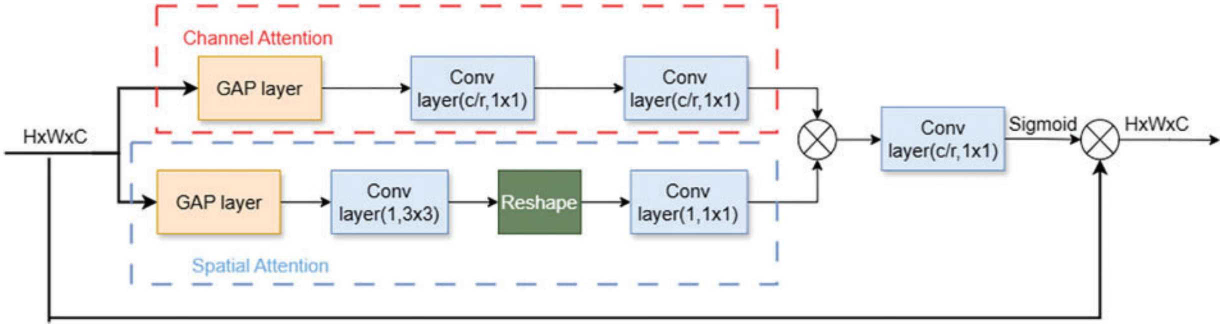


Figure 4: Hybrid attention module structure

The paper defines the input $K^n \in R^{h\times w\times c}$ of HA, where n denotes the layer number of MAMS-net, and h, w, c represent the height, width and channel number of the feature map, respectively. HA mainly generates a weight map $A^n \in R^{h\times w\times c}$ of the same size as K^n, calculated as follows:

$$A^n = X^n \times Y^n \tag{1}$$

where $X^n \in R^{h\times w\times c}$ and $Y^n \in R^{h\times w\times c}$ represent the spatial attention map and the channel attention map, respectively.

Spatial attention focuses on the significance in the spatial dimension, essentially achieved through average pooling operation as follows:

$$X^n = \frac{1}{c}\sum_{i=1}^{c} K^n_{1:h,1:w,i} \tag{2}$$

To further enhance the integration effect of channel attention, we introduce an additional convolutional layer to merge feature map information. The calculation of channel attention is as follows:

$$Y^n_{input} = \frac{1}{h \times w}\sum_{i=1}^{h}\sum_{j=1}^{w} K^n_{i,j,1:c} \tag{3}$$

This operation mainly aggregates the spatial feature information into the channel. The total information of inter-channel modeling in this operation is:

$$Y^n = Relu\left(V_2^{ca} \times Relu\left(V_1^{ca} Y^n_{input}\right)\right) \tag{4}$$

where $V_1^{ca} \in R^{\frac{c}{r} \times c}$ and $V_2^{ca} \in R^{c \times \frac{c}{r}}$ represent the parameter matrix of two convolution layers, and r denotes the fading rate. After merging the two attention maps of space and channel, the convolution layer $1 \times 1 \times c$ was added to calculate the integrated hybrid attention and a sigmoid function is applied to normalize the values between 0.5 and 1.

After obtaining the weight map A^n, a new feature map F can be further obtained, calculated as:

$$F = K^n \times A^n \tag{5}$$

The discriminative and complementary feature maps F thus obtained will be fed into the MSLF module to further extract multi-scale features of the image and enhance the receptive field.

3.3 Multi-Stage Local Fusion

If the discriminative local features obtained through the HA module are directly output as the final result, it would limit the network's ability to compare clues from different scales, preventing the model from adapting well to the irregular and multi-scale characteristics of food images. Therefore, this paper interacts and models the outputs of HA modules under different stages to capture the dependency relationships among feature maps at each stage. The overall structure of the proposed MSLF is shown in Fig. 5. Firstly, the outputs of the HA modules from stage 2, stage 3 and stage 4 are divided into two branches and input into the MSLF, with the first two outputs F^{r1} and F^{r2} serving as response features, and the last output F^v as verification features.

To capture the long-range dependencies of the feature maps, it is necessary to compute the spatial positional relationships between the three outputs of the attention module, as shown in Eq. (6):

$$\theta = W_\theta F^v \tag{6}$$

where W_θ is the weight matrix that needs to be learned. Given the spatial dimensional differences among the outputs of the three attention modules, it is necessary to learn their respective weight matrices individually, as expressed in Eqs. (7) and (8):

$$\varphi = W_\varphi F^{v(r)} \tag{7}$$

$$g = W_g F^r \tag{8}$$

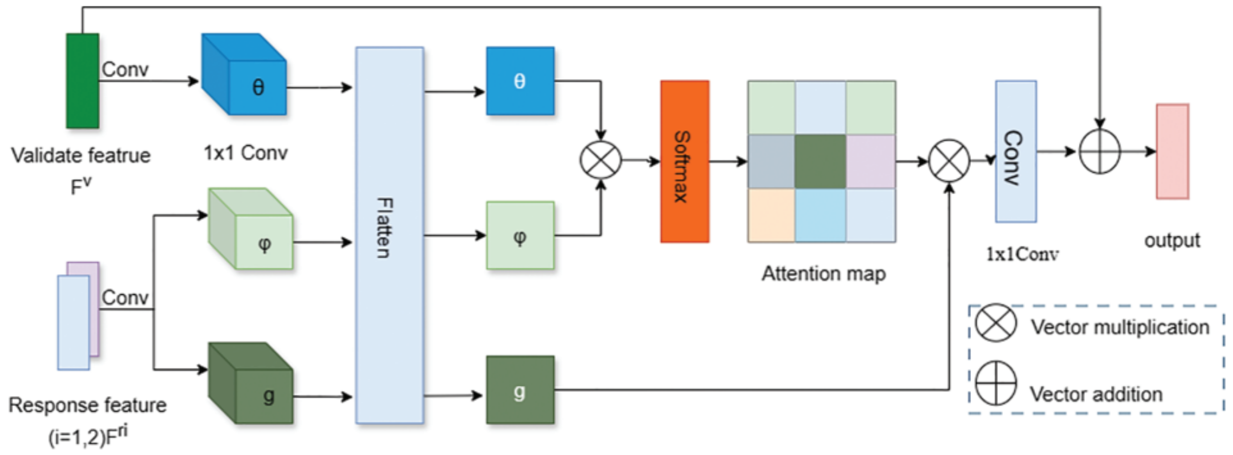


Figure 5: Overall structure of multi-stage local fusion

After projecting the three feature maps into the embedding space, the original feature maps are reduced to half their original channel size, and the height and width of the feature maps are flattened into H × W. The flattened verification feature maps and response feature maps are then subjected to multiplication operations. The purpose of this is to compute the similarity of this matrix, denoted as (f), with the formula for (f) shown as Eq. (9):

$$\mathrm{f}\left(\mathrm{F}^{\mathrm{v}},\mathrm{F}^{\mathrm{r}}\right)=\theta\left(F^{v}\right)^{\mathsf{T}}\varphi\left(F^{r}\right) \tag{9}$$

After applying softmax, an attention mapping is obtained and then multiplied by the response feature matrix. Afterward, a 1 × 1 convolution is added to the original verification feature to obtain the final output of MSLF. The final convolution ensures that the size of the matrix obtained in the previous step remains unchanged. The final output of $\mathrm{MSLF}_{\mathrm{out}}$ is shown as in Eq. (10):

$$\mathrm{MSLF}_{\mathrm{out}}\left(\mathrm{F}^{\mathrm{v}},\mathrm{F}^{\mathrm{r1}},\mathrm{F}^{\mathrm{r2}}\right)=F^{v}+z\left(f\left(\theta\left(F^{v}\right),\varphi_{1}\left(F^{r1}\right)g_{1}\left(F^{r2}\right)\right)\right) \tag{10}$$

where z is the last convolution operation.

As the neural network deepens, the receptive field of the outputs from the later network layers increases, containing more information from the original input image and extracting more feature information, whereas the opposite is true for the outputs of the shallower layers. Therefore, this paper uses the output of stage 4 as the verification feature in conjunction with the shallower features to capture the dependency relationships among the feature maps at each stage. This approach extracts multi-scale information of image features, allowing the network to better adapt to the characteristics of food images.

3.4 Loss Function

To determine the closeness of the actual output to the expected output and to accelerate the convergence of the network training process, this paper selects the multi-class cross-entropy loss. Its mathematical expression is as follows:

$$L_{CE}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K}y_{ic}\log\left(h_{\theta}\left(x_{i}\right)_{c}\right) \tag{11}$$

where N represents the number of training set samples, K is the number of categories; y_{ic} is the one-hot encoding of the sample target value, taking 1 if the correct category of sample x_i is c, otherwise 0; $h_\theta\,(x_i)_c$ represents the predicted probability that sample x_i belongs to category c.

4 Experiments

4.1 Datasets

We have noticed that Bossard et al. [13] have published several large-scale food image datasets. In this paper, we select two typical datasets, ETH Food-101 and ChineseFoodNet [45], for method validation. Since the official distribution of the Food-101 dataset includes only a training set and a test set, we divide one-third of the Food-101 training set into a validation set to determine the optimal model and reduce training time. The model achieving the highest accuracy in the validation set is then used on the test set. To validate our proposed method's capability to explore various food cues in different real environments and mine inconspicuous or missing discriminative features, we created the Chinese meat dataset "Roushi60", which contains 60 categories. the new dataset primarily consists of specialty dishes and foods from various provinces in China. These foods have unique characteristics and have little overlap with the meat dishes featured in the ChineseFoodNet dataset, and the meat in the ChineseFoodNet dataset has similar characteristics. Approximately half of these images come from Google and recipe websites, sourced through direct searches for food names, with about 90 images per category. We manually selected representative images and discarded irrelevant ones. We captured the other half of the images with smartphone cameras from various angles, depicting actual foods in real life. We designed the entire dataset to contain as diverse and complex an environment as possible to simulate the recognition performance of our model in real-world conditions. These chinese meat has many cooking styles, some cooking styles can completely destroy the discriminative characteristics of food, and even have great challenges for human recognition. To ensure that the images resemble natural conditions, background noise information is preserved and manually selected to ensure it is relevant to food image. In Roushi60, the total number of pictures is 5395, and the pictures of each category are randomly divided into training, validation and test at a ratio of 8:1:1. Fig. 6 presents some examples from the Roushi60 dataset. Table 1 provides specific information about the three datasets.

4.2 Model Training

The paper presents the MAMS-net model, which was tested in a series of experiments conducted on a server equipped with a 48 GB NVIDIA A40 GPU and an Intel Xeon CPU. The software used was Python 3.7 and the PyTorch framework.We use the pretrained weights from the ImageNet22k dataset for parameter initialization. During the training phase, the input image size is randomly cropped (1-crop) to 224 × 224, followed by random horizontal flipping for image augmentation. In the testing phase, the image size is adjusted to 256 × 256, then center cropped to 224 × 224. This experiment uses AdamW as the optimizer, with an initial learning rate set to 0.005 and learning rate decay set to 0.0005. A learning rate warm-up is applied in the first epoch, with the learning rate multiplied by 0.1 at the 10th epoch. The batch size is set to 16, and the number of epochs for training each dataset is set to 100, employing early stopping to prevent overfitting.

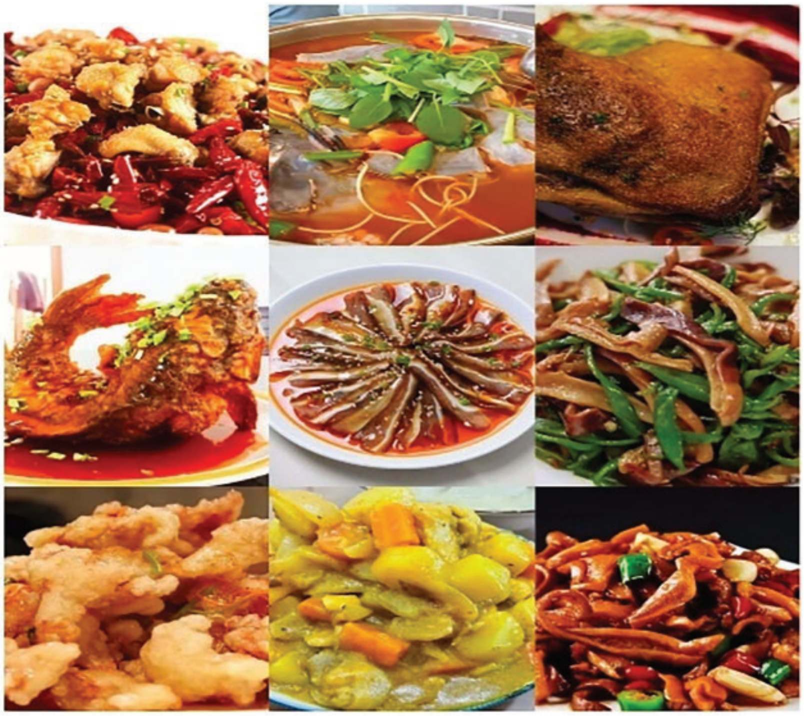


Figure 6: Some examples from the Roushi60 dataset

Table 1: Specific information about the three datasets

Dataset	Classes	Images	Training	Validation	Test
ETH Food-101 [13]	101	101000	53025	22725	25250
ChineseFoodNet [45]	208	167400	145066	20254	20310
Roushi60	60	5395	4316	532	547

4.3 Model Evaluation

The classification effectiveness of the MAMS-net model is evaluated based on the Top-1 and Top-5 accuracy rates on the test. The Top-1 accuracy rate refers to the proportion of images correctly predicted relative to the total number of images in the test, while the Top-5 accuracy rate refers to the proportion of images for which the correct category is among the top five predicted categories by output probability, relative to the total number of images in the test.

5 Results

5.1 Comparison Experiments

This section primarily evaluates the identification performance of the MAMS-net model and verifies the superiority of the methods presented in this paper. This section compares the model recognition effect proposed in this paper with advanced food image recognition methods. Tables 2–4 respectively show the comparison of recognition accuracy between the model proposed in this paper and advanced food image recognition models on the ETH Food-101 dataset, ChineseFoodNet dataset, and Roushi60 dataset. According to the comparative experiments in Tables 3 and 4, the methods proposed in this paper have superior Top-1 accuracy on both the ChineseFoodNet and Roushi60 datasets, with the highest Top-1 and Top-5 accuracies highlighted in bold. From the comparative experiment results in Table 2, although there is a subtle gap in recognition performance on the ETH Food-101 dataset compared to the literature [39], the image resolution input into the network in literature [39] is 448 × 448, which is higher than the resolution of images input in this paper, making the model training process more friendly and reducing the model's inference time. The accuracy of the method proposed in this paper on the three datasets has improved by 1.04%, 3.42%, and 1.36% respectively compared to the original ConvNeXt-B backbone network, demonstrating the effectiveness and superiority of the combined multi-head mixed attention and multi-stage local fusion strategies presented in this paper.

Table 2: Comparison of recognition effects on ETH Food-101 (%)

Method	Top-1 Acc	Top-5 Acc
DCL (ResNet50) [16]	88.90	97.82
WISeR (WRN) [29]	90.27	98.71
PARNet (ResNet101) [33]	90.40	–
IG-CMAN (SENet-154) [34]	90.40	98.42
MSMVFA (ResNet) [35]	90.59	98.25
SGLANet (Vgg) [36]	90.33	98.20
MJR-Net (ResNet50) [37]	90.82	98.32
PRENet (SENet-154) [39]	**91.13**	98.71
ConvNeXt-B [44]	90.08	98.57
TL model (ResNet) [46]	80.00	–
Ours	91.12	**98.77**

Table 3: Comparison of recognition effects on ChineseFoodNet (%)

Method	Top-1 Acc	Top-5 Acc
ResNet50 [27]	75.82	94.22
IG-CMAN (SENet-154) [34]	81.97	97.02
MSMVFA (ResNet) [35]	81.94	96.94
DenseNet121 [42]	76.85	94.91
ConvNeXt-B [44]	79.44	95.71
ChineseFoodNet (DenseNet) [45]	81.43	96.73
Fusion Model (ResNet) [47]	79.8	97.0
MVANet264 [48]	82.42	97.33
Ours	**82.86**	**97.83**

Table 4: Comparison of recognition effects on Roushi60 (%)

Method	Top-1 Acc	Top-5 Acc
ResNet50 [27]	89.12	97.79
ResNet101 [27]	90.65	98.34
DenseNet121 [42]	90.86	98.79
ConvNeXt-B [44]	91.14	98.93
Ours	**92.50**	**99.27**

5.2 Ablation Study

In this part, we validate the effectiveness of our proposed method through extensive experiments. The ablation study comprises two parts: the proposed modules ablation and HA insertion position ablation.

5.2.1 The Proposed Module Ablation

This section analyzes the ablation experiments on the ETH Food-101, ChineseFoodNet, and Roushi60 datasets, with the results of these experiments presented in Table 5. As indicated by Table 5, the modules proposed in this study enhance model performance to some extent. The classification accuracies on the backbone network ConvNeXt for the three datasets are respectively 90.08%, 79.44%, and 91.14%. The proposed Hybrid Attention (HA) module significantly improves model performance, with increases of 0.7%, 1.24%, and 0.64% respectively over the base model. This demonstrates that the Hybrid Attention module is more effective in extracting discriminative and complementary features from images compared to the base model, capturing deeper feature information. Furthermore, the effectiveness of the Multi-Stage Local Fusion (MSLF) module is verified. Based on the backbone network, it increases the classification accuracies for two datasets by 0.75% and 2.12%, respectively. This indicates that the module can integrate features output at multiple stages and extract multi-scale feature information, further distinguishing the differences between categories and better adapting to food image recognition. Finally, combining the two modules to create the MAMS-net model achieves optimal performance on the ETHFood-101 dataset, and similar experimental phenomena are observed on the ChineseFoodNet and Roushi60 datasets, with recognition accuracies of 82.86% and 92.50%, respectively. The recognition effect is higher than that of the individual modules, indicating that the two modules are complementary, compensating for each other's shortcomings and maximizing the recognition accuracy.

Table 5: Module ablation experiments on three datasets

Method	ETH Food-101	ChineseFoodNet	Roushi60
	Top-1 Acc (%)	Top-1 Acc (%)	Top-1 Acc (%)
ConvNeXt-B	90.08	79.44	91.14
ConvNeXt-B + HA	90.78	80.68	91.78
ConvNeXt-B + MSLF	90.83	81.56	91.93
MAMS-net	91.12	82.86	92.50

5.2.2 HA Insertion Position Ablation

This study hypothesizes, based on practical experience, that the HA (hybrid attention) module would achieve optimal results when applied to the intermediate layers of the backbone network. To validate this hypothesis, we inserted the HA module at different positions after each stage and conducted a series of experiments. The ETH Food-101 dataset, being one of the classic food image datasets, was chosen for its representative experimental results, and the findings are presented in Table 6. Initially, the HA module was inserted individually after each stage to assess the classification performance of each attention mechanism separately. According to Table 6, except for a performance decline when HA was inserted after stage 1, improvements were noted after all other stages. The potential decline in performance after stage 1 might be due to the network's shallow layers not yet fully learning the basic features of the image; adding HA at this point might introduce noise information, affecting overall performance. Further testing with various stage combinations revealed that inserting HA after stages 2, stage 3 and stage 4 achieved the best result, with a recognition accuracy of 90.78%. Similarly, the same insertion points for the HA module were adopted for tests on the ETH Food-101 and Roushi60 datasets.

Table 6: Experimental results on ETH Food-101 dataset after HA module insertion at different stages

HA	Acc/%
Stage 1	89.98
Stage 2	90.68
Stage 3	90.70
Stage 4	90.71
Stage 2 + stage 3	90.76
Stage 2 + stage 4	90.76
Stage 3 + stage 4	90.75
Stage 2 + stage 3 + stage 4	**90.78**

5.2.3 Explore the Performance and Model Complexity Analysis in Different ConvNeXt Variants

This section primarily verifies the effectiveness of the method proposed in this paper on various ConvNeXt variants, using models such as ConvNeXt-tiny, ConvNeXt-small, and ConvNeXt-large, which correspond to the abbreviations ConvNeXt-T, ConvNeXt-S, and ConvNeXt-L in the table, respectively. Other experimental setups and parameters remain consistent with those used in the validation on ConvNeXt-base (ConvNeXt-B). The experimental results, as shown in Table 7, indicate that as the ConvNeXt model sizes increase, the recognition accuracy on three datasets also gradually improves. However, while the largest recognition accuracy was achieved using the ConvNeXt-L model, the improvement over the earlier models was marginal, and its parameter count is more than double that of the ConvNeXt-B model. Therefore, considering economic and time costs, this paper selects ConvNeXt-B as the backbone network. The ConvNeXt-B model maintains a high level of recognition accuracy while reducing model inference time. Additionally, the experiments in this section provide diverse options; when time and economic costs are not considerations, opting for larger architectural models can pursue higher recognition accuracy.

Table 7: Experimental results of ConVvNeXt variants and complexity analysis of the model (%)

Method	ETH Food-101	ChineseFoodNet	Roushi60	Params (M)	Flops (G)
	Top-1 Acc	Top-1 Acc	Top-1 Acc		
ConvNeXt-T	89.98	80.54	90.94	35.6	4.62
ConvNeXt-S	90.49	81.93	91.78	56.6	8.82
ConvNeXt-B	91.12	82.86	92.50	95.6	15.52
ConvNeXt-L	91.31	83.06	92.73	204.6	34.52

5.2.4 Effect of ImageNet-1k and ImageNet-22k Pre-Training Weights on Model Recognition

This section primarily discusses the impact of ImageNet-1k and ImageNet-22k weights on the MAMS-net model proposed in this paper, with experimental validation results presented in Table 8. As shown in Table 8, using the pretrained weights from the larger dataset can enhance the model's recognition performance to some extent, and changes in the pretrained weights do not affect the model's parameter size, as validated by experiments in reference [44]. Therefore, without increasing the model's parameter count, this paper opts to use ImageNet-22k pretrained weights for higher recognition accuracy.

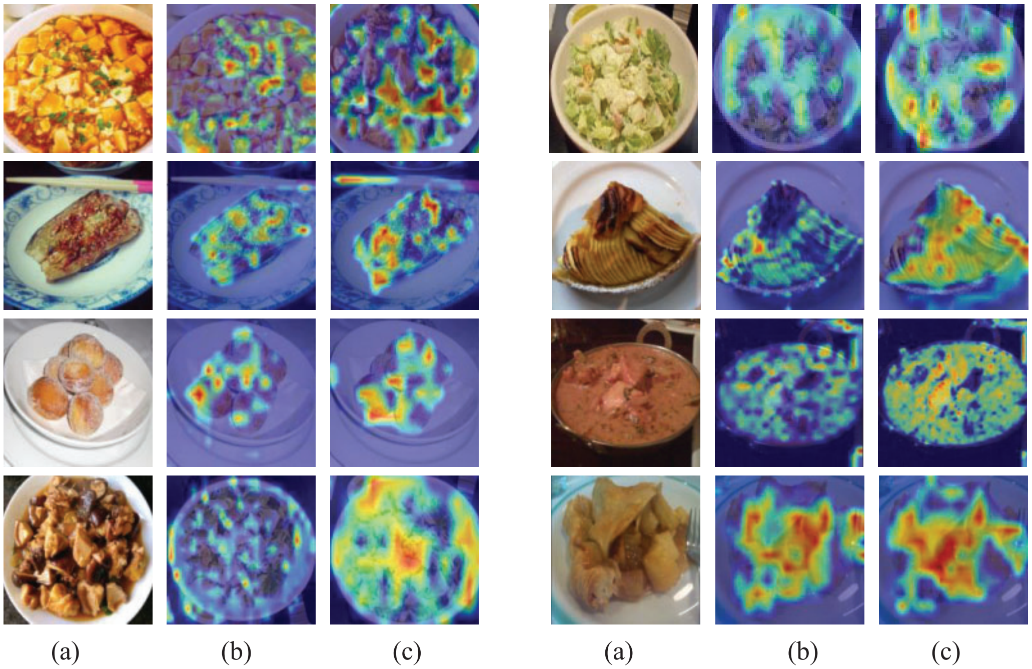
Table 8: The influence of ImageNet1-k and ImageNet-22k pre-training weights on the model

Weight (ImageNet)	ETH Food-101	ChineseFoodNet	Roushi60
	Top-1 Acc	Top-1 Acc	Top-1 Acc
MAMS-net(1k)	91.06	82.74	92.32
MAMS-net(22k)	91.12	82.86	92.50

5.3 Qualitative Evaluation

To further illustrate the effectiveness of the methods described in this paper, we utilized GradCAM [49] for visualization experiments on three datasets, as shown in Fig. 7. The outputs shown are from the last layer of the network, with each row displaying a heatmap of the output from the last convolutional layer of the backbone network under different module additions. Fig. 7a represents the original image, Fig. 7b shows the heatmap with only the HA module added, and Fig. 7c displays the feature heatmap with the MSLF module added on top of the HA module (the MAMS-net model proposed in the paper). The heatmap gradually changes from blue to red, indicating an increasing focus by the network on these areas.

From Fig. 7b, it can be seen that the HA module introduced in this paper enhances the image features by interacting and modeling across spatial and channel dimensions, thereby strengthening visual features. To a certain extent, it discards background noise interference and directly mines apparent fine-grained features, allowing the network to locate discriminative regions. As food images are non-rigid and lack fixed semantic features, distinguishing similar food images is challenging. However, after introducing the MSLF module, the backbone network can learn features of different granularities and acquire a wealth of multi-scale discriminative features. As shown in Fig. 7c, the heatmap obtained after multiple stages of local feature fusion highlights the details of various scales better than simply introducing a mixed attention mechanism (as shown in Fig. 7b). This indicates that the multi-stage local fusion module has learned fine-grained features beyond the attention mechanism, and the two modules mutually enhance each other, allowing the network to precisely locate the discriminative regions of food, reducing the interference of background noise.



(a) Original image (b) Heat map obtained by the HA module (c) Heat map obtained by MAMS-net

Figure 7: Visual result map

However, the method described in the paper exhibits a certain gap in performance compared to state-of-the-art methods on the Food-101 dataset. Our analysis of the images incorrectly identified reveals that, compared to the other two datasets, the Food-101 dataset has a higher occurrence of images with overlapping discriminative regions and mutual obstructions, which most of the misidentified images contain. These features are not effectively recognized by our method. Additionally, the method proposed in this paper requires significant computational power to successfully complete the recognition process, indicating that its practical application is still some distance away. In future research, our research will focus on two aspects: 1) Using ingredient information to guide the recognition of food images, which can alleviate problems caused by discriminative regions being obscured or overlapped to some extent; 2) With the increasing demand for applying food recognition on mobile and edge devices, where the entire process needs to be fast and convenient, we plan to design a lightweight network for food image recognition.

6 Conclusion

In this paper, we design a multi-head hybrid attention (HA) mechanism on the basis of the baseline network ConvNeXt, which focuses on mining discriminative key regions within images to make the network pay more attention to effective features, to some extent, discarding the interference of background information in pictures. To better adapt to the irregular and multi-scale characteristics of food images, we further propose a multi-stage local feature fusion (MSLF) module to interactively model the long-distance dependencies of output feature maps at various stages, learning complementary effective information from different scale features, thereby enhancing network performance. This work adopts a weakly supervised learning approach, eliminating the need for additional annotations such as bounding boxes, and achieves end-to-end training. The effectiveness and superiority of the methods proposed in this paper have been verified in the field of image recognition on three datasets: ETH Food-101, ChineseFoodNet, and Roushi60.

Acknowledgement: Not applicable.

Funding Statement: The support of this research was by Hubei Provincial Natural Science Foundation (2022CFB449), Science Research Foundation of Education Department of Hubei Province (B2020061), are gratefully acknowledged.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Weizhen Chen; data collection: Gonghui Deng; analysis and interpretation of results: Gonghui Deng, Dunzhi Wu; draft manuscript preparation: Gonghui Deng. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset will be made available online after the article is accepted.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Ege and K. Yanai, "Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions," in *Proc. Themat. Workshops ACM Multimed.*, Mountain View, CA, USA, Oct. 23–27, 2017, pp. 365–375.

[2] L. B. Sorensen, P. Moller, A. Flint, M. Martens, and A. Raben, "Effect of sensory perception of foods on appetite and food intake: A review of studies on humans," *Int. J. Obes. Relat. Metab. Disord.*, vol. 27, no. 10, pp. 1152–1166, Oct. 2003. doi: 10.1038/sj.ijo.0802391.

[3] D. Paully, "A simple method for estimating the food consumption of fish populations from growth data and food conversion experiments," *Fish Bull*, vol. 84, no. 4, pp. 827–840, Oct. 1986.

[4] M. Nestle *et al.*, "Behavioral and social influences on food choice," *Nutr. Rev.*, vol. 56, no. 5, pp. 50–64, May 1998. doi: 10.1111/j.1753-4887.1998.tb01732.x.

[5] W. Q. Min, S. Q. Jiang, L. H. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, Oct. 2019. doi: 10.1145/3329168.

[6] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. 2016 IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 27–30.

[7] S. H. Hou, Y. S. Feng, and Z. L. Wang, "VegFru: A domain-specific dataset for fine-grained visual categorization," in *2017 IEEE/CVF Int. Conf. Comput., Vis.*, Venice, Italy, Oct. 2017, pp. 541–549.

[8] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proc. Int. Workshop Multimed. Assist. Diet. Manag.*, Amsterdam, The Netherlands, Oct. 16, 2016, pp. 41–49. doi: 10.1145/2986035.

[9] Z. Y. Min, J. J. Chen, Y. Cao, C. Forde, C. W. Ngo and T. S. Chua, "Food photo recognition for dietary tracking: System and experiment," in *MultiMed. Mode. 24th Int. Conf.*, Bangkok, Thailand, Feb. 5–7, 2018, pp. 129–141.

[10] H. Yang, S. Kang, C. Park, J. Lee, K. Yu and K. Min, "A hierarchical deep model for food classification from photographs," *KSII T. Internet Inf.*, vol. 14, no. 4, pp. 1704–1720, Apr. 2020.

[11] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 2018 Eur. Conf. on Comput. Vis.*, Munich, Germany, Sep. 8–14, 2018.

[12] C. Wah, S. Branson, P. Weilinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200-2011 Dataset.* Colifornial Institute of Technology, CNS-TR-2010-001, pp. 1–8, 2011.

[13] L. Bossard, M. Guilaumin, and L. V. Gool, "Food-101-mining discriminative components with random forests," in *Comput. Vis.-ECCV 2014. 13th Eur. Conf.*, Zurich, Switzerland, Sep. 6–12, 2014, pp. 446–461.

[14] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Comput. Vis.-ECCV 2014. 13th Eur. Conf.*, Zunrich, Switzerland, Sep. 6–12, 2014, pp. 834–849.

[15] M. Lam, B. Mahasseni, and S. Todorovic, "Fine-grained recognition as HSnet search for informative image parts," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2520–2529.

[16] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, USA, Jun. 15–20, 2019, pp. 5157–5166.

[17] J. Wang, X. Yu, and Y. Gao, "Feature fusion vision transformer for fine-grained visual categorization," 2021. Accessed: Oct. 15, 2023. [Online]. Available: https://arxiv.org/abs/2107.02341

[18] Z. Yang, T. Luo, D. Wang, Z. Q. Hu, J. Gao and L. W. Wang, "Learning to navigate for fine-grained classification," in *Proc. 2018 Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 8–14, 2018, pp. 438–454.

[19] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," in *Proc. AAAI Conf. Artif. Intell.*, Hilton Midtown, NY, USA, Feb. 7–12, 2020, pp. 12289–12296.

[20] T. Zhang, D. Chang, Z. Ma, and J. Guo, "Progressive co-attention network for fine-grained visual classification," in *2021 Int. Conf. VCIP*, Munich, Germany, Dec. 5–8, 2021, pp. 1–5.

[21] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum entropy fine grained classification," in *Adv. Neural inf. Proces. Syst.*, Montreal, QC, Canada, Dec. 2–8, 2018, pp. 637–647.

[22] R. R. Selvaraju, K. Desai, J. Johnson, and N. Naik, "Casting your model: Learning to localize improves self supervised representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 20–25, 2021, pp. 11053–11062.

[23] D. Wu *et al.*, "Align yourself: Self-supervised pre-training for fine-grained recognition via saliency alignment," 2021. Accessed: Aug. 8, 2023. [Online]. Available: https://arxiv.org/abs/2106.15788v3

[24] Y. Shu, A. Van, and L. Liu, "Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems," in *Proc. 2023 IEEE Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 11392–11401.

[25] W. Q. Min, S. Q. Jiang, J. T. Sang, H. Y. Wang, X. D. Liu and L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," *IEEE Trans. Multimed.*, vol. 19, no. 5, pp. 1100–1113, May 2020. doi: 10.1109/TMM.2016.2639382.

[26] G. Ciocca, G. Micali, and P. Napoletano, "State recognition of food images using deep features," *IEEE Access*, vol. 8, pp. 32003–32017, Jan. 2021. doi: 10.1109/ACCESS.2020.2973704.

[27] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, USA, Jun. 27–30, 2016, pp. 770–778.

[28] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. 22nd ACM Int. Conf. Multimed.*, Orlando, Florida, USA, Nov. 2014, pp. 1085–1088.

[29] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *2018 IEEE WACV*, LaKe Tahoe, NV, USA, Mar. 12–15, 2018, pp. 567–576.

[30] H. T. Nguyen, Y. Cao, C. W. Ngo, and W. K. Chan, "FoodMask: Real-time food instance counting, segmentation and recognition," *Pattern Recognit.*, vol. 146, no. 2, pp. 1–11, Feb. 2024. doi: 10.1016/j.patcog.2023.110017.

[31] J. Z. Dai, X. J. Hu, M. Li, Y. Li, and S. D. Du, "The multi-learning for food analyses in computer vision: A survey," *Multimedia Tools Appl.*, vol. 82, no. 17, pp. 25615–25650, Jul. 2023. doi: 10.1007/s11042-023-14373-6.

[32] M. J. Luo, W. Q. Min, Z. L. Wang, J. J. Song, and S. Q. Jiang, "Ingredient prediction via context learning network with class-adaptive asymmetric loss," *IEEE Trans. Image Process*, vol. 32, pp. 5509–5523, Oct. 2023. doi: 10.1109/TIP.2023.3318958.

[33] J. Qiu, F. P. W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," in *Proc. Br. Mach. Vis. Conf. BMVC*, Cardiff, UK, Sep. 9–12, 2022, pp. 1–11.

[34] W. Q. Min, L. H. Liu, Z. D. Luo, and S. Q. Jiang, "Ingredient-guided cascaded multi-attention network for food recognition," in *Proc. 27th Int. Conf. Multimed.*, Nice, France, Oct. 2019, pp. 1331–1339.

[35] S. Q. Jiang, W. Q. Min, L. H. Liu, and Z. D. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Trans. Image Process*, vol. 29, pp. 265–276, Jul. 2020. doi: 10.1109/TIP.2019.2929447.

[36] W. Q. Min *et al.*, "ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proc. 28th ACM Int. Conf. Multimed.*, Seattle, WA, USA, Oct. 12–16, 2020, pp. 393–401.

[37] Y. X. Liu, W. Q. Min, S. Q. Jiang, and Y. Rui, "Food image recognition via multi-scale jigsaw and reconstruction network," (in Chinese), *J. Softw.*, vol. 33, no. 11, pp. 4379–4395, 2022. doi: 10.13328/j.cnki.jos.006325.

[38] P. H. Ma, C. P. Lau, N. Yu, A. Li, and J. P. Sheng, "Application of deep learning for image-based Chinese market food nutrients estimation," *Food Chem.*, vol. 373, no. S2, pp. 1–29, Mar. 2022. doi: 10.1016/j.foodchem.2021.130994.

[39] W. Q. Min *et al.*, "Large scale visual food recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9932–9949, Jun. 2023. doi: 10.1109/TPAMI.2023.3237871.

[40] S. Feng, Y. G. Wang, J. H. Gong, X. Li, and S. X. Li, "A fine-grained recognition technique for identifying Chinese food images," *Heliyon*, vol. 9, no. 11, pp. 1–12, Oct. 2023. doi: 10.1016/j.heliyon.2023.e21565.

[41] F. Shuang, Z. X. Lu, Y. Li, C. Han, X. Gu and S. D. Wei, "Foodnet: Multi-scale and label dependency learning-based multi-task network for food and ingredient recognition," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4485–4501, Mar. 2024. doi: 10.1007/s00521-023-09349-4.

[42] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 2261–2269.

[43] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. 2021 IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 10–17, 2021, pp. 10012–10022.

[44] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, "A convnet for the, 2020s," in *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, USA, Jun. 18–24, 2022, pp. 11976–11986.

[45] X. Chen, Y. Zhu, H. Zhou, L. Diao, and D. Wang, "ChineseFoodNet A large-scale image dataset for chinese food recognition," 2017. Accessed: May 9, 2023. [Online]. Available: https://arxiv.org/abs/1705.02743

[46] G. VijayaKumari, P. Vutkur, and P. Vishwanath, "Food classification using transfer learning technique," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 225–229, Jun. 2022. doi: 10.1016/j.gltp.2022.03.027.

[47] L. Hu, W. Zhang, C. Zhou, G. Lu, and H. Bai, "Automatic diet recording based on deep learning," in *2018 Chinese Autom. Congr. (CAC)*, Xi'an, Shanxi, China, Nov. 2018, pp. 3778–3782.

[48] H. Z. Liang, G. H. Wen, Y. Hu, M. N. Luo, P. Yang and Y. X. Xu, "MVANet: Multi-task guided multi-view attention network for chinese food recognition," *IEEE Trans. Multimed.*, vol. 23, pp. 3551–3561, Jan. 2021. doi: 10.1109/TMM.2020.3028478.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Int. J. Comput. Vis.*, Venice, Italy, Feb. 2017, pp. 618–626.