Journal of Food Engineering 380 (2024) 112134

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: www.elsevier.com/locate/jfoodeng

Fine grained food image recognition based on swin transformer

Zhiyong Xiao [a,b,*], Guang Diao [a], Zhaohong Deng [a,b]

[a] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China
[b] State Key Laboratory of Food Science and Resources, Jiangnan University, Wuxi, 214122, China

ARTICLE INFO

Keywords:
Fine-grained food image recognition
Deep learning
Swin transformer
Food health
Local feature enhancement

ABSTRACT

Fine-grained food image recognition is an important research direction in the field of computer vision and machine learning. However, fine-grained food image recognition faces huge challenges when dealing with foods that vary greatly in shape but belong to the same category or subcategories of that food. To improve this problem, this paper proposes a deep convolution module for obtaining local enhanced feature representation and combines it with the global feature representation obtained from Swin Transformer for deep residual, to obtain a deeper enhanced feature representation. An end-to-end fine-grained food universal classifier was also proposed, which can more accurately extract effective feature information from enhanced feature representations and achieve accurate recognition. Our approach can accurately handle foods with widely different shapes but belonging to the same category and is expected to help people better manage their diet and improve their health. Our models were trained and verified on the public fine-grained food datasets Foodx-251 and UEC Food-256 respectively, where the accuracy of the method on the validation set is 81.07% and 82.77% respectively. Compared with other state-of-the-art self-supervised methods, the method proposed in this paper exhibits higher accuracy in fine-grained food image recognition tasks.

1. Introduction

Food not only has a profound impact on human health and nutrition but also plays an important role in defining people's identity, social status, and culture (Khanna, 2009). Therefore, food-related research has become a persistent hot topic (Sajadmanesh et al., 2017). Researchers from different research fields have conducted food-related research from different angles, including food choice, food perception, food consumption, food safety, and food culture, etc. Min et al. (2019) systematically proposed a food computing framework, pointing out that food computing includes multiple tasks such as food perception, identification, retrieval, and recommendation, and serves multiple fields such as medicine, biology, agriculture, food industry, nutrition, and health. Among them, food image recognition is one of its basic and core tasks. From the field of computer vision, food image recognition is divided into coarse-grained image recognition and fine-grained image recognition (Qiu et al., 2022). Fine-grained image recognition is more challenging. Therefore, this paper focuses on solving the difficult problem of fine-grained food image recognition. The difficulties in fine-grained image recognition mainly include: (1) The differences between classes are small. (2) The intra-category differences are large (Min et al., 2019). It is clear from Fig. 1 that the difficulties in recognizing fine-grained food images discussed earlier are evident. For example, the five images in Caprese Salad and Apple Pie are in the same category but the differences between them are huge, which makes it extremely difficult to accurately grasp the category of each food. To solve this problem, Classic image recognition handles all object categories similarly and achieves impressive results comparable to human performance (Dodge and Karam, 2017).

With the emergence of deep learning algorithms and the rise of large models, food image recognition has gradually become a hot research topic (Zhao et al., 2017). With the advent of the Internet age and the digital age, image samples in the data set can be crawled through the Internet. Therefore, the recently released food data sets contain a large number of images, resulting in the emergence of many powerful food recognition algorithms (Zheng et al., 2018). Food image recognition can be roughly divided into coarse-grained food image recognition and fine-grained food image recognition. Research on coarse-grained food image recognition has produced good results, for example, recently based on the improved algorithm of Vision Transformer, the accuracy on the ETHZ Food-101 (Bossard et al., 2014) dataset and Vireo Food-172 (Chen and Ngo, 2016) dataset reached 95.2% and 94.3% of good entries respectively (Gao et al., 2024). However, the research on fine-grained food image recognition is not in-depth enough, so this paper focuses on solving the problem of fine-grained food image recognition. Most works on fine-grained image analysis are divided

* Corresponding author.
E-mail address: zhiyong.xiao@jiangnan.edu.cn (Z. Xiao).

https://doi.org/10.1016/j.jfoodeng.2024.112134
Received 10 January 2024; Received in revised form 12 April 2024; Accepted 11 May 2024
Available online 16 May 2024

Fig. 1. Example of fine-grained food classes from FoodX-251 dataset.

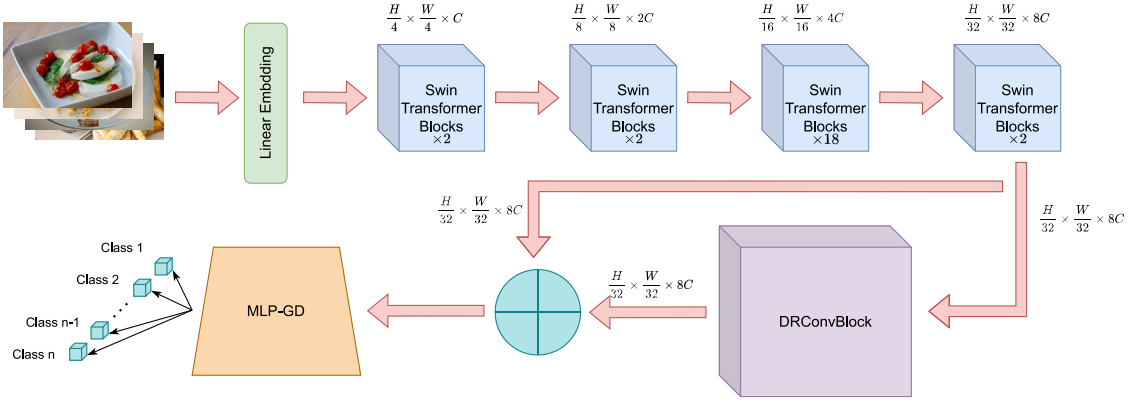

Fig. 2. The overall architecture of Swin-DR. Swin-DR consists of three parts; The first part is a backbone network composed of multiple Swin Transformer Blocks, the second part is the local feature enhancer DRConvBlock, and the last part is the fine-grained food universal classifier MLP-GD.

into fine-grained image recognition (Yu et al., 2019) and fine-grained image retrieval (Xu et al., 2018). Fine-grained image recognition focuses on learning detailed features from images to identify the correct categories. In previous research, a variety of methods have appeared to solve problems in fine-grained image recognition, and many good results have been achieved. The main methods include CNN (convolutional neural network) (Yanai and Kawano, 2015), Vision Transformer, dual-teacher model fusing CNN and Transformer (Xiao et al., 2022), 3D lightweight network using multiscale convolution attention and vision Transformer (Xiao et al., 2024), etc. Ródenas et al. (2022). recently proposed method learning multi-subset of classes for fine-grained food recognition, which achieved the best performance in multiple data sets, including a SoTA score of 79.90% on the fine-grained food dataset foodx-251. The key to fine-grained food image recognition lies in creating automatic methods to identify areas with differences in the image (Hu et al., 2018). However, fine-grained image recognition requires the establishment of a more detailed method that focuses on classifying subsets of categories within the entire class set (Akata et al., 2015). It solves the image recognition problem by focusing on distinguishing difficult-to-distinguish object categories (Lin et al., 2019). However, the previous method mainly had two shortcomings: on the one hand, the network focuses on the most obvious differences between classes and ignores more subtle inter-class variations, and on the other hand, all classes are processed in the same global or local way, no deeper differential information was obtained. Because incorrect recognition often occurs in most similar categories, we cannot simply distinguish all food categories (He and Peng, 2017).

To address the shortcomings in previous fine-grained food image recognition methods, we learned from the experience of previous work and combined swin Transformer (Liu et al., 2021), convolutional neural network (Tan and Le, 2019), and multilayer perceptron (Taud and Mas, 2018; Ji et al., 2023) to make the extracted features more obvious and make it easier for the model to learn the similarities of different subcategories. This paper proposes a more accurate fine-grained food image recognition method by combining the Swin Transformer backbone network with deep residual networks(Swin-DR). Swin-DR can more accurately grasp the feature relationship between sub-categories and parent categories in fine-grained levels, and extract effective feature information in more detail, thereby completing more accurate predictions of food categories. Swin-DR uses Swin Transformer (Liu et al., 2021) as the backbone network, proposes that depthwise separable residual convolutional blocks(DRConvBlock) can further enhance local features, deletes the classifier in the original Swin Transformer, and A multi-layer perceptron based on global average pooling and

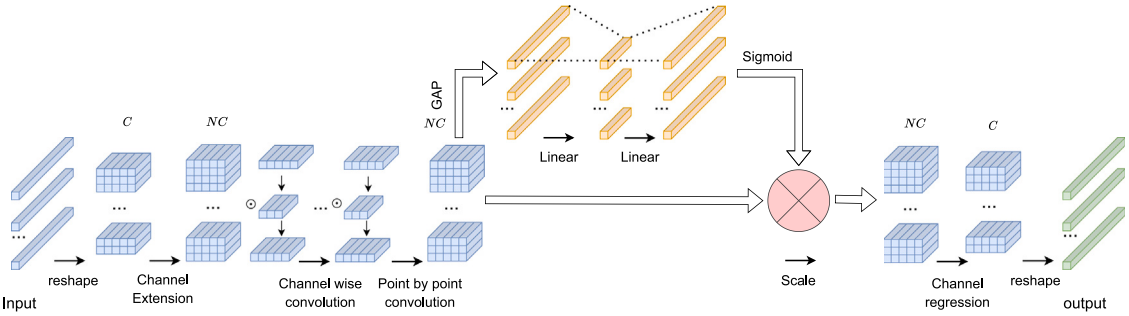

Fig. 3. Structure diagram of DRConvBlock. Reshape refers to spatial dimension transformation. Channel expansion refers to increasing the number of channels by N times through 1×1 convolution. Channel-wise convolution refers to dividing the feature map into NC (number of channels) groups and using 3×3 convolution processing respectively. Point-by-point convolution refers to concatenating the feature maps obtained from grouped convolutions and then processing them through 1×1 convolution. GAP refers to global average pooling. Linear values are fully connected and activated using the GELU activation function. Scale refers to multiplying a vector processed by the sigmoid function with a previous feature map. Channel regression refers to using 1×1 convolution to return the number of channels to C. Note that each convolutional block operation needs to be batch normalized and activated through the ReLU activation function.

dropout(MLP-GD) has been redesigned as an end-to-end fine-grained food image universal classifier. DRConvBlock can further mine the local feature representation from the global feature representation and enhance the representation, thereby extracting enhanced local feature information. MLP-GD can better extract effective feature information from the enhanced feature representation, completing more accurate food category predictions. The experiment showed that Swin-DR successfully improved the fuzzy class and outperformed state-of-the-art methods on both datasets.

In the era of globalization, the styles of food are undergoing tremendous changes, resulting in many different forms of the same food. People are not easily able to distinguish their differences, which leads to poor management of daily diet and may bring certain health problems. How to accurately grasp the relationship and differences between parent and child classes of fine-grained food has become the core of solving this problem. Swin-DR can mine and enhance local feature representations and deeply combines global feature representations with enhanced local feature representations. Finally, an end-to-end food universal classifier is used for feature extraction and recognition. Therefore, Swin-DR can better solve the difficulties in fine-grained food image recognition, and accurately grasp the relationship and differences between parent and child classes. The goal of Swin-DR is to help people better distinguish between similar foods that have significant macro differences, manage their daily diet more reasonably, and improve physical health issues.

2. Methods

2.1. Swin transformer

Liu et al. (2021) proposed Swin Transformer, which is a hierarchical Transformer whose feature representation is calculated by shifting windows. The shift windowing scheme brings higher efficiency by limiting self-attention computation to non-overlapping local windows while also allowing cross-window connections. This hierarchical architecture provides the flexibility to model at various scales with linear computational complexity relative to image size. These features of the Swin Transformer make it compatible with a wide range of vision tasks, including image recognition and dense prediction tasks such as object detection and semantic segmentation. With the rise of the Swin Transformer in the field of computer vision, more and more variants of the Swin Transformer are appearing. For example, Chu et al. (2021) proposed Twins revisited the design of spatial attention, and demonstrated that a well-designed but simple spatial attention mechanism performs well against state-of-the-art schemes. Dong et al. (2022) proposed Cswin proposed a cross-shaped window self-attention mechanism, and changed different layers of the Transformer network that changes the stripe width to achieve powerful modeling capabilities while limiting computational costs.

2.2. Transfer learning

The purpose of transfer learning is to improve the target learner's learning performance in the target domain by transferring knowledge in different but related source domains (Zhuang et al., 2020). Transfer learning can reduce the dependence of building a target learner on a large amount of target domain data. Transfer learning (Cui et al., 2018) has become a very hot research direction in the field of deep learning because of its wide application prospects. In the field of transfer learning, if the target learner is negatively affected by the transferred knowledge, this phenomenon is also called negative transfer (Yang et al., 2018). Whether negative transfer occurs depends on several factors, such as the correlation between the source and target domains, and the learner's ability to discover transferable and beneficial parts of knowledge across domains (Wang et al., 2019). Therefore, when choosing transfer learning, you need to pay more attention to whether the original research field is relevant to the current research field, to avoid negative transfer as much as possible. In addition to this traditional transfer learning, there are also some deep ones, such as reinforcement transfer learning (Taylor and Stone, 2009), lifelong transfer learning (Parisi et al., 2019) and online transfer learning (Zhao et al., 2014), etc. The work of this paper is based on transfer learning, as it can reduce the cost of the experimental process, allowing us to conduct further in-depth research.

2.3. Overall architecture of the approach

The overall architecture of Swin-DR is shown in Fig. 2. Swin-DR consists of three parts: the first part is the backbone network, the second part is the local feature enhancer, and the last part is the fine-grained food universal classifier. Among them, the backbone network is composed of multiple Swin Transformer blocks, the local feature enhancer is DRConvBlock, and the fine-grained food universal classifier is MLP-GD. The introduction of DRConvBlock and MLP-GD will be expanded on in the following text. The overall execution process of the method is as follows: firstly, input the RGB fine-grained food image of 3 channels into the backbone network, and obtain the feature information of the global feature representation through processing by the backbone network. Then, the feature information is passed into DRConvBlock for local feature enhancement, thereby obtaining the feature information that enhances the local feature representation. Then, the enhanced local feature information is combined with the global feature information to obtain the enhanced feature information. Finally, the enhanced feature information is fed into MLP-GD for feature extraction and category prediction. Swin-DR combines global feature information with enhanced local feature information, thus possessing the dual advantages of feature representation and enhanced local feature representation. It can

better grasp the relationship between subclasses and parents in fine-grained food and more accurately grasp category information, thus better solving the difficulties in fine-grained food recognition and achieving more accurate category prediction.

2.4. DRConvBlock

This article has already introduced the main architecture of Swin DR in the previous section and will introduce the local feature enhancer DRConvBlock in this section. Swin Transformer performs well on coarse-grained image recognition, but its grasp of local features at fine-grained levels is not detailed enough. Therefore, to enable the network to better obtain local feature information, Swin-DR added a local feature enhancer DRConvBlock in front of the classifier. Compared with the Transformer's original image processing method, after passing DRConvBlock, Swin-DR can get a more detailed and deeper enhanced local feature representation. As shown in Fig. 3, DRConvBlock consists of multiple parts, including channel expansion convolution, depth-separable convolution, squeeze-and-excitation module, and channel regression convolution. The channel expansion convolution is composed of 1×1 convolution and is used to amplify the feature information channel by N times. Depthwise separable convolution is composed of 3×3 channel-wise convolution and 1×1 point-wise convolution. Channel regression convolution consists of 1×1 convolution and is used to restore N times the channels to the original channels. Squeeze-and-excitation consists of a global average pooling layer, two fully connected layers, and a Sigmoid layer. In addition to the structure, batch normalization and ReLU function activation are performed on the feature map after each convolutional block, and the activation function used in the squeeze and excitation stage is GELU.

Fig. 4 shows the structure of the Swin Transformer block. The original architectural formula of the Swin Transformer Block part is as follows:

$$\begin{aligned} \hat{\mathbf{z}}^{l} &= W - MSA(LN(Z^{l-1})) + Z^{l-1} \\ \mathbf{z}^{l} &= MLP(LN(\hat{Z}^{l})) + \hat{Z}^{l} \\ \hat{\mathbf{z}}^{l+1} &= SW - MSA(LN(Z^{l})) + Z^{l} \\ \mathbf{z}^{l+1} &= MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \end{aligned} \tag{1}$$

Among them, z^{l-1} is the result of the $L-1$ Swin Transformer Block, Z^{l} is the result of passing the Lth Swin Transformer block, W-MSA() is the window attention mechanism, and SW-MSA() is the moving window attention mechanism, LN() is layer normalization, MLP() is multi-layer perceptron. After passing through multiple Swin Transformer blocks, a relatively complete global feature representation can be obtained. However, the feature information obtained solely through the self-attention mechanism is not accurate enough for fine-grained food image recognition. Therefore, this paper proposes the local feature enhancer DRConvBlock to obtain more detailed feature information. The DRConvBlock formula is as follows:

$$\begin{aligned} \mathbf{Z} &= V_{nc} * X = \sum_{i=1}^{nC} V_{nc}^{i} * X^{k} \\ \mathbb{Z} &= \sigma(\delta(Z)) \end{aligned} \tag{2}$$

$\mathbf{X} = [x_1, x_2, \ldots, x_c]$ is the global feature information obtained from the Swin Transformer backbone network. $\mathbf{Z} = [z_1, z_2, \ldots, z_{nc}]$ is the result of $\mathbf{X}$ enlarging the channel n times through 1×1 convolution. $\mathbb{Z}$ is the result of feature map $\mathbf{Z}$ being batch normalized and activated by the RelU activation function. where δ is batch normalization and σ is the ReLu activation function.

$$\begin{aligned} \mathbf{Y_i} &= V_{nc} * \mathbb{Z} = \sum_{i=1}^{C} V_{nc}^{i} * \mathbb{Z}^{k} \\ \mathbb{Y} &= Concat(y_1 \odot w_1, y_2 \odot w_2, \ldots, y_{nc} \odot w_{nc}) \end{aligned} \tag{3}$$

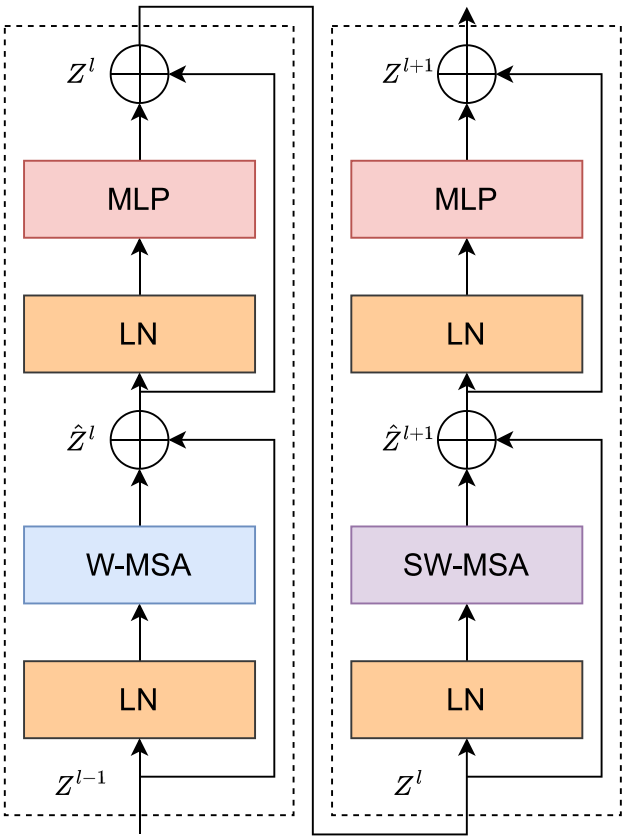

Fig. 4. Detailed structure of the Swin Transformer block. W-MSA denotes window-based self-attention, and SW-MSA represents shifted window-based self-attention. $\hat{Z}^{i}$ and Z^{i} denote the output feature map of the ith block after (S)W-MSA and MLP, respectively.

$\mathbf{Y_i}$ is to divide the feature map $\mathbb{Z}$ into nc groups and perform group convolution separately. The convolution kernel size of the group convolution is 3×3. Weighted concatenate the feature maps y_i and w_i of group convolution, and then concatenate all the results together to obtain $\mathbb{Y}$.

$$\begin{aligned} \mathbf{U_i} &= \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,W} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,W} \\ \vdots & \vdots & \ddots & \vdots \\ u_{H,1} & u_{H,2} & \cdots & u_{H,W} \end{bmatrix} \\ u_{i,j} &= y_{i,j} \odot W \end{aligned} \tag{4}$$

U_i is the result of the weighted connection between the feature map $y_i = [yi_1, yi_2, \ldots, yi_{nc}]$ obtained by the ith group convolution and the ith weight $W_i = [wi_1, wi_2, \ldots, wi_{nc}]$. The final $U = [u_1, u_2, \ldots, u_{nc}]$ is the feature information composed of nc U_i feature maps.

$$\begin{aligned} \mathbf{G} &= F_{sq}(U) = \frac{1}{HxW} \sum_{i=1}^{H} \sum_{j=1}^{W} U(i,j) \\ \mathbf{S} &= F_1(G, W) = \sigma(W1 \bullet G + b1) \\ \mathbb{S} &= F_2(S, W) = \sigma(W2 \bullet S + b2) \\ \mathbb{U} &= Sigmoid(\mathbb{S}) \end{aligned} \tag{5}$$

$G = [g_1, g_2, \ldots, g_{nc}]$ is the feature vector obtained by performing GAP (global average pooling) on the feature map U. The purpose is to reduce the dimension and retain feature information. $S = [s_1, s_2, \ldots, s_{nc}]$ is a weighted connection of the feature vector G and the weight $W1 = [w1_1, w1_2, \ldots, w1_{nc}]$, plus the bias b1 and activated through the GELU activation function. Where σ is the GELU activation function. $\mathbb{U}$ is the

feature vector after $\mathbb{S}$ is processed by the Sigmoid function.

$$\begin{aligned}\mathbf{P} &= U \odot \mathbb{U} \\ \mathbb{X} &= \delta(\sigma(V_c * P)) = \delta(\sigma(\sum_{i=1}^{c} V_c^i * P^k)) \\ \mathbb{Y} &= X \oplus \mathbb{X}\end{aligned} \tag{6}$$

$P = [p_1, p_2, \ldots, p_c]$ is a brand new feature map obtained by matrix multiplication of feature vector **U** and feature map $\mathbb{U}$. $X = [x_1, x_2, \ldots, x_c]$ is the result of channel regression of feature **P** through 1×1 convolution (channel number returned from NC to C) and undergoing batch normalization, activated by ReLU activation function. Where σ is batch normalization and δ is the ReLU activation function. The final enhanced feature information $\mathbb{Y} = [y_1, y_2, \ldots, y_c]$ of the fusion of global features and local features is the residual connection of the global feature map X and the enhanced local feature map $\mathbb{X}$.

This paper has verified the effectiveness and practicality of local feature enhancer theory in the experimental section below and demonstrated its superiority in combination with the Swin Transformer with the significant increase of various evaluation indicators.

2.5. End-to-end fine-grained universal classifier

In the native Swin Transformer, the classifier is only a single FC (Fully Connected Layer), which is not enough to accurately extract feature information and complete good category prediction. Therefore, this paper proposes a novel fine-grained food image universal classifier MLP-GD. As shown in Fig. 5, Swin-DR replaces a single FC with MLP (Multi-Layer Perception) and performs GAP (Global Average Pooling) before transferring feature information to MLP, reducing the dimensionality of the feature information and removing excess information. Swin-DR also used a dropout with a deletion rate of 0.1 between each FC to prevent overfitting due to excessive computational complexity, and the number of neurons in each FC decreased sequentially, making the classifier more stable and predicting more accurately. The original Swin Transformer calculation formula is as follows:

$$\mathbf{Y} = f(W \bullet X + b) \tag{7}$$

In the original Swin Transformer, the output result $Y = [y_1, y_2, \ldots, y_n]$ is the result of the input feature information $X = [x_1, x_2, \ldots, x_c]$ through a weighted connection with the weight $W = [w_1, w_2, \ldots, w_c]$ and then adding the bias b. where n is the number of categories of fine-grained food. MLP-GD improvements based on it are as follows:

$$\begin{aligned}\mathbf{G} &= F_{sq}(X) = \frac{1}{HxW} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j) \\ \mathbf{Y} &= F_1(G, W) = \delta(\sigma(W1 \bullet G + b1)) \\ \mathbb{Y} &= F_2(Y, W) = \delta(\sigma(W2 \bullet Y + b2))\end{aligned} \tag{8}$$

G is the result of the global feature information $X = [x_1, x_2, \ldots, x_c]$ obtained from the Swin Transformer backbone network after GAP dimensionality reduction. δ is the dropout function, and σ is the GELU activation function. $Y = [y_1, y_2, \ldots, y_h]$ is the weighted connection between G and the weight $W1 = [w1_1, w1_2, \ldots, w1_c]$ based on dropout, with bias $b1$ added, and then activated by the GELU activation function. In the same way, The final output result $\mathbb{Y}$ is the result of a more detailed and accurate feature extraction based on Y. Compared to a single FC, MLP-GD can extract more detailed feature information, thus being able to predict the category of fine-grained food more accurately.

This paper validated the effectiveness and practicality of the end-to-end fine-grained food image classifier in the experimental section below and demonstrated its feasibility in combination with the Swin Transformer as various evaluation metrics increased.

3. Experiments

3.1. Datasets

This paper uses two common fine-grained food image datasets to validate the effectiveness of Swin-DR. All datasets in the experiment were divided using official standard methods.

- FoodX-251 (Kaur et al., 2019) contains 251 fine-grained classes with 118k training, 12k validation and 28k test images. Human-verified labels are made available for the training and test images. The classes are fine-grained and visually similar, for example, different types of cakes, sandwiches, puddings, soups, and pasta.
- UEC FOOD 256 (Kawano and Yanai, 2014) contains 256-kind fine-grained food photos. Each food photo has a bounding box indicating the location of the food item in the photo. Most of the food categories in this dataset are popular foods in Japan and other countries. Therefore, some categories might not be familiar to other people than Japanese.

3.2. Evaluation metrics

To verify the performance of the proposed method in this article, three commonly used image recognition evaluation metrics were adopted, namely Top-1 accuracy(Acc.), F1 score(F1.), and precision (pre.). The F1 score is the harmonic mean of the precision rate and the recall rate. The precision rate is how many of the results predicted to be positive samples are correctly classified. The accuracy is the proportion of the number of correctly classified samples to the total number of samples. The evaluation metrics are defined as follows.

$$\mathbf{Acc} = \frac{\sum_{i=0}^{N-1} \left(f(x_i) == y_i \right)}{N} \tag{9}$$

$$\mathbf{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\mathbf{Pre} = \frac{TP}{TP + FP} \tag{11}$$

$$\mathbf{F1} = \frac{2 \cdot pre \cdot Recall}{pre + Recall} \tag{12}$$

In the definitions of the major evaluation indicators mentioned above, TP (True Positive) is defined as the number of positive samples predicted by the system and positive samples. FP (False Positive) is defined as the number of false samples predicted by the system but true samples. TN (True Negative) is defined as the number of false samples predicted by the system and false samples. FN (False Negative) is defined as the number of false samples that the system predicts to be true samples.

Comparison method: This paper adopts various advanced self-supervised methods to compare with Swin-DR in experiments, and the experimental environment and equipment are all under the same conditions, ensuring the fairness of the experiment. The ablation experiment is a separate and splicing experiment conducted on each component of Swin-DR under the same conditions.

3.3. Implementation details

All experiments in this paper were conducted on Nvidia GPU 2080TI. Swin-DR uses the weights trained by Swin Transformer on the ImageNet-22k dataset as Swin-DR pre-training weights. This paper divided the datasets Foodx-251 and UEC Food-256 into training and validation sets according to official standards. The other parameter settings in the experiment are shown in Table 1. In addition to these parameters, this paper has also combined multiple data augmentation methods to make the execution of each method more accurate and effective. Among them are RandomResizedCrop, RandomHorizontalFlip, TrivialAugmentWide, RandomErasing, AutoAugment, etc. The data augmentation methods summarized in this paper are more suitable for general data augmentation methods for fine-grained food image recognition.

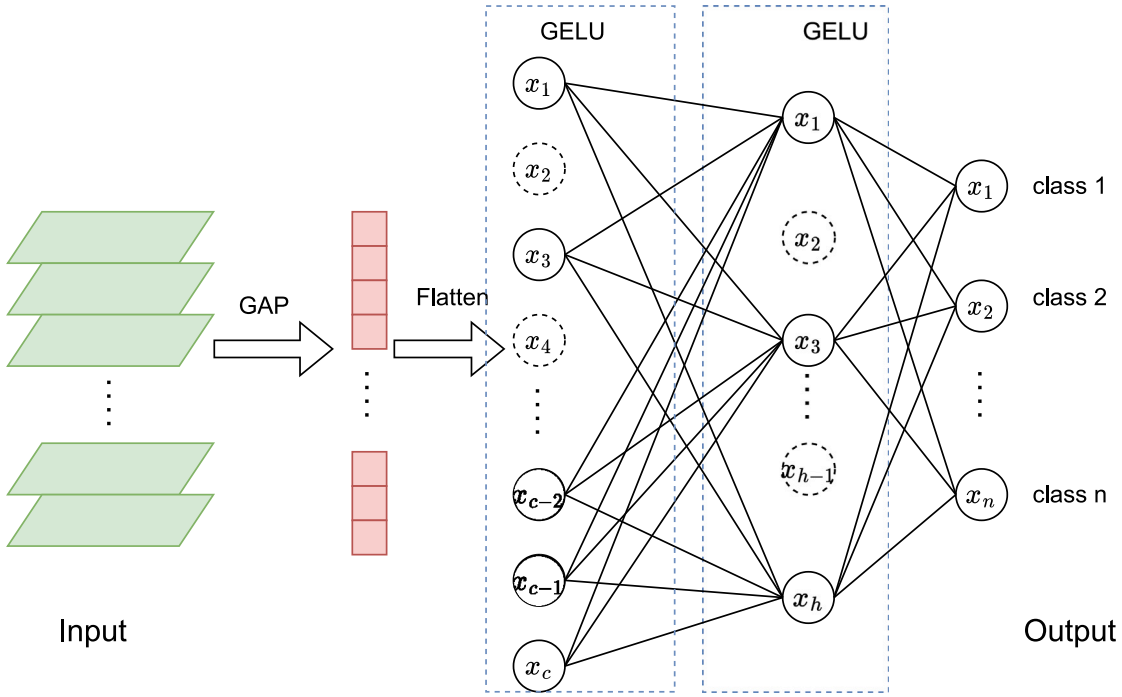

Fig. 5. Detailed structure diagram of MLP-GD. Where GAP is the global average pooling. GELU is the activation function. flatten is dimension expansion. The dotted circles represent neuron nodes that do not participate in calculations. Each connection line will have a weight value, so each connection line is a weighted connection.

Table 1
Detailed settings for pre-training, fine-tuning, and transfer learning.

Task	Model	Datasets	Input	Epochs	Batch size	Optimizer	LR	LR decay	Weight decay	Warmup epochs
Pretrain	Swin Transformer	ImageNet-22K	224	90	4096	AdamW	1.00E−03	Cosine	0.01	5
Finetune	Swin-DR	Foodx-251	224	50	8	AdamW	5.00E−04	Cosine	0.05	0
Finetune	Swin-DR	UEC Food-256	224	50	8	AdamW	5.00E−04	Cosine	0.05	0

Table 2
Swin-DR compares results with other models on the Foodx-251 and UEC Food-256 datasets. The short horizontal line represents the indicators that were not evaluated in the cited article.

Methods	Epochs	Resolution	FoodX251			UECFood-256		
			Acc. (%)	Pre. (%)	F1. (%)	Acc. (%)	Pre. (%)	F1. (%)
ResNet-50	50	224 × 224	72.13	72.05	72.93	75.54	75.44	76.06
ResNet-101	50	224 × 224	73.11	73.1	74.00	75.65	75.57	76.3
TResNet-L	50	224 × 224	74.84	74.82	75.61	76.18	76.01	76.74
TResNet-XL	50	224 × 224	74.59	74.51	75.27	76.35	76.17	76.77
EfficiectNet-b7	50	224 × 224	73.54	73.44	74.15	74.84	74.52	75.05
ConvNext-B	50	224 × 224	77.84	77.75	78.35	78.79	78.6	79.11
ConvNext-L	50	224 × 224	78.02	77.92	78.52	79.44	79.25	79.74
ViT-B	50	224 × 224	77.46	77.35	78.02	78.90	78.91	79.66
ViT-L	50	224 × 224	79.51	79.36	79.88	80.67	80.61	81.86
SwinT-B	50	224 × 224	78.57	78.55	79.17	80.81	80.74	81.61
SwinT-L	50	224 × 224	79.58	79.53	80.11	81.77	81.72	82.52
Swinv2-B	50	192 × 192	77.36	77.33	77.98	80.30	80.27	80.92
Swinv2-L	50	192 × 192	78.31	78.30	78.89	80.22	80.23	80.86
DeiT-S	50	224 × 224	72.62	72.63	73.39	75.28	75.28	75.98
DeiT-B	50	224 × 224	75.91	75.91	76.65	77.41	77.38	78.07
DeiTv2-B	50	224 × 224	75.52	75.41	76.03	78.05	78.06	78.70
DeiTv2-L	50	224 × 224	77.54	77.43	78.14	79.07	78.96	79.58
Twins-B	50	224 × 224	75.82	75.77	76.44	77.61	77.53	78.11
Twins-L	50	224 × 224	76.04	75.93	76.6	78.13	78.00	78.65
Cait-S	50	224 × 224	76.41	76.40	77.08	77.87	77.76	78.46
CSWin-L (Ródenas et al., 2022)	50	224 × 224	79.90	–	–	–	–	–
VOLO-D5 (Ródenas et al., 2022)	50	224 × 224	79.51	–	–	–	–	–
Inception V3 (Hassannejad et al., 2016)	50	224 × 224	–	–	–	76.17	–	–
WRN (Martinel et al., 2018)	50	224 × 224	–	–	–	79.76	–	–
Swin-DR	**50**	**224 × 224**	**81.07**	**80.98**	**81.48**	**82.77**	**82.41**	**83.12**

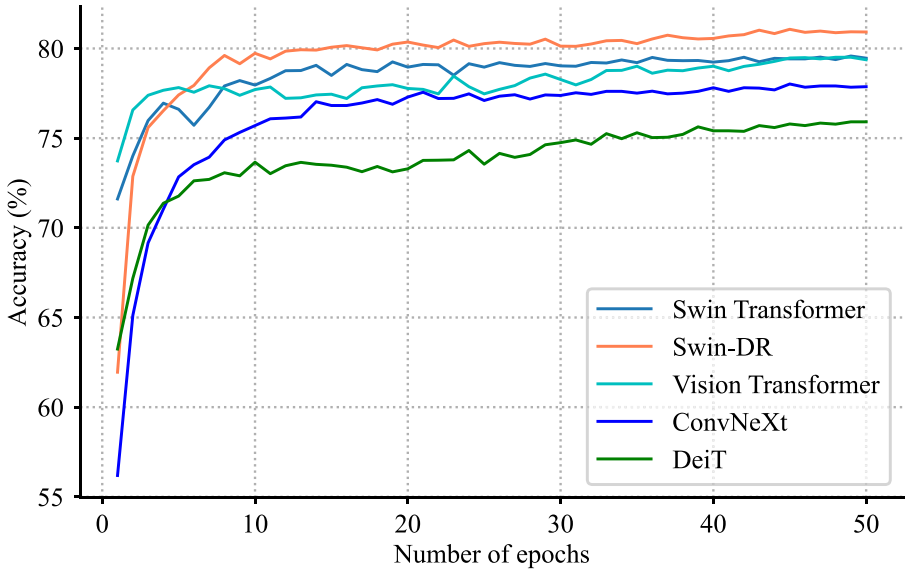

Fig. 6. Comparison results between Swin-DR and other advanced self-supervised methods.

4. Results

4.1. Performance comparison to the SoTA

This paper conducted a comparative experiment between Swin-DR and all the state-of-the-art self-supervised recognition methods on fine-grained food datasets FoodX-251 and UEC Food-256, using three conventional image recognition evaluation metrics as evaluation criteria. The three evaluation indicators are Top-1 accuracy, F1 score, and precision. For the fairness and accuracy of the experimental results, all evaluation results were measured on a single Nvidia 2080TI GPU, with mixed precision acceleration and batch size set to 8 during the experiment. These advanced self-supervised recognition methods are (1) Data-efficient image Transformers (DeiT, DeiTv2) (Touvron et al., 2021a, 2022); (2) Swin Transformer (SwinT, Swinv2) (Liu et al., 2021, 2022a); (3) ConvNeXt (Liu et al., 2022b); (4) Going deeper with Image Transformer (CaiT) (Touvron et al., 2021b); (5) Revisiting the Design of Spatial Attention in Vision Transformers(Twins) (Chu et al., 2021); (6) High Performance GPU-Dedicated Architecture (TResNet) (Ridnik et al., 2021); (7) Vision Transformer (ViT) (Dosovitskiy et al., 2020); (8) Rethinking Model Scaling for Convolutional Neural Networks (EfficientNet) (Tan and Le, 2019) (9) Learning Multi-Subset of Classes for Fine-Grained Food Recognition (CSwin, VOLO) (Ródenas et al., 2022).

From the Table 2, it can be seen that the convolutional neural network has the highest accuracy of 78.02% and 79.44% on two datasets, Vision Transformer and its variants have the highest accuracy of 79.58% and 81.77%, CSwin on multiple subsets can reach the highest accuracy of 79.90%, and Swin-DR achieved the best results of 81.07% and 82.77% on both datasets. Fig. 6 shows the trend of accuracy changes between Swin-DR and some state-of-the-art recognition methods during training and validation on the fine-grained food dataset FoodX-251. Swin-DR shows a significant increase in accuracy during training and validation compared to other methods, and the improvement is also relatively stable. Compared to convolutional neural networks, Swin-DR can not only obtain local feature representations but also mine global feature representations, thereby more accurately extracting feature information. Compared to Vision Transformer and its variants, Swin-DR can mine more detailed local feature representations and enhance them, thereby extracting more accurate feature information and more accurately grasping the subtle differences and connections in fine-grained food images. In addition, Swin-DR is the result of the combination of Swin Transformer and CNN deep residuals. Compared to the ordinary combination method, Swin-DR can obtain enhanced feature information, making it stronger than other combination methods of Transformer and CNN in fine-grained food recognition tasks.

The fine-grained food image recognition task involves the relationship between multiple subclasses and parents, and there are significant macro differences between subclasses and parents. Other recognition methods cannot accurately mine the connections and differences between subclasses and parent classes of fine-grained food. However, Swin-DR can extract local feature representations from global feature representations and enhance them, and effectively combine global feature representations with enhanced local feature representations to obtain enhanced feature representations. Therefore, Swin-DR can more accurately extract effective feature information and grasp the connections and differences between parent and child classes more accurately, ultimately solving the problems in fine-grained food image recognition and making more accurate recognition. These experimental results show that Swin-DR not only has high accuracy and good fitting, but also has strong generalization ability to unknown samples, and can accurately infer data patterns and make correct predictions. Therefore, the method can more accurately handle food of different sizes and shapes but of the same type. As shown in Fig. 1, the images of various apple pies in the Foodx-251 data set have huge differences in appearance, some are strip-shaped, and some are pie-shaped, but they are indeed in the same category. Swin-DR can better help everyone distinguish the same type of food with widely different shapes, allowing everyone to better manage their diet and improve their health.

4.2. Ablation investigations

To verify the effectiveness and rationality of each component of Swin-DR, this paper conducted a series of ablation experiments on datasets FoodX-251 and UEC Food-256. From the experimental results in Table 3, it can be seen that the accuracy of the native Swin Transformer without adding any modules on the two datasets is 79.58% and 81.77%, respectively. After adding the local feature enhancer DRConvBlock to the benchmark model, it can be seen that there is a significant improvement, with accuracy rates of 80.74% and 82.28%, respectively, indicating that the module can significantly enhance feature information. After adding the fine-grained food image universal classifier MLP-GD to the benchmark model, there was also a significant improvement, with accuracy rates of 80.19% and 82.07%, respectively,

Table 3
Ablation experiment of Swin-DR (Ours) on the Foodx-251 and UEC Food-256 datasets.

Methods			Foodx-251			UEC Food-256		
Swin transformer	DRConvBlock	MLP-GD	Acc. (%)	Pre. (%)	F1. (%)	Acc. (%)	Pre. (%)	F1. (%)
✓	✗	✗	79.58	79.53	80.11	81.77	81.72	82.52
✓	✓	✗	80.74	80.69	81.29	82.28	82.08	82.76
✓	✗	✓	80.19	80.16	80.93	82.07	81.95	82.93
✓	✓	✓	81.07	80.98	81.48	82.77	82.41	83.12

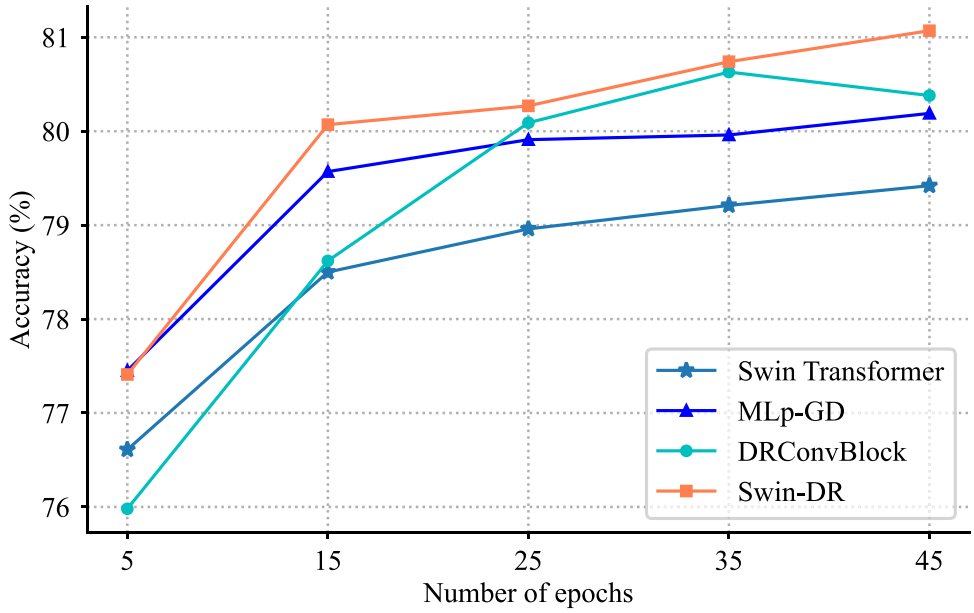

Fig. 7. Figure of ablation experiment results on Foodx-251.

indicating that MLP-GD can better extract feature information. Finally, after adding the enhancer DRConvBlock and the universal classifier MLP-GD to the benchmark Swin Transformer, the accuracy significantly increased and reached the best levels of 81.07% and 82.77%, respectively.

Through Fig. 7, we can visually observe the changing trends of each component during the training and validation process, as well as the differences and improvements between each component and the native benchmark Swin Transformer during the training process. This can more accurately prove the effectiveness of each component in Swin-DR and the rationality of the combination of components. The results of this series of ablation experiments can further prove that each component of Swin-DR is not only theoretically feasible but also achievable, and has achieved very good results. Ultimately, it can demonstrate the advantages of Swin-DR in processing fine-grained food image recognition tasks.

5. Conclusion and future work

Fine-grained recognition is a challenging computer vision problem, where the inter-class differences between similar classes are very small. To solve the difficulties in fine-grained food image recognition, this paper proposes a new fine-grained food image recognition method Swin-DR that combines deep residual convolution and Swin Transformer and proposes a new local feature enhancer DRConvBlock for mining deeper local feature information from global feature representations. Finally, this paper also proposes an end-to-end universal fine-grained food image classifier MLP-GD that can more accurately extract effective feature information from enhanced feature representations. In the result section of this paper, a series of comparative experiments were conducted to demonstrate the superiority of Swin-DR over other state-of-the-art self-supervised recognition methods, and a series of ablation experiments were also performed to demonstrate the effectiveness of each component of Swin-DR and the rationality of its combination.

Under the influence of globalization and modern lifestyles, people's eating habits have undergone tremendous changes, and health problems such as malnutrition and obesity are becoming increasingly serious. It is not difficult to see from the results of comparative experiments and ablation experiments that Swin-DR can better solve the problems and challenges in fine-grained food image recognition. Swin-DR not only fits well but also has stronger generalization and inference abilities to make correct predictions in unknown samples. Therefore, Swin-DR is expected to help people better manage their diet and improve their health through the recognition of fine-grained food images. Unfortunately, Swin-DR has only been tested on fine-grained food image datasets and has not been further explored on a wider range of fine-grained image datasets. Therefore, the future work is to extend this method to a wider range of fields and practical applications in computer vision.

CRediT authorship contribution statement

Zhiyong Xiao: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology. **Guang Diao:** Software, Investigation, Formal analysis, Data curation, Conceptualization. **Zhaohong Deng:** Writing – review & editing, Visualization, Validation, Software, Investigation, Data curation.

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.

Data availability

Data will be made available on request.

Acknowledgments

The authors are grateful to the reviewers for their valuable comments, which have greatly improved the paper. This work was supported by the Natural Science Foundation of Jiangsu Province (China) under Grant BK20190079 and the National Natural Science Foundation of China under Grant 62176105.

References

Akata, Zeynep, Reed, Scott, Walter, Daniel, Lee, Honglak, Schiele, Bernt, 2015. Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2927–2936.

Bossard, Lukas, Guillaumin, Matthieu, Van Gool, Luc, 2014. Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. Springer, pp. 446–461.

Chen, Jingjing, Ngo, Chong-Wah, 2016. Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 32–41.

Chu, Xiangxiang, Tian, Zhi, Wang, Yuqing, Zhang, Bo, Ren, Haibing, Wei, Xiaolin, Xia, Huaxia, Shen, Chunhua, 2021. Twins: Revisiting the design of spatial attention in vision transformers. Adv. Neural Inf. Process. Syst. 34, 9355–9366.

Cui, Yin, Song, Yang, Sun, Chen, Howard, Andrew, Belongie, Serge, 2018. Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4109–4118.

Dodge, Samuel, Karam, Lina, 2017. A study and comparison of human and deep learning recognition performance under visual distortions. In: 2017 26th International Conference on Computer Communication and Networks. ICCCN, IEEE, pp. 1–7.

Dong, Xiaoyi, Bao, Jianmin, Chen, Dongdong, Zhang, Weiming, Yu, Nenghai, Yuan, Lu, Chen, Dong, Guo, Baining, 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gao, Xinle, Xiao, Zhiyong, Deng, Zhaohong, 2024. High accuracy food image classification via vision transformer with data augmentation and feature augmentation. J. Food Eng. 365, 111833.

Hassannejad, Hamid, Matrella, Guido, Ciampolini, Paolo, De Munari, Ilaria, Mordonini, Monica, Cagnoni, Stefano, 2016. Food image recognition using very deep convolutional networks. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management. pp. 41–49.

He, Xiangteng, Peng, Yuxin, 2017. Fine-grained image classification via combining vision and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5994–6002.

Hu, Ligang, Zhang, Wei, Zhou, Chenwang, Lu, Guanze, Bai, Haoran, 2018. Automatic diet recording based on deep learning. In: 2018 Chinese Automation Congress. CAC, IEEE, pp. 3778–3782.

Ji, Chao, Deng, Zhaohong, Ding, Yan, Zhou, Fengsheng, Xiao, Zhiyong, 2023. RMMLP: Rolling MLP and matrix decomposition for skin lesion segmentation. Biomedical Signal Processing and Control 84, 104825.

Kaur, Parneet, Sikka, Karan, Wang, Weijun, Belongie, Serge, Divakaran, Ajay, 2019. Foodx-251: a dataset for fine-grained food classification. arXiv preprint arXiv:1907.06167.

Kawano, Y., Yanai, K., 2014. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV). pp. 3–17.

Khanna, Sunil K., 2009. Food and culture: a reader (2nd ed.), by carole counihanand penny van esterik.. Ecol. Food Nutr. 48 (2), 157–159.

Lin, Zhongqi, Mu, Shaomin, Huang, Feng, Mateen, Khattak Abdul, Wang, Minjuan, Gao, Wanlin, Jia, Jingdun, 2019. A unified matrix-based convolutional neural network for fine-grained image classification of wheat leaf diseases. IEEE Access 7, 11570–11590.

Liu, Ze, Hu, Han, Lin, Yutong, Yao, Zhuliang, Xie, Zhenda, Wei, Yixuan, Ning, Jia, Cao, Yue, Zhang, Zheng, Dong, Li, et al., 2022a. Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, Lin, Stephen, Guo, Baining, 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Zhuang, Mao, Hanzi, Wu, Chao-Yuan, Feichtenhofer, Christoph, Darrell, Trevor, Xie, Saining, 2022b. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Martinel, Niki, Foresti, Gian Luca, Micheloni, Christian, 2018. Wide-slice residual networks for food recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 567–576.

Min, Weiqing, Jiang, Shuqiang, Liu, Linhu, Rui, Yong, Jain, Ramesh, 2019. A survey on food computing. ACM Comput. Surv. 52 (5), 1–36.

Parisi, German I, Kemker, Ronald, Part, Jose L, Kanan, Christopher, Wermter, Stefan, 2019. Continual lifelong learning with neural networks: A review. Neural Netw 113, 54–71.

Qiu, Jianing, Lo, Frank P-W, Sun, Yingnan, Wang, Siyao, Lo, Benny, 2022. Mining discriminative food regions for accurate food recognition. arXiv preprint arXiv:2207.03692.

Ridnik, Tal, Lawen, Hussam, Noy, Asaf, Ben Baruch, Emanuel, Sharir, Gilad, Friedman, Itamar, 2021. Tresnet: High performance gpu-dedicated architecture. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1400–1409.

Ródenas, Javier, Nagarajan, Bhalaji, Bolaños, Marc, Radeva, Petia, 2022. Learning multi-subset of classes for fine-grained food recognition. In: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management. pp. 17–26.

Sajadmanesh, Sina, Jafarzadeh, Sina, Ossia, Seyed Ali, Rabiee, Hamid R, Haddadi, Hamed, Mejova, Yelena, Musolesi, Mirco, Cristofaro, Emiliano De, Stringhini, Gianluca, 2017. Kissing cuisines: Exploring worldwide culinary habits on the web. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 1013–1021.

Tan, Mingxing, Le, Quoc, 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Taud, Hind, Mas, J.F., 2018. Multilayer perceptron (MLP). Geom Model Land Chang Scen 451–455.

Taylor, Matthew E., Stone, Peter, 2009. Transfer learning for reinforcement learning domains: A survey. J. Mach. Learn. Res. 10 (7), 1633–1685.

Touvron, Hugo, Cord, Matthieu, Douze, Matthijs, Massa, Francisco, Sablayrolles, Alexandre, Jégou, Hervé, 2021a. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Touvron, Hugo, Cord, Matthieu, Jégou, Hervé, 2022. Deit iii: Revenge of the vit. In: European Conference on Computer Vision. Springer, pp. 516–533.

Touvron, Hugo, Cord, Matthieu, Sablayrolles, Alexandre, Synnaeve, Gabriel, Jégou, Hervé, 2021b. Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 32–42.

Wang, Zirui, Dai, Zihang, Póczos, Barnabás, Carbonell, Jaime, 2019. Characterizing and avoiding negative transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11293–11302.

Xiao, Zhiyong, Su, Yixin, Deng, Zhaohong, Zhang, Weidong, 2022. Efficient combination of CNN and transformer for dual-teacher uncertainty-guided semi-supervised medical image segmentation. Comput. Methods Programs Biomed. 226, 107099.

Xiao, Zhiyong, Zhang, Yuhong, Deng, Zhaohong, Liu, Fei, 2024. Light3DHS: A lightweight 3D hippocampus segmentation method using multiscale convolution attention and vision transformer. NeuroImage 292, 120608.

Xu, Peng, Yin, Qiyue, Huang, Yongye, Song, Yi-Zhe, Ma, Zhanyu, Wang, Liang, Xiang, Tao, Kleijn, W Bastiaan, Guo, Jun, 2018. Cross-modal subspace learning for fine-grained sketch-based image retrieval. Neurocomputing 278, 75–86.

Yanai, Keiji, Kawano, Yoshiyuki, 2015. Food image recognition using deep convolutional network with pre-training and fine-tuning. In: 2015 IEEE International Conference on Multimedia and Expo Workshops. ICMEW, IEEE, pp. 1–6.

Yang, Ze, Luo, Tiange, Wang, Dong, Hu, Zhiqiang, Gao, Jun, Wang, Liwei, 2018. Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 420–435.

Yu, Jun, Tan, Min, Zhang, Hongyuan, Rui, Yong, Tao, Dacheng, 2019. Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans. Pattern Anal. Mach. Intell. 44 (2), 563–578.

Zhao, Bo, Feng, Jiashi, Wu, Xiao, Yan, Shuicheng, 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. Int. J. Autom. Comput. 14 (2), 119–135.

Zhao, Peilin, Hoi, Steven C.H., Wang, Jialei, Li, Bin, 2014. Online transfer learning. Artificial Intelligence 216, 76–102.

Zheng, Min, Li, Qingyong, Geng, Yangli-ao, Yu, Haomin, Wang, Jianzhu, Gan, Jinrui, Xue, Wenyuan, 2018. A survey of fine-grained image categorization. In: 2018 14th IEEE International Conference on Signal Processing. ICSP, IEEE, pp. 533–538.

Zhuang, Fuzhen, Qi, Zhiyuan, Duan, Keyu, Xi, Dongbo, Zhu, Yongchun, Zhu, Hengshu, Xiong, Hui, He, Qing, 2020. A comprehensive survey on transfer learning. Proc. IEEE 109 (1), 43–76.