Journal of Food Engineering 365 (2024) 111833

Contents lists available at ScienceDirect

Journal of Food Engineering

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

High accuracy food image classification via vision transformer with data augmentation and feature augmentation

Xinle Gao, Zhiyong Xiao *, Zhaohong Deng

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China

ARTICLE INFO

Keywords:
Food image classification
Deep learning
Vision transformer
Data augmentation
Feature enhancement
Food health

ABSTRACT

Food image classification is an important research direction in the field of computer vision and machine learning. However food image classification faces great challenges when dealing with foods with similar shapes but different nutritional values. In order to improve this problem, this paper proposes a high-accuracy food image classification with data augmentation and feature enhancement through vision transformer (AlsmViT), which can accurately handle foods with similar shapes but different nutritional values, which is expected to help people better manage their diet and improve their health. Our approach incorporates Augmentplus, LayerScale, and multi-layer perception mechanisms for feature local enhancement. Our models are trained and validated on the public datasets Food-101 and Vireo Food-172, respectively, where the accuracy of the AlsmViT-L model validation set is 95.17% and 94.29%, respectively. Compared with other state-of-the-art self-supervised methods, our proposed method exhibits higher accuracy in food image classification tasks.

1. Introduction

With the popularity of globalization and modern lifestyle, eating habits have changed dramatically (Schulenkorf and Siefken, 2019). However, the changes are not always in a healthier direction. At the same time, health problems such as malnutrition and obesity are on the rise around the world (Ingram et al., 2020), negatively impacting people's health and well-being. Therefore, there is an increasingly urgent need for food classification and diet monitoring. Food image classification is an important research direction in the field of computer vision and machine learning (Ozturk et al., 2023; Konstantakopoulos et al., 2023b). It involves classifying food images into different categories and is often used in applications such as diet monitoring, restaurant menu management, and food nutrition analysis (Nadeem et al., 2023). However, the food image classification task faces challenges, especially when dealing with foods with similar shapes but different nutritional values (Min et al., 2019). Fortunately, with the development of technology, the Convolutional Neural Network (CNN) was born, and the method of predicting food categories based on the Convolutional Neural Network can help to distinguish foods with similar shapes but different nutritional values.

VijayaKumari et al. (2022) used the transfer learning technology Efficientnet-B0 to apply to the convolutional neural network. The model was verified on the food data set Food-101 with an accuracy rate of 80%. Then Chaitanya et al. (2023) proposed to use the pre-trained Inception v3 CNN model to inspire the original customized CNN framework through transfer learning. With this pre-trained model, the learning process is improved and therefore more efficient. The model is used for the recognition and classification of food images, and its accuracy rate for the dataset UEC Food-100 is 78.77%. The Fotios (Konstantakopoulos et al., 2023a) team proposed an image-based automatic diet evaluation system for Mediterranean food, which uses EfficientNet-B2 for pre-training models and their weight evaluation, and classifies food images in the MedGRFood dataset, with Top-1 accuracy of 83.8% and Top −5 is 97.6% accurate. Convolutional neural networks, as currently effective deep learning classification models, have become the dominant method for various image classification tasks, such as in medicine (Khan et al., 2023; Aytaç et al., 2022), geography (Shi et al., 2022; Abdelraouf et al., 2022), and agriculture (Azgomi et al., 2023; Ganguly et al., 2022; Düsenberg et al., 2023). However, it is challenging to apply convolutional neural networks to large-scale food images because of the complexity of category diversity and the high similarity between shapes in large-scale food images, which makes the image classification task very difficult. Therefore, the application of convolutional neural networks to large-scale food image classification still faces many challenges and problems, and further research and improvement are needed.

The proposal of Vision Transformer (Dosovitskiy et al., 2020) in 2020 solves the problem of image category diversity and high shape similarity. Vision Transformer, referred to as ViT, is a deep learning

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

https://doi.org/10.1016/j.jfoodeng.2023.111833
Received 19 June 2023; Received in revised form 19 October 2023; Accepted 9 November 2023
Available online 11 November 2023

model for image classification. The model divides the image into a set of small patches, uses the self-attention mechanism in these small patches to capture the local features in the image, and finally generates a global image representation. This method can achieve relatively good performance in image classification tasks. Through semi-supervised learning, Vision Transformer can better utilize the information of unlabeled image data, improving the performance and generalization ability of the model (Xiao et al., 2022). Therefore,Vision Transformer has aroused widespread interest and has important theoretical and practical value. Vision Transformer has aroused widespread interest and has important theoretical and practical value. Sheng et al. (2022) proposed a lightweight transformer-based deep neural network for food image recognition, which can achieve effective recognition of food images with fewer parameters and lower computational costs. They conduct extensive experiments on three popular food datasets, demonstrating that their approach achieves state-of-the-art performance in applying lightweight neural networks to food image recognition. Knott et al. (2023) proposed an image machine learning program based on a pre-trained visual transformer. It is easier to implement than the current standard method of training convolutional neural networks (CNN). The model achieved the best competitive in competitive classification accuracy on two datasets: apple defect detection and banana ripeness estimation. However, vision transformer also has many shortcomings: ViT has a high dependence on large sample data sets (for example, ImageNet-1k and ImageNet-21k). Due to its large number of parameters, overfitting may occur for small sample data sets (for example, Food-101 and Vireo Food-172). Because small sample data sets cannot provide sufficient sample diversity to cover the full distribution and variation of the data. In this case, the ViT model will rely too much on limited samples and fail to capture broader data patterns, resulting in overfitting. In addition, some studies have shown that as the depth of the vision transformer network increases, its accuracy in image classification tasks can still continue to improve, indicating that the network has not reached saturation. For example, the depth from DeiT-S to DeiT-B increases from 12 layers to 24 layers, and the Top-1 accuracy rate increases from 79.9% to 81.8% (Touvron et al., 2021a); the depth from Swin-T to Swin-S increases from 6 layers to 18 layers, and the Top-1 The accuracy rate increased from 81.2% to 83.2% (Liu et al., 2021).

Therefore, this paper proposes a high-accuracy food image classification with data augmentation and feature enhancement through vision transformer. This method includes Augmentplus, LayerScale and multi-layer perception mechanism of feature local enhancement, referred to as AlsmViT. It solves the problem of Vision Transformer overfitting and premature saturation, and improves the accuracy of Vision Transformer classification. The method can accurately handle foods with similar shapes but different nutritional values, which is expected to help people better manage their diets and improve their health. Our models were trained and validated on the datasets Food-101 (Bossard et al., 2014) and Vireo Food-172 (Chen and Ngo, 2016), respectively, where the accuracy of the validation set of the AlsmViT-L model was 95.17% and 94.29%, respectively, which improved the accuracy by 5.26% and 5.12%, respectively, relative to the ViT-L model.

2. Related work

2.1. Vision transformer

Vision Transformer was proposed by Dosovitskiy et al. The model uses Transformer's self-attention mechanism to divide the image into image blocks and mark these blocks with CLS tokens and image memory tokens. Vision Transformer outperforms traditional convolutional neural network models on large-scale image classification tasks. The proposal for Vision Transformer has aroused widespread interest, and many scholars have joined in the improvement of Vision Transformer. For example, Yuan et al. (2021) introduced a new Token-to-Token attention mechanism, which can divide images into tokens in different ways and add additional convolutional layers to extract local features. A layered Vision Transformer model proposed by Liu et al. (2021), which introduces a translation window for each attention module, and uses grouped convolution modules for cross-window information exchange, thus effectively solving the multi-scale image semantic information expressed problem. An attention-free Transformer model proposed by Zhai et al. (2021) This model uses separable convolution to replace the self-attention mechanism, thereby reducing the number of model parameters and computational complexity. As the Vision Transformer becomes more and more popular, it is also being used in the food field. Sheng et al. (2022) proposed a lightweight Transformer-based deep neural network for food image classification. This network constructs an efficient food image recognition network through transformer grouping and token shuffling and effectively combines Transformers to extract global features. The advantage of MobileNet is to extract local features, thereby reducing model parameters and computational costs. Zhou et al. (2023) proposed a new multi-scale fusion module (MSFM) for high-level features and low-level features to learn multi-scale features. In addition, the spatial attention module (SAM) of the transformer encoder is utilized to capture salient object features in the image to improve the model performance. These models are less general than ordinary transformers but often perform well in certain computer vision tasks because their architectural priors reduce the need to learn task biases from scratch.

2.2. Data augmentation

For supervised training, data augmentation provided by automated design programs, such as Auto-Augment (Cubuk et al., 2019) or RandAugment (Cubuk et al., 2020), is usually employed. Auto-Augment was proposed by Google in 2018 to automatically search for data enhancement strategies through AutoML. The strategy is applied to subsets of the training set, and different subsets of different subsets, enabling the automatic search for the best augmentation strategy. However, one problem with Auto-Augment is that the search space is huge, which searches only in proxy tasks. So Google proposed a simpler data enhancement strategy RandAugment in 2019. Its main idea is to randomly select a series of basic data enhancement operations and randomly set the strength of each operation to generate a set of enhancement strategies. Although the search space of RandAugment is extremely small, it is still necessary to determine the optimal N and M for different data sets, which still have a large experimental cost. Then the University of Freiburg proposed TrivialAugment (Müller and Hutter, 2021). Compared with the previous data enhancement strategy, this method is parameterless, and each image only uses data enhancement once, so it does not require any search space. To better understand how to fully utilize the potential for the Transformer, Touvron et al. (2022) proposed a new data augmentation method 3-Augment. This method is a simple data augmentation method inspired by self-supervised learning. Surprisingly, when using ViT, this method performs better than the usual automatic/learned data augmentation used to train ViT but is prone to overfitting when training with small datasets.

As data augmentation further developed, researchers applied it to real food systems, as shown in Table 1. Sivaranjani et al. (2019) enhanced the cashew nut dataset through simple flip, rescale, shear range and rotation. Phiphiphatphaisit and Surinta (2020) used a variety of data augmentation techniques (for example, rescaling, rotation, width shift, height shift, horizontal flip, shear, and zoom.) to enhance the ETH Food-101 dataset, which is comparable to the data of Sivaranjani et al. (2019) Enhancement methods are more complex. Aguilar et al. (2021) proposed an uncertainty-aware data augmentation method that estimates and uses epistemic uncertainty to guide model training. Compared with the above two data augmentation methods, the advantage of this method is that it can generate new synthetic images from real images that are difficult to classify existing in the training data based on

Table 1
Data augmentation application cases in real food systems.

Year	References	Datasets	Methods
2019	Sivaranjani et al. (2019)	Cashew	Flip, Rescale, Shear range and Rotation.
2020	Phiphiphatphaisit and Surinta (2020)	Food-101	Rescaling, Rotation, Width shift, Height shift, Horizontal flip, Shear, and Zoom.
2021	Aguilar et al. (2021)	MAFood-121	Uncertainty-aware Data Augmentation (UDA).
2022	Zhang et al. (2022)	ZD958H and ND616H	Data enhancement method based on generative adversarial network (GAN).
2023	He et al. (2023)	Food101-LT	CAM-based exemplar augmentation.

Table 2
Ablation of our data augmentation strategy using ViT-B on datasets Food-101 and Vireo Food-172.

Data-Augmentation			Datasets	
RandomResizedCrop	TrivialAugmentWide	RandomErasing	Food-101	Vireo Food-172
✗	✗	✗	83.0	85.1
✓	✗	✗	89.1	90.3
✗	✓	✗	87.6	88.7
✗	✗	✓	84.6	88.7
✓	✓	✓	**90.5**	**90.6**

cognitive uncertainty. Zhang et al. (2022) used the data augmentation method of generative adversarial network (GAN) to enhance haploid corn kernels. The advantage of this method is that GAN can generate new samples with diversity based on the training data. By generating new samples from noise, GAN can increase the number and variation of samples in the data set, providing more sample choices. He et al. (2023) proposed a CAM-based sample enhancement method that uses class activation mapping (CAM) to identify the most important regions from instance rare class images, and then cut and paste the identified regions to the image by executing CutMix. Compared with the above four data enhancement methods, the advantage of this method is that it can retain high-quality semantic information.

2.3. Deeper architectures

Deeper architectures generally lead to better performance, but this complicates their training process. Architectures and optimizers must be tuned to train them properly. For example, the Fixup (Zhang et al., 2019) method can be used to train deep residual networks without normalization layers. And also solves the problem of exploding and vanishing gradients at the beginning of training by properly tuning the standard initialization. With proper regularization, Fixup enables residual networks without normalization to achieve state-of-the-art performance in image classification and machine translation. The method opens up new possibilities for theory and applications. The Skip-init (De and Smith, 2020) method can train deep residual networks without normalization. This method is mainly to apply the activation function of the layer before batch normalization, which can solve the problem of bias introduced by batch normalization, and make the residual block better approximates the identity transformation. The T-Fixup (Huang et al., 2020) method allows training without warmup or layer normalization. The method achieves state-of-the-art accuracy and allows training very deep models with over 1000 MLPs/attention blocks. The ReZero (Bachlechner et al., 2021) method helps deep neural networks propagate signals more efficiently and preserve dynamic contours. ReZero is applied to various residual architectures, including fully connected networks, Vision Transformer, and ResNet, and a significant improvement in convergence speed is observed.

3. Methods

3.1. Data augmentation

Since the advent of AlexNet, the data augmentation procedures used to train neural networks have been significantly modified. Interestingly, the same data augmentation, such as Auto-Augment or RandAugment, is widely used in ViT. Given that the architectural priors and biases in these architectures are very different, the augmentation strategy may not be tuned and may overfine given the large number of choices involved in the selection.

We propose a data augmentation method combining 3-Augment (Touvron et al., 2022), TrivialAugment (Müller and Hutter, 2021) and RandomErasing (Zhong et al., 2020), and rewrite the RandomResizedCrop function, which we call Augmentplus. Augmentplus solves the problem that the 3-Augment method is easy about overfine on small data sets and improves the accuracy of the validation set.

We randomly selected a picture in the food-101 dataset for the data enhancement diagram, as shown in Fig. 1. We consider the following transformations:

(1) RandomResizedCrop: We inherit the RandomResizedCrop class of transforming and rewriting the get_params function. We use the get_image_size functions to read the image size, and we remove the 10 loops in the original get_params function to reduce the amount of calculation.

(2) RandomHorizontalFlip: Randomly flip the image horizontally.

(3) TrivialAugmentWide: This applies a single augmentation to each image.

(4) GausslanBlur: Slightly changes the details in the image.

(5) gray_scale: Increases the grayscale and pays more attention to the shape.

(6) Solarization: Adds strong noise to colors to be more robust to changes in color intensity and thus pays more attention to shapes.

(7) ColorJitter: Increase the brightness of the picture.

(8) RandomErasing: Fill a certain area in the picture with the same pixel value, thereby covering the picture information in the area, forcing the model to learn the features outside the area for recognition, to a certain extent, avoiding the model from falling into a local optimum, thereby improving the performance of the model Generalization.

To show the data enhancement effect of Augmentplus, we provide the ablation of different data enhancement components of ViT-B on the dataset Food-101 and Vireo Food-172, as shown in Table 2. Our method shows better results.

3.2. Deeper image transformers with LayerScale

In order to solve the problem that the depth of the Vision Transformer image classification transformer is not enough and it is prone to premature saturation, we introduced the LayerScale (Touvron et al., 2021b) method proposed by Touvron et al. Compared with Vision Transformer's original image transformer method, LayerScale can effectively improve the training of deeper architectures. LayerScale is a neural network layer normalization technique that adds a learnable diagonal matrix to the output of each residual block. The initial value of this diagonal matrix is close to zero. By adding this simple layer after

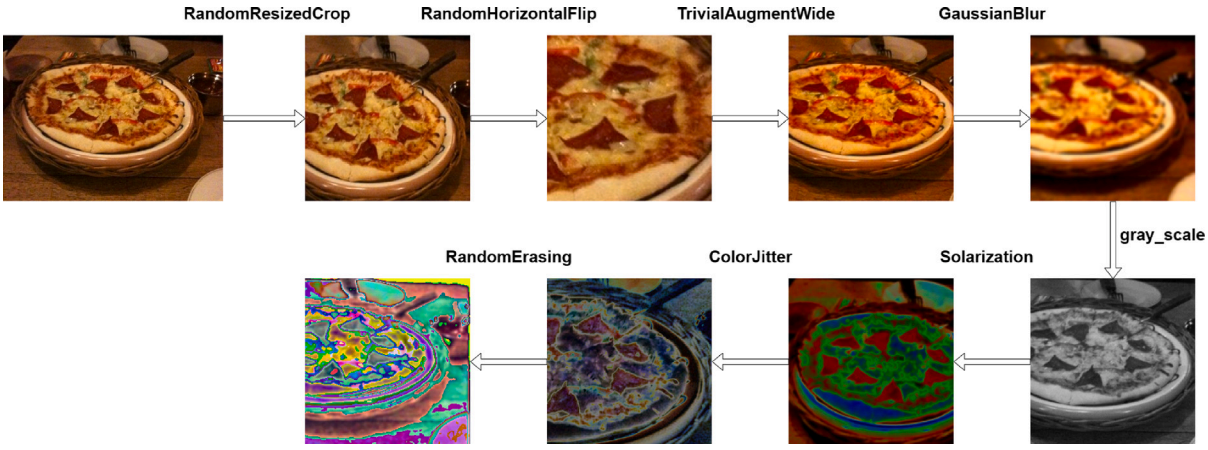

Fig. 1. Augmentplus data enhancement step-by-step illustration.

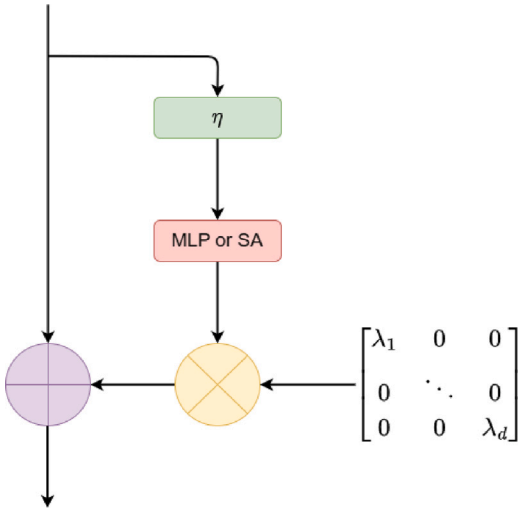

Fig. 2. LayerScale structure.

each residual block, LayerScale can improve the training dynamics, enabling the network to better cope with deep and high-capacity image tasks, resulting in better training results. This approach can benefit neural networks in deep layers and improve their performance.

LayerScale is composed of a matrix–vector, multi-layer perception mechanism, or multi-head attention mechanism. Matrix vector can better increase the convergence of multi-layer perception mechanism or multi-head attention mechanism. LayerScale is to group weight updates related to the same output channel. Formally, LayerScale is the product of the diagonal matrices output by each residual block, as shown in Fig. 2.

The original architectural formula for the Vision Transformer Block section is as follows:

$$x'_l = x_l + SA(x_l) \tag{1}$$

$$x_{l+1} = x'_l + MLP(x'_l) \tag{2}$$

Among them, x_l is the input parameter, x'_l is the residual result, SA() is the multi-head attention mechanism, and MLP() is the multi-layer perception mechanism. The result of the multi-head attention mechanism and forward propagation is directly calculated as the residual, and the result value is too large, resulting in a slower convergence speed. The LayerScale architecture introduces matrix vectors, which can reduce residual results and enhance feature extraction. The formula is as follows:

$$x'_l = x_l + diag(\lambda_{l,1}, \ldots, \lambda_{l,d}) \times SA(\eta(x_l)) \tag{3}$$

$$x_{l+1} = x'_l + diag(\lambda'_{l,1}, \ldots, \lambda'_{l,d}) \times MLP(\eta(x'_l)) \tag{4}$$

Among them, parameters $\lambda'_{l,1}$ and $\lambda'_{l,d}$ are learnable weights, and η is layer normalization. The diagonal values are all initialized to a fixed small value ε: we set it to $\varepsilon = 1e{-}4$. The formulation is similar to other normalization strategies ActNorm (Kingma and Dhariwal, 2018) or LayerNorm (Ba et al., 2016) but performed on the output of the residual block. However, we seek a different effect: ActNorm is a data-dependent initialization that calibrates the activations so that they have zero mean and unit variance. In contrast, we initialize the diagonal with small values such that the initial contribution of the residual branch to the function implemented by the transformer is small. LayerScale provides more variety in optimization than just tuning an entire layer by a single learnable scalar as in ReZero (Bachlechner et al., 2021), Skip-Init (De and Smith, 2020), Fixup (Zhang et al., 2019), and T-Fixup (Huang et al., 2020). As we will show empirically, providing each channel with a degree of freedom to do so is a decisive advantage of LayerScale over existing methods.

In theory, adding these weights does not affect the expressive power of the architecture, because it can be integrated into the existing SA and MLP layer matrix without affecting the role of the architecture.

3.3. Multi-layer perception mechanism for feature local enhancement

The MLP performs point-by-point operations, which are applied to each marker individually. It consists of two linear transformations with a nonlinear activation in between:

$$MLP(x) = \sigma(xW_1 + b_1)W_2 + b_2 \tag{5}$$

where $W_1 \in \mathbb{R}^{C\times K}$ is the weight of the first layer, projecting each token to a higher dimension K. $W_2 \in \mathbb{R}^{K\times C}$ is the weight of the second layer. $b_1 \in \mathbb{R}^K$ and $b_2 \in \mathbb{R}^C$ are deviations. $\sigma(\cdot)$ is the nonlinear activation of GELU in ViT.

As a complement to the SA module, the MLP module performs dimension expansion and nonlinear transformation on each token, thereby enhancing the representation ability of the token. However, the spatial relationship between visually important tokens is not considered. This causes the original ViT to require a large amount of training data to learn these inductive biases.

To combine the advantages of GRN (Woo et al., 2023) to increase feature aggregation and the advantages of CNN to extract local information with Transformer's ability to establish long-range dependencies, we propose a multi-layer perception mechanism with feature local enhancement (MLP-GC), the structure is shown in Fig. 3. The structure

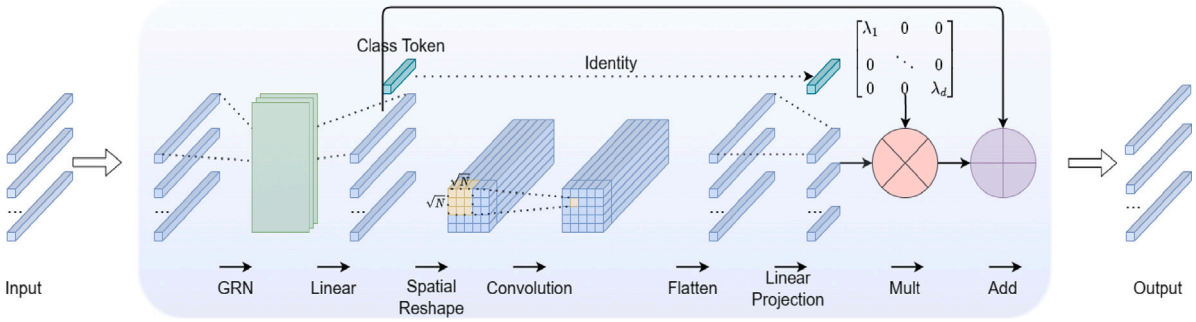

Fig. 3. Schematic diagram of the multi-layer perception mechanism for feature local enhancement.

first obtains the output of the multi-head attention mechanism as the input of the fully connected layer, and then the input of the fully connected layer is optimized by global response normalization and the result is input to the next fully connected layer. Third, split the fully connected layer into Class Tokens and Tokens, and reshape the Tokens space. Fourth, the spatially reshaped Tokens are convolved and flattened. Fifth, the flattened Tokens and Class Tokens are spliced and linearly projected. Sixth, multiply the Linear Projection after the linear projection by a diagonal matrix and add the fully connected layer of the second step.

The multi-layer perception mechanism module of feature local enhancement specifically performs the following process. First, given the label $x_t^h \in \mathbb{R}^{(N+1)\times C}$ generated from the previous SA module, the label is linearly mapped from low-dimensional to high-dimensional through Linear, and then nonlinearly activated by GELU, and the label becomes $x_t^l \in \mathbb{R}^{(N+1)\times C}$. Second, mark $x_t^l \in \mathbb{R}^{(N+1)\times C}$ undergoes feature aggregation through the global response normalization layer to keep the mark unchanged, and the global response normalization layer sets the learning parameters γ and β to zero. Third, mark $x_t^l \in \mathbb{R}^{(N+1)\times C}$ is linearly projected from high-dimensional to low-dimensional through Linear to obtain mark $x_t^s \in \mathbb{R}^{(N+1)\times C}$. Fourth, we split the tokens $x_t^s \in \mathbb{R}^{(N+1)\times C}$ into patch tokens $x_p^s \in \mathbb{R}^{(N+1)\times C}$ and class tokens $x_c^s \in \mathbb{R}^C$, and extend the patch token embeddings to a higher dimension of $x_p^s \in \mathbb{R}^{(N\times e)\times C}$, where e is the expansion ratio. Fifth, patch labels are restored to the image of $x_p^r \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times(e\times C)}$ in the spatial dimension based on their position relative to the original image. Sixth, we perform depth-wise convolution with kernel size k on these restored patch tokens, padding is $(k-1)/2$, the stride is 1, groups are the number of output feature channels, and the enhancement and adjacent k^2-1 order Cards indicate relevance and get $x_p^d \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times(e\times C)}$. Seventh, these patch labels are flattened into a sequence of $x_p^f \in \mathbb{R}^{N\times(e\times C)}$. Eighth, sequence $x_p^f \in \mathbb{R}^{N\times(e\times C)}$ is concatenated with class token $x_c^s \in \mathbb{R}^C$ to obtain $x_t^{h+1} \in \mathbb{R}^{(N+1)\times C}$. Ninth, introduce the diagonal matrix multiplied by $x_t^{h+1} \in \mathbb{R}^{(N+1)\times C}$ to get $(x_t^{h+1})' \in \mathbb{R}^{(N+1)\times C}$, and the value of the diagonal is initialized to a fixed value ε: we set it to $\varepsilon = 0.1$. Tenth, add markers $(x_t^s \in \mathbb{R}^{(N+1)\times C})$ and $(x_t^{h+1})' \in \mathbb{R}^{(N+1)\times C}$ to get $(x_t^{h+1})'' \in \mathbb{R}^{(N+1)\times C}$. These programs can be written as:

$$x_t^l = GELU(LN(Linear(x_t^h))) \tag{6}$$

$$x_t^l = \gamma \times x_t^l \times N(G(x_t^l)) + \beta + x_t^l \tag{7}$$

$$x_t^s = Linear(x_t^l) \tag{8}$$

$$x_c^s, x_p^s = Split(x_t^s) \tag{9}$$

$$x_p^r = SpatialReshape(x_p^s) \tag{10}$$

$$x_p^d = Conv(x_p^r) \tag{11}$$

$$x_p^f = Flatten(x_p^d) \tag{12}$$

$$x_t^{h+1} = Concat(x_c^s, x_p^f) \tag{13}$$

$$(x_t^{h+1})' = diag(\lambda_{l,1}, \ldots, \lambda_{l,d}) \times x_t^{h+1} \tag{14}$$

$$(x_t^{h+1})'' = (x_t^{h+1})' + x_t^s \tag{15}$$

4. Results

4.1. Experiments settings

4.1.1. Datasets

We use the Food-101 (Bossard et al., 2014) dataset and the Vireo Food-172 (Chen and Ngo, 2016) dataset for training and validation of experiments. Food-101 is a very commonly used food image dataset, which includes 101 categories of food images, each category contains 1000 images. There are a total of 101,000 images, each labeled with the correct category. This dataset covers common cuisines from all over the world, including various vegetables, fruits, meals, pasta, meat, seafood, desserts, etc. Created by researchers at the University of Texas at Austin, the dataset aims to provide a unified test benchmark for food image classification tasks in the field of computer vision.

Vireo Food-172 is a classification dataset containing 172 categories, where each category corresponds to a different culinary food or dish. The dataset contains approximately 110,000 images, each of which has been manually labeled for its specific category. The pictures of the Vireo Food-172 data set come from the Internet and social media platforms, such as Instagram (Rich et al., 2016), Yelp (Zhou and Lin, 2016), etc. The images are of varying resolution and quality, with some noise and distortion, so it is a challenging dataset.

4.1.2. Implementation details

All our experiments are performed on an NVIDIA GeForce RTX 2080Ti GPU. We use the weights of the DeiTv2 model trained on the dataset ImageNet-21k as the pre-trained model weights. The Food-101 dataset is split into a training set and a validation set based on the official text. The Vireo Food-172 dataset has been officially split into a training set and a validation set without having to split it yourself. Other parameter settings are shown in Table 3.

4.2. Ablation investigations

In this section, we will introduce the ablation experimental settings and results of AlsmViT-L (Ours) on the Food-101 and Vireo Food-172 datasets, as shown in Table 4.

First, we set ViT without any improvement measures as the baseline. Then we added the data enhancement method Augmentplus on the basis of the baseline, and the accuracy rate increased, indicating that our data enhancement method is very effective. We added LayerScale on the basis of baseline and Augmentplus, which greatly increased the accuracy rate, and also reduced the amount of parameters and

Table 3
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	DeiTv2	ImageNet-21k	224	300	2048	LAMB	3.E−03	Cosine	0.02	5
Finetune	AlsmViT (Ours)	Food-101	224	40	8	AdamW	5.E−04	Cosine	0.05	0
Finetune	AlsmViT (Ours)	Vireo Food-172	224	40	8	AdamW	5.E−04	Cosine	0.05	0

Table 4
Ablation experiment of AlsmViT-L (Ours) on the Food-101 and Vireo Food-172 datasets.

Methods			Food-101				Vireo Food-172			
Augmentplus	LayerScale	MLP-GC	Params (M)	FLOPs (G)	Peak memory (MiB)	Acc. (%)	Params (M)	FLOPs (G)	Peak memory (MiB)	Acc. (%)
✗	✗	✗	305.7	60.0	7404	89.9	305.7	60.0	7404	89.2
✓	✗	✗	304.8	90.0	7393	91.7	304.9	60.0	7394	90.7
✓	✓	✗	303.2	59.7	7472	94.2	303.3	59.7	7473	93.5
✓	✓	✓	303.4	59.7	9262	**95.2**	303.5	59.7	92.7	**94.3**

Table 5
AlsmViT (Ours) compares results with other models on the Food-101 and Vireo Food-172 datasets.

Methods	Epochs	Resolution	Food-101				Vireo Food-172			
			Params (M)	FLOPs (G)	Peak memory (MiB)	Acc. (%)	Params (M)	FLOPs (G)	Peak memory (MiB)	Acc. (%)
ViT-B	40	224 × 224	87.2	17.0	2340	88.33	87.2	17.0	2340	88.84
ViT-L	40	224 × 224	305.7	60.0	7404	89.91	305.7	60.0	7404	89.17
DeiT-T	40	224 × 224	5.5	1.1	313	86.14	5.5	1.1	313	88.24
DeiT-S	40	224 × 224	21.6	4.2	796	88.56	21.7	4.2	796	90.14
DeiT-B	40	224 × 224	85.7	16.9	1971	90.54	85.8	16.9	1972	91.62
MAE-B	40	224 × 224	85.7	16.9	2283	89.95	85.8	16.9	2283	90.84
MAE-L	40	224 × 224	303.2	59.7	6150	92.52	303.3	59.7	6151	92.75
DeiTv2-S	40	224 × 224	21.6	4.2	826	90.49	21.7	4.2	826	90.94
DeiTv2-M	40	224 × 224	38.3	7.5	1260	91.79	38.3	7.5	1261	91.80
DeiTv2-B	40	224 × 224	85.7	16.9	2339	92.19	85.8	16.9	2338	92.21
DeiTv2-L	40	224 × 224	303.2	59.7	7472	94.20	303.3	59.7	7470	93.70
ConvnexTv2-N	40	224 × 224	15.0	2.4	843	86.66	15.0	2.4	843	87.63
ConvnexTv2-T	40	224 × 224	27.9	4.5	1386	87.51	27.9	4.5	1387	88.84
ConvnexTv2-B	40	224 × 224	87.6	15.4	3402	92.80	87.7	15.4	3402	91.92
ConvnexTv2-L	40	224 × 224	196.3	34.4	6168	94.28	196.4	34.4	6169	93.02
Swin-T	40	224 × 224	27.6	4.4	964	90.61	27.6	4.4	964	91.12
Swin-S	40	224 × 224	48.9	8.5	1578	91.91	48.9	8.5	1579	91.78
Swin-B	40	224 × 224	86.8	15.2	2383	92.25	86.9	15.2	2384	92.28
Swin-L	40	224 × 224	195.1	34.1	4435	93.71	195.2	34.1	4437	92.95
Swinv2-T	40	256 × 256	21.2	4.4	2083	91.32	21.2	4.4	2085	90.97
Swinv2-S	40	256 × 256	37.2	8.6	3856	92.24	37.3	8.6	3857	91.67
Swinv2-B	40	256 × 256	66.0	15.2	5933	93.11	66.1	15.2	5934	92.59
Swinv2-B*	40	192 × 192	66.0	8.5	3335	92.32	66.1	8.5	3338	92.35
Swinv2-L*	40	192 × 192	148.2	19.1	6062	93.10	148.3	19.1	6064	92.97
AlsmViT-B (Ours)	40	224 × 224	85.8	16.9	2897	93.60	85.9	16.9	2898	92.95
AlsmViT-L (Ours)	40	224 × 224	303.4	59.7	9262	**95.17**	303.5	59.7	9207	**94.29**

Table 6
Four evaluation metrics of ConvNeXtv2-L, Swin-L, Swinv2-L* and AlsmViT-L (Ours) on Food-101 and Viero Food-172 datasets.

Methods	Epochs	Resolution	Food-101				Vireo Food-172			
			Acc. (%)	Pre. (%)	Rec. (%)	F1. (%)	Acc. (%)	Pre. (%)	Rec. (%)	F1. (%)
ConvNeXtv2-L	40	224 × 224	94.28	94.30	94.28	94.28	93.02	93.01	93.02	92.97
Swin-L	40	224 × 224	93.71	93.72	93.71	93.70	92.95	92.96	92.95	92.91
Swinv2-L*	40	192 × 192	93.10	93.12	93.10	93.10	92.97	92.99	92.97	92.94
AlsmViT-L (Ours)	40	224 × 224	**95.17**	**95.20**	**95.17**	**95.17**	**94.29**	**94.29**	**94.29**	**94.25**

calculations of the model, indicating that LayerScale can enable ViT to train deeper high-capacity images. Finally, we added MLP-GC on the basis of baseline, Augmentplus, and LayerScale, and the optimizer was replaced by AdamW from the original SGD, and the accuracy rate increased again, indicating that MLP-GC can enhance the feature extraction ability of ViT. In the end, we got the best results.

4.3. Comparison with other methods

To demonstrate the high accuracy of the proposed method, we compare our method with baselines on the Food-101 and Vireo Food-172 datasets and six state-of-the-art self-supervised classification methods. The six advanced self-supervised classification methods are (1) Data-efficient image Transformers (DeiT) (Touvron et al., 2021a); (2) Masked Autoencoders (MAE) (He et al., 2022); (3) Data-efficient image Transformers v2 (DeiTv2) (Touvron et al., 2022); (4) ConvNeXtv2 (Woo et al., 2023); (5) Swin Transformer(Swin) (Liu et al., 2021); (6) Swin Transformer v2(Swinv2) (Liu et al., 2022).

We compared the AlsmViT model with the above six advanced self-supervised classification methods in the data sets Food-101 and Vireo Food-172. As shown in Table 5, we show the model parameter amount, computational effort, peak memory and accuracy. Peak memory is measured on a single Nvidai 2080Ti GPU with a fixed batch size of 8 and mixed precision. The accuracy (Acc) can show that the AlsmViT model performs better than the above-mentioned six advanced self-supervised classification methods. This is because the food image classification task may involve a variety of different foods, sometimes with similar local features but different global information. The AlsmViT model

is able to better capture this global information and thereby more accurately distinguish different food categories. However, compared to ConvNeXtv2-L, Swin-L and Swinv2-L*, the AlsmViT-L model has the disadvantages of relatively large number of model parameters, calculation amount and peak memory. Therefore, in addition to Accuracy, we add three commonly used image classification indicators: Precision (Pre), Recall (Rec) and F1-score (F1) to evaluate the performance of the model. As shown in Table 6, we show 4 evaluation results of the AlsmViT-L model and ConvNeXtv2-L, Swin-L and Swinv2-L* on Food-101 and Vireo Food-172. The evaluation results show that the AlsmViT-L model is also better than ConvNeXtv2-L, Swin-L and Swinv2-L* in Pre, Rec and F1. This shows that the AlsmViT model not only has high accuracy and good fitting, but also has strong generalization ability to unseen samples, and can accurately infer data patterns and make correct predictions. The model is therefore able to accurately handle foods that are similar in shape but have different nutritional values. For example, bread and pizza in the Food-101 data set are both round or round-like foods and have certain similarities in appearance. However, eating too much pizza can lead to obesity and malnutrition. Reducing the intake of high-calorie foods is expected to help people better manage their diet and improve their health.

5. Conclusion and future work

This paper introduces a high-accuracy food image classification method with data augmentation and feature enhancement through vision transformer (AlsmViT), designed to handle foods with similar shapes but different nutritional values. 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. Therefore, this method is expected to help people better manage their diet and improve their health by identifying and classifying food images.

Our method consists of the following parts: the data augmentation method Augmentplus, which improves the generalization ability of the data by randomly cropping the image and enhancing the image itself; The depth image converter LayerScale improves the accuracy of the model by training deeper high-capacity images; multi-layer perception mechanism with feature local enhancement (MLP-GC), by introducing global response normalization layer, convolutional neural network, diagonal matrix, and residual module into MLP, it improves the feature extraction ability of the model. Comparison with other state-of-the-art self-supervised methods on public datasets Food-101 and Vireo Food-172. Our proposed method exhibits higher accuracy, precision, recall and f1-score, in food image classification tasks. It shows that this method not only fits well, but also has strong generalization ability to unseen samples, and can accurately infer data patterns and make correct predictions. But our model only showed high-performance results on the food dataset and did not try to experiment on the more general image classification dataset. So our future work is to extend the AlsmViT model to a wider range of image classification and computer vision applications on the basis of food image classification.

CRediT authorship contribution statement

Xinle Gao: Investigation, Methodology, Software, Writing – original draft. **Zhiyong Xiao:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing. **Zhaohong Deng:** Formal analysis, Validation, Visualization, 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 under Grant BK20190079, Natural Science Foundation of China under Grant 62176105 and the Key Research and Development Program of China under Grant 2022YFE0112400.

References

Abdelraouf, Amr, Abdel-Aty, Mohamed, Wu, Yina, 2022. Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways. IEEE Trans. Intell. Transp. Syst. 23 (10), 18546–18556.

Aguilar, Eduardo, Nagarajan, Bhalaji, Khantun, Rupali, Bolaños, Marc, Radeva, Petia, 2021. Uncertainty-aware data augmentation for food recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 4017–4024.

Aytaç, Utku Can, Güneş, Ali, Ajlouni, Naim, 2022. A novel adaptive momentum method for medical image classification using convolutional neural network. BMC Med. Imaging 22 (1), 1–12.

Azgomi, Hossein, Haredasht, Fatemeh Roshannia, Motlagh, Mohammad Reza Safari, 2023. Diagnosis of some apple fruit diseases by using image processing and artificial neural network. Food Control 145, 109484.

Ba, Jimmy Lei, Kiros, Jamie Ryan, Hinton, Geoffrey E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bachlechner, Thomas, Majumder, Bodhisattwa Prasad, Mao, Henry, Cottrell, Gary, McAuley, Julian, 2021. Rezero is all you need: Fast convergence at large depth. In: Uncertainty in Artificial Intelligence. PMLR, pp. 1352–1361.

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.

Chaitanya, A., Shetty, Jayashree, Chiplunkar, Priyamvada, 2023. Food image classification and data extraction using convolutional neural network and web crawlers. Procedia Comput. Sci. 218, 143–152.

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.

Cubuk, Ekin D., Zoph, Barret, Mane, Dandelion, Vasudevan, Vijay, Le, Quoc V., 2019. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123.

Cubuk, Ekin D., Zoph, Barret, Shlens, Jonathon, Le, Quoc V., 2020. Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703.

De, Soham, Smith, Sam, 2020. Batch normalization biases residual blocks towards the identity function in deep networks. Adv. Neural Inf. Process. Syst. 33, 19964–19975.

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.

Düsenberg, Björn, Schmidt, Jochen, Sensoy, Ikay, Bück, Andreas, 2023. Flowability of plant based food powders: Almond, chestnut, chickpea, coconut, hazelnut and rice. J. Food Eng. 111606.

Ganguly, Shreyan, Bhowal, Pratik, Oliva, Diego, Sarkar, Ram, 2022. BLeafNet: a Bonferroni mean operator based fusion of CNN models for plant identification using leaf image classification. Ecol. Inform. 69, 101585.

He, Kaiming, Chen, Xinlei, Xie, Saining, Li, Yanghao, Dollár, Piotr, Girshick, Ross, 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

He, Jiangpeng, Lin, Luotao, Ma, Jack, Eicher-Miller, Heather A., Zhu, Fengqing, 2023. Long-tailed continual learning for visual food recognition. arXiv preprint arXiv:2307.00183.

Huang, Xiao Shi, Perez, Felipe, Ba, Jimmy, Volkovs, Maksims, 2020. Improving transformer optimization through better initialization. In: International Conference on Machine Learning. PMLR, pp. 4475–4483.

Ingram, John, Ajates, Raquel, Arnall, Alex, Blake, Lauren, Borrelli, Rosina, Collier, Rosemary, de Frece, Annabel, Häsler, Barbara, Lang, Tim, Pope, Harley, et al., 2020. A future workforce of food-system analysts. Nat. Food 1 (1), 9–10.

Khan, Sher Shermin Azmiri, Prova, Ayesha Aziz, Acharjee, Uzzal Kumar, 2023. MRI-based brain tumor image classification using CNN. Asian J. Res. Comput. Sci. 15 (1), 1–10.

Kingma, Durk P., Dhariwal, Prafulla, 2018. Glow: Generative flow with invertible 1x1 convolutions. arXiv preprint arXiv:1807.03039.

Knott, Manuel, Perez-Cruz, Fernando, Defraeye, Thijs, 2023. Facilitated machine learning for image-based fruit quality assessment. J. Food Eng. 345, 111401.

Konstantakopoulos, Fotios S., Georga, Eleni I., Fotiadis, Dimitrios I., 2023a. An automated image-based dietary assessment system for mediterranean foods. IEEE Open J. Eng. Med. Biol. 4, 45–54.

Konstantakopoulos, Fotios S., Georga, Eleni I., Fotiadis, Dimitrios I., 2023b. A review of image-based food recognition and volume estimation artificial intelligence systems. IEEE Rev. Biomed. Eng. 1–17.

Liu, Ze, Hu, Han, Lin, Yutong, Yao, Zhuliang, Xie, Zhenda, Wei, Yixuan, Ning, Jia, Cao, Yue, Zhang, Zheng, Dong, Li, et al., 2022. 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.

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

Müller, Samuel G., Hutter, Frank, 2021. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 774–782.

Nadeem, Muhammad, Shen, Henry, Choy, Lincoln, Barakat, Julien Moussa H., 2023. Smart diet diary: Real-time mobile application for food recognition. Appl. Syst. Innov. 6 (2), 53.

Ozturk, Samet, Bowler, Alexander, Rady, Ahmed, Watson, Nicholas J., 2023. Near-infrared spectroscopy and machine learning for classification of food powders during a continuous process. J. Food Eng. 341, 111339.

Phiphiphatphaisit, Sirawan, Surinta, Olarik, 2020. Food image classification with improved MobileNet architecture and data augmentation. In: Proceedings of the 3rd International Conference on Information Science and Systems. pp. 51–56.

Rich, Jaclyn, Haddadi, Hamed, Hospedales, Timothy M., 2016. Towards bottom-up analysis of social food. In: Proceedings of the 6th International Conference on Digital Health Conference. pp. 111–120.

Schulenkorf, Nico, Siefken, Katja, 2019. Managing sport-for-development and healthy lifestyles: The sport-for-health model. Sport Manag. Rev. 22 (1), 96–107.

Sheng, Guorui, Sun, Shuqi, Liu, Chengxu, Yang, Yancun, 2022. Food recognition via an efficient neural network with transformer grouping. Int. J. Intell. Syst. 37 (12), 11465–11481.

Shi, Cuiping, Zhang, Xinlei, Sun, Jingwei, Wang, Liguo, 2022. Remote sensing scene image classification based on self-compensating convolution neural network. Remote Sens. 14 (3), 545.

Sivaranjani, A., Senthilrani, S., Ashokumar, B., Murugan, A. Senthil, 2019. CashNet-15: an optimized cashew nut grading using deep CNN and data augmentation. In: 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE, pp. 1–5.

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.

VijayaKumari, G., Vutkur, Priyanka, Vishwanath, P., 2022. Food classification using transfer learning technique. Glob. Transit. Proc. 3 (1), 225–229.

Woo, Sanghyun, Debnath, Shoubhik, Hu, Ronghang, Chen, Xinlei, Liu, Zhuang, Kweon, In So, Xie, Saining, 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142.

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.

Yuan, Li, Chen, Yunpeng, Wang, Tao, Yu, Weihao, Shi, Yujun, Jiang, Zi-Hang, Tay, Francis E.H., Feng, Jiashi, Yan, Shuicheng, 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 558–567.

Zhai, Shuangfei, Talbott, Walter, Srivastava, Nitish, Huang, Chen, Goh, Hanlin, Zhang, Ruixiang, Susskind, Josh, 2021. An attention free transformer. arXiv preprint arXiv:2105.14103.

Zhang, Hongyi, Dauphin, Yann N., Ma, Tengyu, 2019. Fixup initialization: Residual learning without normalization. arXiv preprint arXiv:1901.09321.

Zhang, Liu, Nie, Qing, Ji, Haiyan, Wang, Yaqian, Wei, Yaoguang, An, Dong, 2022. Hyperspectral imaging combined with generative adversarial network (GAN)-based data augmentation to identify haploid maize kernels. J. Food Comp. Anal. 106, 104346.

Zhong, Zhun, Zheng, Liang, Kang, Guoliang, Li, Shaozi, Yang, Yi, 2020. Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 13001–13008.

Zhou, Wei, Dou, Peng, Su, Tao, Hu, Haifeng, Zheng, Zhijie, 2023. Feature learning network with transformer for multi-label image classification. Pattern Recognit. 136, 109203.

Zhou, Feng, Lin, Yuanqing, 2016. Fine-grained image classification by exploring bipartite-graph labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1124–1133.