

---

# FEATURE-ENHANCED TRESNET FOR FINE-GRAINED FOOD IMAGE CLASSIFICATION \*

---

**LuLu Liu**

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

**Zhiyong Xiao\***

School of Artificial Intelligence and Computer Science  
Jiangnan University  
Wuxi, 214122, China.  
`{Zhiyong Xiao}zhiyong.xiao@jiangnan.edu.cn`

## ABSTRACT

Food is not only a core component of humans' daily diets, but also an important carrier of cultural heritage and emotional bonds. With the development of technology, the need for accurate classification of food images has grown, which is crucial for a variety of application scenarios. However, existing Convolutional Neural Networks (CNNs) face significant challenges when dealing with fine-grained food images that are similar in shape but subtle in detail. To address this challenge, this study presents an innovative method for classifying food images, named Feature-Enhanced TResNet (FE-TResNet), specifically designed to address fine-grained food images and accurately capture subtle features within them. The FE-TResNet method is based on the TResNet model and integrates Style-based Recalibration Module (StyleRM) and Deep Channel-wise Attention (DCA) technologies to enhance feature extraction capabilities. In experimental validation on Chinese food image datasets ChineseFoodNet and CNFOOD-241, the FE-TResNet method significantly improved classification accuracy, achieving rates of 81.37% and 80.29%, respectively, demonstrating its effectiveness and superiority in fine-grained food image classification.

**Keywords** Food Image Classification · CNN · FE-TResNet · Deep Learning · Feature enhancement.

## 1 Introduction

The rapid advancement of computer vision technology has led to the broad application of image classification techniques across a spectrum of fields, including biometric identification, product categorization, and the stringent monitoring of food safety. Fine-grained image classification, a particularly complex and specialized task, has emerged as a pivotal area with significant potential in these critical applications [1, 2]. Unlike conventional image classification, fine-grained classification presents significant challenges due to the high degree of similarity between categories, including the nuanced differences in the appearance of various fruit types and the pronounced distinctions among individuals within the same category, as evidenced by the significant variations in color, size, and shape among different apple cultivars, adding to the complexity of the classification endeavor [3]. Food image classification highlights the inherent challenges in fine-grained classification [4], as complexity is not limited to the food items themselves; the context provided by the background and the perspective determined by the camera angle also exert a profound influence on classification outcomes. Traditional research in food image classification largely relies on manual feature extraction methods [5, 6, 7], that often fail to capture the full spectrum of food characteristics, thereby limiting classification accuracy. In contrast, deep learning techniques, leveraging the hierarchical structure of neural networks, can automatically extract intricate features from food images, significantly enhancing classification precision [8, 9, 10].

Since the advent of deep learning, it has received widespread attention from researchers. Several robust models, such as ResNet [11, 12, 13], Vision Transformer (ViT) [14], Inception [15], EfficientNet [16], and MobileNet [17, 18], have been developed and widely applied to tasks such as image segmentation [19, 20, 21, 22, 23], classification [24], and recognition [4]. Zhiyong Xiao et al. introduced an innovative semi-supervised learning optimization strategy based on

---

\**Citation:* Lulu Liu, Zhiyong Xiao. Feature-Enhanced TResNet for Fine-Grained Food Image Classification



Figure 1: These are example images from food dataset. The above two lines are pictures of tofu. Due to different cooking and ingredients, different tofu foods are presented, from left to right, from top to bottom, in order: Mapo tofu, home-style tofu, fried tofu, tofu flower and stinky tofu; The bottom two lines are potato pictures, which are successively hot and sour shredded potatoes, mashed potatoes, fried potatoes and potato braised beans. Besides, they are placed in different utensils. Different shooting environments and other factors lead to changes in the visual appearance of Chinese food images, which brings challenges to visual food classification.

the teacher-student paradigm that integrates the strengths of Convolutional Neural Networks (CNNs) and Transformers [25]. This integration significantly enhanced the efficiency of utilizing large volumes of unlabeled medical images and improved the efficacy of model segmentation outcomes. Chao Ji et al. proposed an MLP-based model employing matrix decomposition and rolling tensor technology for skin lesion segmentation, replacing the self-attention mechanism of Transformers [26]. The research not only resulted in superior model performance with a reduced parameter count but also demonstrated the innovative application of traditional methods can lead to groundbreaking results in the field of deep learning. Moreover, the adaptability of certain models to image segmentation tasks has been elegantly demonstrated through innovative modifications. This approach merged the capabilities of CNNs and ViTs to craft the Light3DHS model, designed for the segmentation of 3D hippocampal structures [27]. The Light3DHS model not only refined segmentation accuracy but also made a substantial impact on brain disease research, fostering deeper clinical investigations.

Recently, in the realm of food image classification and recognition, Xinle Gao et al. successfully tackled the intricate task of classifying food images with similar morphologies using sophisticated data and feature enhancement strategies, with the Vision Transformer (AlsmViT) [28]. Also in 2024, Zhiyong Xiao et al. introduced a deep convolutional module that, after being integrated into the comprehensive feature representation derived from Swin Transformer, notably enhanced the depth and intricacy of both local and global feature representations [29]. Translating the hierarchical parallelism and evolutionary optimization philosophy of hybrid parallel genetic algorithm to food classification can transcend the limitations of conventional CNNs, particularly excelling in scenarios with class-imbalanced and fine-grained food recognition [30]. A lightweight fine-grained food recognition model integrates an FIP-Attention module for modeling complex ingredient-dish relationships and an FCR-Classifier for refining texture-color features, achieving state-of-the-art performance on multiple benchmark datasets and enabling practical deployment in a mobile dietary monitoring application [31].

Building on these seminal models, the TResNet model has been introduced, achieving a reduction in both the number of parameters and computational complexity (FLOPs). This model sustains high efficiency in GPU-based training and inference processes and offers increased throughput, significantly boosting the accuracy of neural networks. TResNet enhances ResNet50's high throughput and residual architecture and further incorporates the Squeeze-and-Excitation (SE) attention mechanism [32]. This mechanism enables adaptive learning of channel-wise features within the neural network while retaining the original features, thereby further enhancing TResNet's performance. TResNet has become a popular method for computer vision tasks [33, 34], addressing multi-label classification and fine-grained partitioning

[35, 36]. In a study by Zelin Xu et al. on the diagnosis of Alzheimer’s Disease, TResNet was employed as the backbone network, with SK adjustments replacing SE to dynamically adjust the size of the Convolutional kernel for capturing features from various receptive fields, achieving an 86.9% accuracy rate in AD diagnosis [37]. Later, Changhyun Kim et al. combined TResNet and Feature Pyramid Network (FPN) to process multi-label classification tasks in medical images, demonstrating robust performance across different lesion sizes [38]. In recent work, Ji-Hyeon Lee et al. utilized a Gaussian filter to remove noise from imaging results in brain tumor classification within magnetic resonance imaging and applied the Patterned-GridMask method in TResNet, achieving a classification accuracy of 97.74% [35]. The TResNet model is not only widely useful in medicine but also well represented in the category of fine granularity. Also in 2023, Dichao Liu et al. explored fine-grained visual classification on FGVC-Aircraft, Stanford Cars, and Food-11, achieving good accuracy [39].

In fact, not every part of an image holds the same significance for classification tasks. Critical is pinpointing and concentrating on regions that are intimately linked to the classification objectives, as well as capturing the interdependencies among pixels. Within Convolutional Neural Networks (CNNs), the Convolutional operation primarily focuses on areas within a confined receptive field, neglecting the crucial connections that distant pixels may share. Thus, capturing long-range dependencies is vital within the architecture of neural networks. In the realm of fine-grained visual categorization, attention mechanisms have emerged as a pivotal subject of exploration, supporting models in surmounting the adversities stemming from intrinsic similarities between classes and variations within a class, and concentrating on the critical details within localized areas [40]. This study introduces a novel image classification approach termed FE-TResNet, predicated on the TResNet backbone network to enhance feature extraction and multi-scale feature fusion more effectively. The FE-TResNet network comprises two core components: a feature extractor and a feature classifier. The feature extractor includes three distinct parts: the backbone network feature extractor, StyleRM, and DCA. Enhancing the features derived from the backbone network feature extractor with stylized weights allows us to determine feature importance through varying stylized weight values, thereby enhancing fine-grained control over feature stylization. Furthermore, multi-scale feature fusion is achieved by combining features from different channels, preserving spatial dimensions while integrating features from various directions. This approach not only improves the model’s ability to discern global and local features but also enriches the dimensionality of feature representation. For the feature classifier, a generic fine-grained food classifier is utilized to process features that have been processed by the feature extractor. Experiments were conducted on fine-grained food image classification using the ChineseFoodNet and CNFOOD-241 datasets with the FE-TResNet model. The experimental results showed that the FE-TResNet model achieved classification accuracies of 81.37% and 80.29% on these two datasets, respectively. These results significantly outperform existing techniques, thereby validating the model’s exceptional performance and potential in image classification tasks.

#### To summarize, our key contributions are:

- This paper proposes a customized FE-TResNet network to achieve fine-grained image classification.
- StyleRM resolves several key issues, including low utilization of stylization feature extraction parameters, limited capture of complex stylization features, and loss of spatial information. This adaptation facilitates the model’s processing of the texture and structure of the input data.
- DCA enables multi-scale feature fusion by skillfully combining information from different channels. This strategy not only captures objects and details that may be overlooked during the convolution process but also effectively meets the demand for a large number of pixel points in the processing of high-resolution images. Consequently, DCA significantly enhances the flexibility of the network model, enabling more efficient processing and analysis of image data, and improved performance in image recognition and classification tasks.

This paper is explained by the following structure. Section II introduces the structure of the FE-TResNet model. In section III, dataset is used for image classification experiments. Section IV overviews the results and analysis of ablation experiments are introduced in detail. lastly, Section V concludes a summary of the full text and the prospect of future work.

## 2 Related work

### 2.1 TResNet

Since AlexNet’s inception, neural network architectures have trended toward increased depth. The common belief was that deepening the network would proportionally enhance its feature extraction capabilities and, consequently, boost model accuracy. Yet, experimental findings have indicated that there is a threshold beyond which model precision

plateaus. This problem is further compounded by escalating training and testing errors, along with the phenomena of gradient vanishing and network degradation. This insight led to a shift in focus toward widening the network's breadth. The ResNet model was developed in response to these challenges, not only expanding the network's width but also pioneering the introduction of residual connections. These connections effectively avoid the pitfalls of gradient disappearance and network degradation inherent in deep neural networks, paving the way for the creation of even more profound architectures. ResNet rapidly became the foundational model for a variety of computer vision tasks. With the evolution of model variants, there was a collective push to refine residual models further, with the dual goals of lessening the GPU throughput requirements and accelerating computational velocity. The TresNet architecture represents a monumental leap from the ResNet50 model, boasting a 3% improvement in recognition accuracy on the ImageNet dataset. The TresNet architecture introduces three diverse architectural variations, distinguished by their depth and channel count [41], with the most extensive parameter set totaling 77.1M. To further refine this framework, this study optimized the Novel Block-Type Selection layer by incorporating a single convolutional layer for feature extraction post the Anti-Alias Downsampling layer. This optimization not only preserves the integrity of the original features but also reduces the parameter count by 36,022, significantly boosting the model's operational efficiency. A detailed diagram illustrating the components of the model is shown in Figure 2.

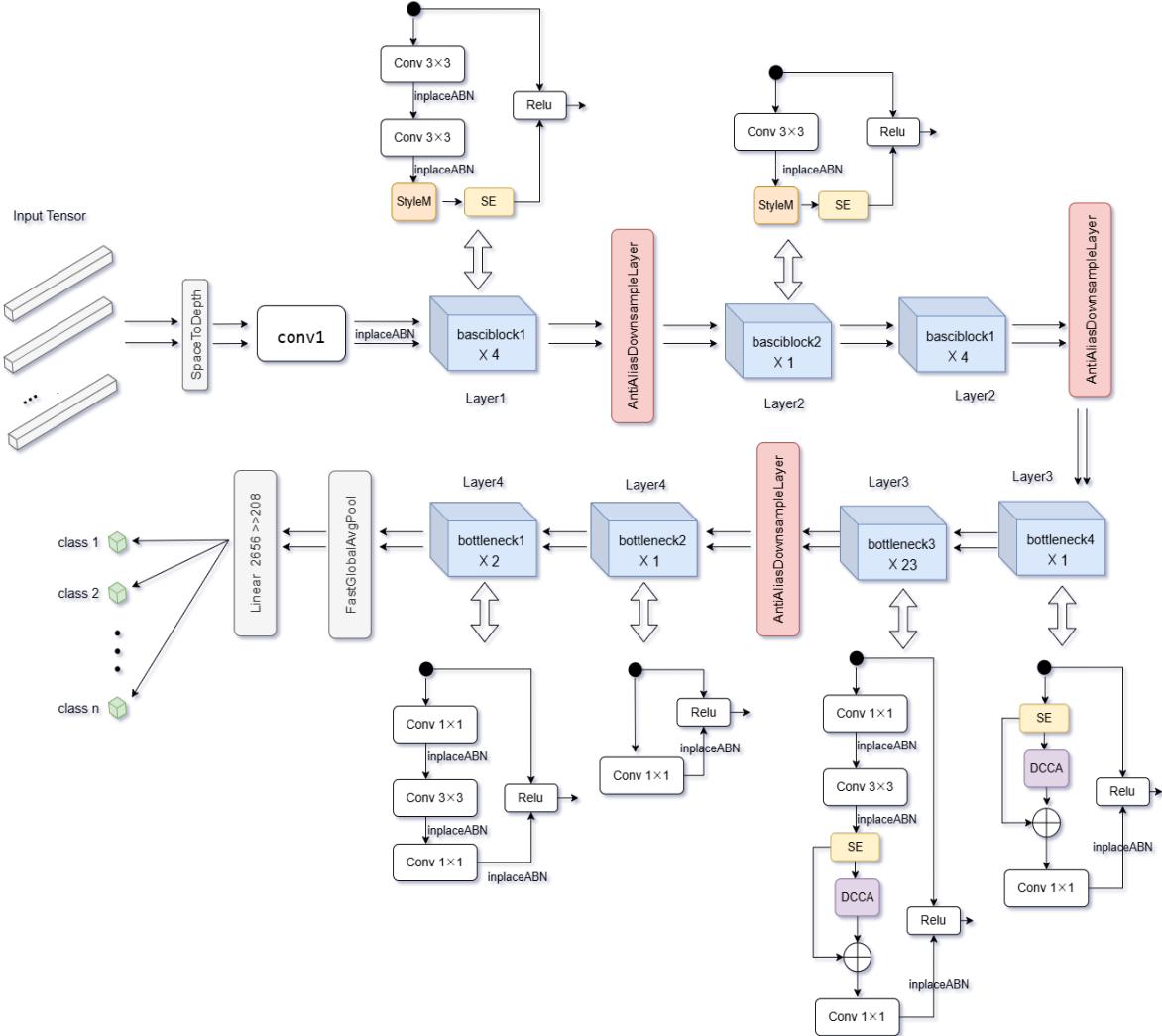


Figure 2: Overall structure of the model. Detailed design of Basicblock and Bottleneck series blocks. The two on the top are basicblock blocks containing the StyleRM and SE attention modules; The four on the footer are bottleneck blocks, and the bottleneck block on the right includes SE and DCA.

## 2.2 StyleRM

Since the studies by Gatys et al., research has overwhelmingly shown that Convolutional Neural Networks (CNNs) are adept at processing not just the morphological aspects of diverse images but also the intrinsic textural attributes, commonly referred to as styles [42, 43, 44]. Recent studies have revealed the intimate nexus between stylistic elements and the CNNs' feature representation, underscoring the pivotal role that textural characteristics play in the mechanism of feature extraction within CNNs [45, 46]. In 2019, Huang et al. proposed the Style-based Recalibration Module for extracting style features, which dynamically assesses the importance of each style and reweights the feature maps accordingly, guiding the network to focus on meaningful style features while ignoring those that are less significant [47]. The module operates through two key steps: Style Pooling and Style Integration. Style Pooling consolidates feature responses across spatial dimensions through global average pooling and global standard deviation pooling, extracting style features from each channel. Style Integration then employs fully connected layer operations to generate specific style weights based on the extracted style features, and by reintegrating the feature maps, it emphasizes or suppresses these feature information, enhancing the representational capacity of CNNs. However, adjusting the stylization of features in each channel by multiplying with a single global content feature coefficient (cfc), a method that, while simple, fails to fully leverage the network's parameter potential. Furthermore, the stylized feature extraction process relies solely on basic multiplication and batch normalization operations, eventually compressed into a  $1 \times 1$  feature map, which limits the model's ability to capture complex style features and may lead to the loss of spatial information. To address these limitations, this study implemented convolutional layers on the original basis to capture spatial information in feature maps. This enhancement not only allows the model to discern more complex stylized features more deeply but also retains important spatial information, thus better understanding and processing the texture and structure in the input data. Furthermore, this approach reduces the risk of overfitting and enhances the overall performance of the model.

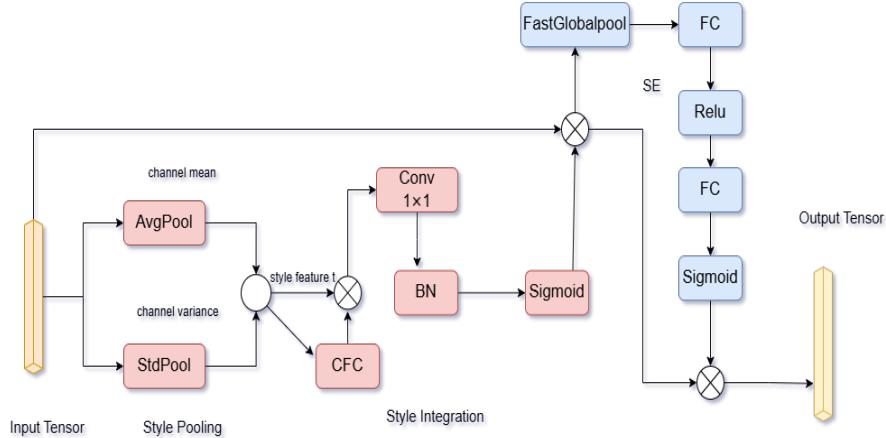


Figure 3: Detailed design diagram of the StyleRM and SE modules in series.

## 2.3 DCA

Capturing internal interdependencies within deep neural networks is of paramount importance. In the context of audio and linguistic data, recurrent operations have been the predominant strategy for addressing these dependencies. For image data, this is typically achieved through the application of Convolutional operations in a deeply stacked and repetitive framework. However, these conventional methods are often constrained to processing dependencies within localized spatial or temporal scopes, posing inherent limitations. Essentially, there are interdependencies present between pixels that are spatially remote within an image, and these long-range dependencies are capable of encapsulating valuable contextual insights. To address this challenge, the concept of Non-Local operations was conceived [48]. This concept incorporates an efficient, simplistic, and universally applicable non-local operational mechanism that directly seizes long-range dependencies by assessing the mutual interactions between any two points in the image, irrespective of their spatial arrangement. The principle underlying Non-Local operations is that the response at any given pixel within the feature map is an aggregate of the weighted features from all other pixels, leading to considerable computational complexity of  $O((H \times W) \times (H \times W))$ , with  $H$  and  $W$  representing the dimensions of the feature map. To enhance the Non-Local approach, the Criss-Cross operation has been developed. This operation introduces a deviation from the standard Non-Local methodology by restricting each pixel's associations to only  $H+W-1$  points within its corresponding row and column, rather than engaging with every pixel across the map. Utilizing

two sequential Criss-Cross operations serves as an effective surrogate for the Non-Local operation, facilitating the acquisition of comprehensive contextual information while significantly lowering the computational complexity to  $O((H \times W) \times (H + W - 1))$ . However, this method is not sufficiently efficient when processing high-resolution images that require more pixel points, and it struggles to capture small objects and details in the images. Expanding upon CCA, this study introduce the technique of depthwise separable convolution, which decomposes standard convolution into depthwise convolution and pointwise convolution. This enables each input channel to independently apply a convolutional kernel and merge features from different channels, achieving multi-scale feature fusion. This facilitates more effective capture of directional features in the images while enhancing the perception of local features. This enhancement not only improves the model's recognition accuracy for small objects and details but also enhances the overall model performance.

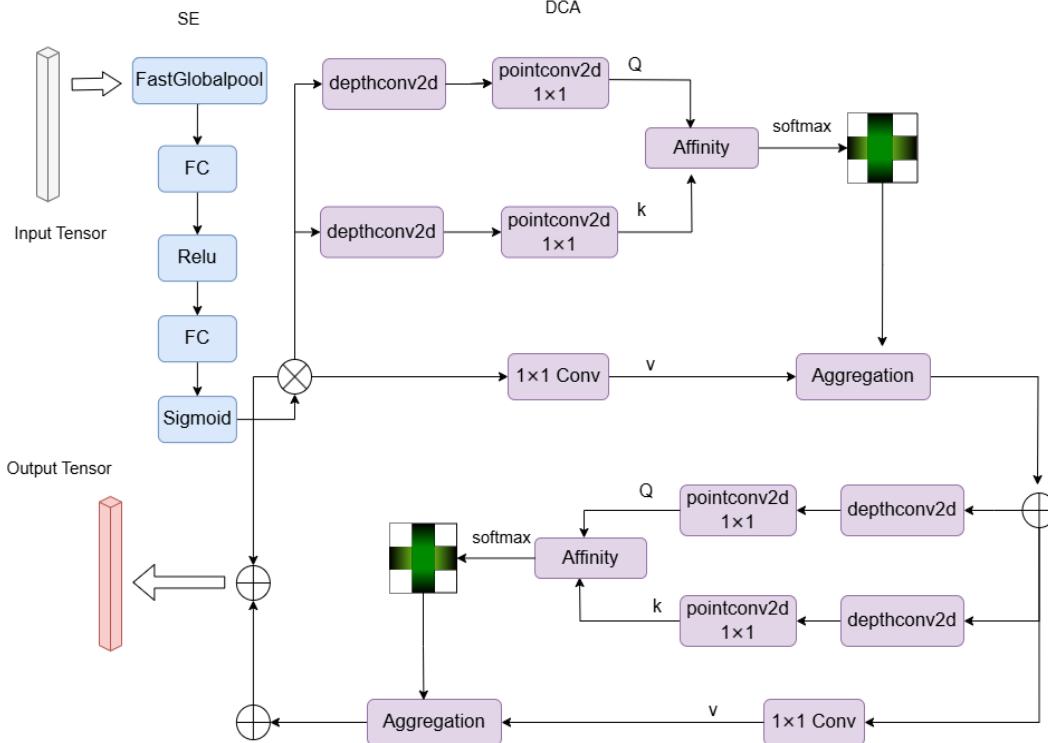


Figure 4: Detailed design diagram of the SE and DCA modules connected in series.

### 3 Materials and method

#### 3.1 Data Preparation

This study delves into two formidable large-scale food image datasets: ChineseFoodNet [49] and CNFOOD-241 [50]. Both datasets have been generously made publicly available by the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, and the University of the Chinese Academy of Sciences for academic research. They have been meticulously divided into training and testing sets and can be accessed through specified links. It should be emphasized that while these datasets are intended for scholarly investigation, they are not to be employed for any commercial use without the appropriate rights and permissions. An example food image is shown in Figure 1. The ChineseFoodNet dataset contains 185628 food images across 208 distinct categories [51]. Comprising images sourced from online recipes and menus, as well as photographs of real dishes and menus captured from everyday life [51]. The images not only demonstrate the rich culinary culture of China but also reflect the dietary habits of the Chinese people, highlighting the diversity and evolution of Chinese cuisine.

The CNFOOD-241 dataset, developed based on ChineseFoodNet, is a comprehensive dataset of Chinese food images that includes 241 categories with a total of 191811 images, all standardized to a size of  $600 \times 600$  pixels. The Chinese cuisine is categorized into five main groups based on the combination of meat and vegetables used in the dishes: mixed meat and vegetables, staple foods, meats, vegetarian diets, and soups. Specifically, the mixed meat and vegetables group

comprises 31 types, representing 16% of the dataset. The staple foods group contains 46 types, accounting for 22%. The meat dishes group consists of 67 types, making up 32% of the dataset. The vegetarian group includes 46 types, also constituting 22%. Finally, the soup group contains 19 types, which account for 9% [50].

### 3.2 The proposed network structure

The feature extractor of the FE-TResNet model consists of the backbone network feature extractor, StyleRM, and DCA. Sample images are fed into the backbone network, and after convolutional processing by the basic blocks of the backbone network, the feature information is mapped to the StyleRM module. Style weights of varying magnitudes are applied to the original feature maps to capture more complex features. Subsequently, the feature maps undergo processing by the remaining SE (Squeeze-and-Excitation) blocks and activation functions of the backbone network's basic blocks to further learn deeper feature representations. As this process unfolds, the model learns richer feature representations while reducing the risk of overfitting. Thereafter, the feature information resulting from convolutional processing by the bottleneck blocks is mapped to the DCA for local feature enhancement, fusing global and local features. Ultimately, following processing by the remaining structure of the bottleneck blocks, the model completes the fusion of multi-scale features while maintaining spatial dimensions, providing greater flexibility to the model. The design of the FE-TResNet model's feature extractor more effectively manages intra-class variability and inter-class similarity in fine-grained food images, achieving more accurate category prediction. This novel methodology offers a new solution for the field of image classification, particularly in food image recognition. The FE-TResNet model structure is shown in Figure 2.

### 3.3 The fusion of StyleRM and SE

In the realm of style transfer learning, this study employ an innovative approach to extract style information from intermediate Convolutional mappings. By leveraging the mean and standard deviation to perform dimensionality reduction on the feature maps, this study extract stylistic information from each channel. Subsequently, through a concise and efficient combination involving a Convolutional layer, batch normalization, and activation functions, this study integrate and adjust the weights for each channel to recalibrate the feature maps effectively. This process not only enhances the model's ability to capture stylistic features but also improves the outcomes of style transfer. To be specific, given input feature maps  $x \in \mathbb{X}^{B \times C \times H \times W}$  the style features  $t \in \mathbb{X}^{B \times C \times 2}$  are calculated by:

$$\beta_b c = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W x_b chw \quad (1)$$

$$\alpha_b c = \sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (x_b chw - \beta_b c)^2} \quad (2)$$

$$t_b c = [\beta_b c, \alpha_b c] \quad (3)$$

Given the style representation  $t \in \mathbb{X}^{B \times C \times 2}$  as an input, the style integration operator performs channel-wise encoding using learnable parameters  $w \in \mathbb{X}^{C \times 2}$ , therefore,  $z_b c = w_c \times t_b c$  where  $z \in \mathbb{X}^{B \times C \times 1}$  represents the encoded style features. Then use BN and activation function operations to get the style weight, the specific implementation formula:

$$\beta_c^{(z)} = \frac{1}{N} \sum_{n=1}^N z_b c \quad (4)$$

$$\alpha_c^{(z)} = \sqrt{\frac{1}{N} \sum_{n=1}^N (z_b c - \beta_c^{(z)})^2} \quad (5)$$

$$z_b c = \gamma_c \left( \frac{z_b c - \beta_c^{(z)}}{\alpha_c^{(z)}} \right) + \beta_c \quad (6)$$

where  $\gamma, \beta \in \mathbb{X}^c$  are affine transformation parameters, and  $g_b c$  represents the channe-wise style weights. Finally, the initial input  $x$  is recalibrated with  $g_b c$  feature maps to emphasize or suppress their information. The Squeeze and Excitation block is a computing unit.  $x \in \mathbb{X}^{B \times C \times H \times W}$  can be constructed based on  $x' \in \mathbb{X}^{B \times C' \times H' \times W'}$ . Firstly, spatial feature compression is performed on the feature map to realize global average pooling in spatial dimension. The feature map of  $C \times 1 \times 1$  is obtained, and a weight vector  $Y$  is obtained through two fully connected layers and an activation function. Finally, the weight vector and the original data are multiplied  $x \times Y$ , and the feature map of the channel with important style features is finally obtained. The structure of the module is shown in Figure 3

### 3.4 The fusion of SE and DCA

The squeezed and Excitation block trained feature graph is taken as the input  $R \in \mathbb{X}^{B \times C \times H \times W}$  of the first DCA model. First, two feature graphs Q and K are obtained by two deep separable convolution $1 \times 1$ . After affinity and softmax, this study get an attention map A. The implementation is as follows, the spatial dimension of Q and K is  $\mathbb{X}^{C' \times H \times W}$ , where  $C' < C$ , at every position u in Q, this study can get a vector  $Q_u \in \mathbb{X}^{C'}$ , meanwhile, in K, With u on the same line can get on the same column  $H + W - 1$  the characteristics of the point set  $K_u \in \mathbb{X}^{(H+W-1) \times C'}$ , Affinity formula for  $D_u = Q_u \times K_u^T$ , where  $D \in \mathbb{X}^{(H+W-1) \times H \times W}$ , Applying the softmax function on the D channel gives the attention map  $A \in \mathbb{X}^{(H+W-1) \times H \times W}$ . The feature adaptive graph  $V \in \mathbb{X}^{C \times H \times W}$  obtained by the third  $1 \times 1$  Convolution in the initial H, similarly,  $V_u$  and  $V_{u'} (V_{u'})$  are feature sets of  $H + W - 1$  points on the same row and column as u can be obtained. Finally, through the Aggregation operation, the  $V_u \times A_u$  of each  $u'$  on the feature graph V is added to the initial feature R, so that  $R'$  is obtained (the sparse connection of each pixel and the global feature). If you want to achieve a dense connection like non-local, you will perform another DCA operation, taking the above  $R'$  as R, to obtain  $R''$  captures the long distance dependence of all pixels. Because the two DCA operations share parameters, the resulting feature map is dense and rich in context information without generating additional parameters. The module structure is shown in Figure 4.

### 3.5 Deeper architectural layers

Deeper networks tend to perform better prior to encountering depth-related bottlenecks in model architecture. However, as the depth of networks increases, challenges such as vanishing and exploding gradients emerge, complicating the training process. Consequently, employing a diverse array of strategies and more sophisticated optimization algorithms is essential to bolster the efficiency of deep learning networks. Gao Huang et al. introduced the innovative concept of stochastic depth [52]. The approach begins training by utilizing a profoundly deep network, strategically discarding a random subset of layers in each mini-batch and circumventing them with an identity function, elegantly overcoming the limitations of residual networks in terms of depth. In 2018, Zijun Zhang et al. advanced the field with the introduction of the normalized direction-preserving Adam (ND-Adam) method [53]. The technique fine-tunes the direction and magnitude of weight updates with greater precision than Adam, effectively mitigating the generalization gap. The following year, Liyuan Liu et al., recognizing issues with adaptive learning rates, proposed warmup works as a variance reduction strategy, thereby enhancing its efficacy and robustness [54]. In recent work, Hao Shao et al. have enhanced deep classification by integrating a linear enhancement of logits within the Softmax function, fostering intra-class compactness and inter-class divergence, and thus, elevating the model's generalization in classification tasks [55].

## 4 Experiment and Result analysis

### 4.1 Experiment preparation

The experiments were conducted using the deep learning framework provided by the PyTorch library, executed on an NVIDIA GeForce RTX 2080Ti GPU. The datasets used were carefully split into training and validation sets. Before training on the training set, input images were randomly cropped and scaled to match the uniform pixel size required by the model. This approach ensures that the model observes different parts of the images with each iteration, effectively preventing overfitting. Moreover, this study implemented automated data augmentation via AutoAugment, diversifying the training dataset through a combination of rotations, cropping, and scaling. Regarding the validation set, after adjusting the input images to a fixed size, a region was cropped from the center using the CenterCrop transformation. These preprocessing steps facilitated the application of image data to the model training. Additional experimental parameters are detailed in Table 1.

Table 1: Parameters in the experimental process

Task	Model	Datasets	Input	Epochs	Batch size	Optimizer	LR	LR decay	weight decay	Warmup epochs
Pretrain	TResNet	ImageNet-21K	224	300	24	SGD	1.E-01	Cosine	0.01	5
Finetune	FE-TResNet	ChineseFoodNet	224	100	48	Adam	1.E-04	Cosine	1.E-5	0
Finetune	FE-TResNet	CNFOOD-241	224	100	48	Adam	1.E-04	Cosine	1.E-5	0

Note: In the table mentioned, "Task" represents the type of problem the model is designed to solve. "Model" represents the different models used. "Datasets" represents the various datasets employed. "Input" refers to the size of the data fed into the model. "Epochs" indicates the number of training cycles. "Batch Size" refers to the number of samples fed into the model during each training iteration. "Optimizer" is used to update the model's weights. "LR" is learning Rate. "LR Decay" reduces the learning rate as training progresses. "Weight Decay" is regularization technique. "Warmup Epochs" stabilizes the training process and prevents large weight updates at the start.

## 4.2 Comparative Parameters

In the realm of deep learning, assessing a model's excellence transcends the Conventional metrics such as parameter count, floating-point operations, and memory footprint. It also encompasses a suite of specific metrics that delineate the model's predictive prowess. Among these are Precision, which quantifies the ratio of true positive predictions to all positive predictions made by the model; Recall, which measures the proportion of actual positive cases correctly identified; and the F1 score, a harmonic mean that balances both Precision and Recall, offering a singular measure of a model's accuracy in binary classification contexts. For classification tasks, the evaluative criteria expand to include Top-1 Accuracy, reflecting the model's ability to precisely predict the most likely category, and Top-5 Accuracy, indicating the model's capacity to list the correct category within its top five predictions. The precise computational formulas for these metrics are articulated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

$$\text{Top - 1 Accuracy} = \frac{P1}{PA} \quad (10)$$

$$\text{Top - 5 Accuracy} = \frac{P5}{PA} \quad (11)$$

In the context of deep learning, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are pivotal terms that define the outcomes of classification models. In the context of the formulas, P1 denotes the count of instances where the model's most probable prediction aligns perfectly with the actual class. P5 refers to the number of cases where the true class is included within the model's top five predicted categories. Lastly, PA signifies the overall count of all instances evaluated in the classification task. With these metrics and their corresponding formulas, one can discern the performance of various models in image classification tasks, particularly in fine-grained categorization. By comparing the experimental results across multiple models, it becomes evident which network architecture excels in fine-grained image classification tasks. The model demonstrating the most outstanding performance is then selected as the benchmark. Subsequently, the methodological enhancements are integrated into the training process, building upon the strengths of this benchmark model. This approach ensures a systematic and empirical method for identifying and refining the most effective deep learning architectures for the nuanced challenges of fine-grained classification. By leveraging the best-performing model as a foundation and incorporating the innovative techniques, this study aim to push the boundaries of what is achievable in the domain of image classification.

## 4.3 Comparative experiments and analysis

This paper have harnessed a suite of esteemed deep learning model architectures to assess two extensive datasets pertinent to Chinese cuisine. (1) The Residual Networks, ResNet50 [11] and ResNet101 [12], are pivotal in mitigating the degradation phenomena within deep networks, averting issues of gradient vanishing and explosion throughout the training regimen; (2) The Dense Convolutional Networks, namely DenseNet121, DenseNet161, DenseNet169, and DenseNet201 [56], amalgamate features across layers on the channel dimension to rival the efficacy of residual networks; (3) The TResNet architecture, comprising TResNet-M, TResNet-L, and TResNet-XL [41], emerged from the quest to enhance the GPU training efficiency of ResNet models; (4) The EfficientNets, including EfficientNetB0 through B3 [16], and its variant EfficientNetV2-S [57], leverage compound scaling and AutoML to dynamically optimize CNNs across various dimensions; (5) The progression of Inception models, InceptionV2 [58], InceptionV3, and InceptionV4 [58, 15], has been geared towards the decoupling of Convolutions, harnessing parallel group Convolutions of assorted sizes to garner richer feature sets; (6) Xception [59, 60], an innovative deep Convolutional architecture inspired by Inception, supplants traditional Convolutions with depthwise separable Convolutions; (7) MobileNet [17, 18], designed for mobile applications, also relies on depthwise separable Convolutions, prioritizing model lightweightness in contrast to Xception's emphasis on performance.

To ensure a fair comparative analysis, uniform image preprocessing was applied to all datasets before experimentation, with consistent configurations maintained throughout. The model, FE-TResNet, was compared with the aforementioned renowned deep learning models on the ChineseFoodNet and CNFOOD-241 datasets. The experimental outcomes, detailed in Table 2, include Parameters (Par.), Floating-Point operations (FL.), Memory Usage (Mem.), Top-1 Accuracy (Top-1 Acc.), and Top-5 Accuracy (Top-5 Acc.). The Top-1 Acc. results show that FE-TResNet outperforms its architectural counterparts in classification precision, adeptly capturing both global and local features and effectively

discerning intra-class variations and inter-class similarities within food imagery. However, FE-TResNet also faces the inherent challenges of high parameter volume, substantial computational demand, and memory intensity, which may elongate training durations.

Table 2: Each model is presented with 50 rounds of best results

Methods	Resolution	ChineseFoodNet					CNFOOD-241				
		Par.(M)	FL.(G)	Mem.(MiB)	Top-1 Acc. (%)	Top-5 Acc. (%)	Par.(M)	FL.(G)	Mem.(MiB)	Top-1 Acc. (%)	Top-5 Acc. (%)
Resnet50	224 × 224	23.93	4.12	3870	78.24	95.65	23.93	4.12	3866	76.49	94.70
Resnet101	224 × 224	42.92	7.84	5120	78.76	95.81	42.92	7.84	5122	76.78	95.02
DenseNet121	224 × 224	7.17	2.88	5156	78.46	95.73	7.17	2.88	5154	76.51	94.70
DenseNet161	224 × 224	26.93	7.82	8348	79.78	96.24	26.93	7.82	8366	78.42	95.61
DenseNet169	224 × 224	12.83	3.42	6002	78.73	95.80	12.83	3.42	5796	77.02	95.11
DenseNet201	224 × 224	18.49	4.37	7182	79.44	96.05	18.49	4.37	7190	77.57	95.38
TResNet-M	224 × 224	29.76	5.74	2996	80.82	96.41	29.76	5.74	2996	78.48	95.76
TResNet-L	224 × 224	54.06	10.88	4366	80.71	96.50	54.06	10.88	4362	79.31	95.90
TResNet-XL	224 × 224	76.33	15.17	5512	80.85	96.72	76.33	15.17	5524	79.85	96.12
EfficientNetB0	224 × 224	4.27	0.40	3654	78.69	95.87	4.27	0.40	3660	76.33	95.01
EfficientNetB1	240 × 240	6.78	0.59	4550	78.39	95.73	6.78	0.59	4548	76.49	95.00
EfficientNetB2	260 × 260	7.99	0.68	4690	78.79	95.84	7.99	0.68	4690	77.20	95.37
EfficientNetB3	300 × 300	11.02	0.99	5676	79.16	95.99	11.02	0.99	5676	77.58	95.47
EfficientNetv2-S	224 × 224	20.44	2.87	5494	81.11	96.69	20.44	2.87	5500	79.61	96.22
InceptionV2	299 × 299	54.63	6.5	4786	78.19	95.55	54.63	6.5	4786	76.86	95.07
InceptionV3	299 × 299	22.21	2.85	3058	75.87	94.61	22.21	2.85	3056	74.28	93.80
InceptionV4	299 × 299	41.46	12.31	4370	77.54	95.26	41.46	12.31	4334	76.30	94.78
MoblienetV2	224 × 224	2.49	0.32	3312	74.38	94.26	2.49	0.32	3310	71.31	93.37
MoblienetV3-s	224 × 224	1.73	0.01	1720	71.19	92.60	1.73	0.01	1718	67.96	91.03
MoblienetV3-l	224 × 224	4.47	0.23	2592	76.44	94.98	4.47	0.23	2592	74.30	93.90
Xception41	299 × 299	25.37	5.03	6460	78.41	95.55	25.37	5.03	6104	76.60	95.02
Xception65	299 × 299	38.30	7.57	7876	78.37	95.84	38.30	7.57	7952	77.09	95.06
Xception71	320 × 320	40.72	9.84	9852	79.06	95.96	40.72	9.84	9769	77.20	95.18
FE-TResNet(study)	224 × 224	82.95	15.82	8006	81.37	97.86	82.95	15.82	8006	80.29	97.97

Note: In the table, "Par.(M)" denotes the number of parameters in millions (Params). "FL(G)" represents the number of floating-point operations required for the model to execute, in billions (Floating Point Operations). "Mem.(MiB)" refers to the memory usage of the model during operation, in megabytes (Megabytes). "Top-1 Acc. (%)" and "Top-5 Acc. (%)" indicate the accuracy of the model's prediction matching the actual category and the accuracy of the actual category being included in the top five most likely categories predicted by the model, respectively.

#### 4.4 Ablation experiment and analysis

The results in Table 2 for Top-1 Acc. and Top-5 Acc. clearly show that, among the ChineseFoodNet and CNFOOD-241 datasets, the TResNet-XL model outperforms others in fine-grained image classification, a testament to its refined architecture derived from ResNet50 and the bolstering presence of the SE module. Its memory consumption is relatively low among the seven models, albeit with a higher volume of parameters. As a whole, TResNet demonstrates pronounced superiority in the realm of fine-grained image classification tasks. Enhancing the TResNet-XL model's capabilities, this study have integrated the StyleRM module, which accentuates stylistic features, and the DCA module, facilitating inter-pixel connectivity across the model's channels. The Top-1 Acc. results in Table 3 show that the incorporation of these modules has markedly improved the TResNet-XL model's fine-grained classification capabilities on the datasets, substantiating the precision of the method in the realm of food image classification. Upon examining the Parameters (Par.), Floating-Point operations (FL.), and Memory usage (Mem.), the method does not impose a significant computational burden or parameter increment on the foundational model, thereby underscoring the formidable adaptability of the FE-TResNet model.

In pursuit of a comprehensive evaluation of the efficiency of the proposed methodology, this study expanded the evaluation to encompass the F1 score, Precision, and Recall on the ChineseFoodNet and CNFOOD-241 datasets for ResNet101, DenseNet161, EfficientNetB2, EfficientNetB3, InceptionV3, Xception65, and the own FE-TResNet method. The results presented in Table 4 indicate that the method surpasses the other seven models across these metrics, indicating an adept balance between the identification of positive samples and the exclusion of negative ones, and an effective utilization of data features for distinguishing between categories. In summary, the FE-TResNet model not

only excels in precision and adaptability but also demonstrates robust generalization capabilities, adeptly discerning minute variations within sample data. This establishes the FE-TResNet model as a potent solution for the fine-grained classification of food images.

Table 3: Ablation experiment of FE-TResNet on the ChineseFoodNet and CNFOOD-241 datasets.

Methods		ChineseFoodNet				CNFOOD-241					
StyleRM	DCA	Par.(M)	FL.(G)	Mem.(MiB)	Top-1 Acc.(%)	Top-5 Acc.(%)	Par.(M)	FL.(G)	Mem.(MiB)	Top-1 Acc.(%)	Top-5 Acc.(%)
✗	✗	76.33	15.17	5512	80.85	96.72	76.33	15.17	5524	79.85	96.12
✓	✗	76.33	15.17	6192	81.22	97.13	76.33	15.17	6178	80.15	97.23
✗	✓	82.94	15.82	7362	80.91	97.03	82.94	15.82	7360	79.92	96.56
✓	✓	82.95	15.82	8006	81.37	97.86	82.95	15.82	8006	80.29	97.97

Note: In the table, "Par.(M)" denotes the number of parameters in millions (Params). "FL(G)" represents the number of floating-point operations required for the model to execute, in billions (Floating Point Operations). "Mem.(MiB)" refers to the memory usage of the model during operation, in megabytes (Megabytes). "Top-1 Acc.%" and "Top-5 Acc.%" indicate the accuracy of the model's prediction matching the actual category and the accuracy of the actual category being included in the top five most likely categories predicted by the model, respectively.

Table 4: Analysis of other aspects of image classification performance.

Methods	Resolution	ChineseFoodNet				CNFOOD-241			
		Accuracy(%)	F1(%)	Precision(%)	Recall(%)	Accuracy(%)	F1(%)	Precision(%)	Recall(%)
Resnet101	224 × 224	78.76	76.34	77.25	76.13	76.78	75.24	75.34	76.24
DenseNet161	224 × 224	79.78	77.60	78.43	77.26	78.42	76.88	76.81	77.81
EfficientNetB2	260 × 260	78.79	76.69	77.62	76.48	77.20	75.55	75.58	76.73
EfficientNetB3	300 × 300	79.16	77.01	77.67	77.02	77.58	76.08	76.19	77.11
InceptionV3	299 × 299	75.87	73.48	74.41	73.16	74.28	73.59	74.54	73.21
Xception65	299 × 299	78.37	76.80	77.51	76.48	77.09	75.99	76.18	76.91
FE-TResNet(study)	224 × 224	81.37	80.37	80.88	80.18	80.29	79.61	79.51	80.27

Note: In the table, "Accuracy(%)" is consistent with the aforementioned "Top-1 Acc.%" . "Precision(%)" represents the precision of the model's predictions that correctly identify the positive class. "Recall(%)" indicates the recall rate, which is the proportion of the actual positive class that is correctly predicted by the model. "F1(%)" denotes the harmonic mean of precision and recall.

## 5 Conclusion and future work

This study introduces a high-precision food image classification method based on TResNet, named FE-TResNet. Through feature enhancement technology, the method can accurately capture subtle features in fine-grained food images, effectively handling food images that are similar in shape but differ in subtle details. To enhance the model's generalization ability, automated data augmentation techniques were utilized. Additionally, incorporating the StyleRM module significantly improved the model's performance in extracting low-level features by assigning different weights to the data. Meanwhile, leveraging DCA technology integrated multi-channel information, which not only promoted the model's capability in multi-scale feature fusion but also addressed the need for processing a large number of pixel points in high-resolution images, thereby enhancing the model's feature extraction performance at the high-level stage. Experimental results on complex food datasets ChineseFoodNet and CNFOOD-241 showed superior accuracy for the model in fine-grained image classification tasks compared to other self-supervised models. Results indicated that the model could better cope with intra-class variability and inter-class similarity, achieving more precise category prediction. Therefore, this method has the potential to help people record and adjust dietary habits, and improve health levels, by identifying and classifying food images. Although FE-TResNet has shown excellent classification performance on food datasets, it remains untested on image datasets in other domains. In the future, based on the foundation of food classification, this paper aim to enhance the model to be more extensible and apply it to a broader range of computer vision fields to solve more complex classification problems.

## Ethical approval

The study does not involve any human or animal testing.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Data availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

## References

- [1] Chi-Sheng Chen, Guan-Ying Chen, Dong Zhou, Di Jiang, and Dai-Shi Chen. Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning. *arXiv preprint arXiv:2402.15761*, 2024.
- [2] Xi Meng, Yingchun Yuan, Guifa Teng, and Tianzhen Liu. Deep learning for fine-grained classification of jujube fruit in the natural environment. *Journal of Food Measurement and Characterization*, 15(5):4150–4165, 2021.
- [3] Berker Arslan, Sefer Memiş, Elena Battini Sönmez, and Okan Zafer Batur. Fine-grained food classification methods on the uec food-100 database. *IEEE Transactions on Artificial Intelligence*, 3(2):238–243, 2021.
- [4] Zhiyong Xiao, Ruke Ling, and Zhaohong Deng. Foodeswin: A high-accuracy food image recognition model for dietary assessment. *Journal of Food Composition and Analysis*, 139:107110, 2025.
- [5] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.
- [6] Aditya Vailaya, Anil Jain, and Hong Jiang Zhang. On image classification: City images vs. landscapes. *Pattern recognition*, 31(12):1921–1935, 1998.
- [7] Zhiyong Xiao and Salah Bourennane. Constrained nonnegative matrix factorization and hyperspectral image dimensionality reduction. volume 5, pages 46–54, 2014.
- [8] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.
- [9] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019.
- [10] Mayank Arya Chandra and SS Bedi. Survey on svm and their application in image classification. *International Journal of Information Technology*, 13(5):1–11, 2021.
- [11] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016.
- [12] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern recognition*, 90:119–133, 2019.
- [13] Brett Koonce and Brett Koonce. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72, 2021.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

- [19] Chen Liu, Zhiyong Xiao, and Nianmao Du. Application of improved convolutional neural network in medical image segmentation. volume 9, pages 1593–1603, 2019.
- [20] Baoxin Qian, Zhiyong Xiao, and Wei Song. Application of improved convolutional neural network in lung image segmentation. *Journal of Frontiers of Computer Science and Technology*, pages 1358–1367, 2020.
- [21] Zhiyong Xiao, Kanghui He, Jianjun Liu, and Weidong Zhang. Multi-view hierarchical split network for brain tumor segmentation. *Biomedical Signal Processing and Control*, 69:102897, 2021.
- [22] Zhiyong Xiao, Nianmao Du, Jianjun Liu, and Weidong Zhang. Sr-net: A sequence offset fusion net and refine net for undersampled multislice mr image reconstruction. *Computer Methods and Programs in Biomedicine*, 202:105997, 2021.
- [23] Zhiyong Xiao, Yang Li, and Zhaohong Deng. Food image segmentation based on deep and shallow dual-branch network. *Multimedia Systems*, 31:85, 2025.
- [24] Yiheng Qian, Zhiyong Xiao, and Zhaohong Deng. Fine-grained crop pest classification based on multi-scale feature fusion and mixed attention mechanisms. *Frontiers in Plant Science*, 16:1500571, 2025.
- [25] Zhiyong Xiao, Yixin Su, Zhaohong Deng, and Weidong Zhang. Efficient combination of cnn and transformer for dual-teacher uncertainty-guided semi-supervised medical image segmentation. *Computer Methods and Programs in Biomedicine*, 226:107099, 2022.
- [26] Chao Ji, Zhaohong Deng, Yan Ding, Fengsheng Zhou, and Zhiyong Xiao. Rmmmp:rolling mlp and matrix decomposition for skin lesion segmentation. *Biomedical Signal Processing and Control*, 84:104825, 2023.
- [27] Zhiyong Xiao, Yuhong Zhang, Zhaohong Deng, and Fei Liu. Light3dhs: A lightweight 3d hippocampus segmentation method using multiscale convolution attention and vision transformer. *NeuroImage*, 292:120608, 2024.
- [28] Xinle Gao, Zhiyong Xiao, and Zhaohong Deng. High accuracy food image classification via vision transformer with data augmentation and feature augmentation. *Journal of Food Engineering*, 365:111833, 2024.
- [29] Zhiyong Xiao, Guang Diao, and Zhaohong Deng. Fine grained food image recognition based on swin transformer. *Journal of Food Engineering*, 380:112134, 2024.
- [30] Zhiyong Xiao, Xu Liu, Jingheng Xu, Qingxiao Sun, and Lin Gan. Highly scalable parallel genetic algorithm on sunway many-core processors. *Future Generation Computer Systems*, pages 679–691, 2021.
- [31] Zhiyong Xiao, Yida Sun, and Zhaohong Deng. Fgfoodnet: Ingredient-perceived fine-grained food recognition for dietary monitoring. *JOURNAL OF FOOD MEASUREMENT AND CHARACTERIZATION*, 2025 JUN 28 2025.
- [32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [33] Ming Chen, Guijin Wang, Jing-Hao Xue, Zijian Ding, and Li Sun. Enhance via decoupling: Improving multi-label classifiers with variational feature augmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1329–1333. Institute of Electrical and Electronics Engineers (IEEE), 2021.
- [34] Inder Pal Singh, Oyebade Oyedotun, Enjie Ghorbel, and Djamilia Aouada. Iml-gcn: Improved multi-label graph convolutional network for efficient yet precise image classification. In *AAAI-22 Workshop Program-Deep Learning on Graphs: Methods and Applications*, 2022.
- [35] Dichao Liu, Longjiao Zhao, Yu Wang, and Jien Kato. Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognition*, 140:109550, 2023.
- [36] Dichao Liu, Longjiao Zhao, Yu Wang, and Jien Kato. Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognition*, 140:109550, 2023.
- [37] Zelin Xu, Hongmin Deng, Jin Liu, and Yang Yang. Diagnosis of alzheimer’s disease based on the modified tresnet. *Electronics*, 10(16):1908, 2021.
- [38] Yu Zheng. Research on x-ray image classification algorithm of covid-19 based on fs-tresn et model. In *2022 10th International Conference on Information Systems and Computing Technology (ISCTech)*, pages 598–604. IEEE, 2022.

- [39] Dichao Liu, Longjiao Zhao, Yu Wang, and Jien Kato. Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognition*, 140:109550, 2023.
- [40] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 852–860, 2022.
- [41] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1400–1409, 2021.
- [42] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.
- [43] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.
- [44] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [45] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.
- [46] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
- [47] HyunJae Lee, Hyo-Eun Kim, and Hyeonseob Nam. Srm: A style-based recalibration module for convolutional neural networks. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 1854–1862, 2019.
- [48] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
- [49] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.
- [50] Bokun Fan, Weiqi Li, Liang Dong, Jingzhen Li, and Zedong Nie. Automatic chinese food recognition based on a stacking fusion model. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4, 2023.
- [51] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.
- [52] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.
- [53] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.
- [54] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.
- [55] Hao Shao and Shunfang Wang. Deep classification with linearity-enhanced logits to softmax function. *Entropy*, 25(5):727, 2023.
- [56] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.
- [57] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

- [58] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [59] Fran ois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [60] Joao Carreira, Henrique Madeira, and Joao Gabriel Silva. Xception: A technique for the experimental evaluation of dependability in modern computers. *IEEE Transactions on Software Engineering*, 24(2):125–136, 1998.
- [61] Ji-hyeon Lee, Jung-woo Chae, and Hyun-chong Cho. Improved classification of different brain tumors in mri scans using patterned-gridmask. *IEEE Access*, 2024.