Food Chemistry 492 (2025) 145276

Contents lists available at ScienceDirect

Food Chemistry

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

FSF-ViT: Image augmentation and adaptive global-local feature fusion for Few-Shot Food classification

Jinhong Li [a,b,d,1], Huiying Xu [a,b,1,*], Xinzhong Zhu [a,b], Jiping Xiong [e], Xiaolei Zhang [c]

[a] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Zhejiang, 321004, China
[b] School of Computer Science and Technology, Zhejiang Normal University, Zhejiang, 321004, China
[c] Academic Affairs Office, Zhejiang Normal University, Zhejiang, 321004, China
[d] Information Engineering College, Jinhua University of Vocational Technology, Zhejiang, 321004, China
[e] College of Physics and Electronic Information Engineering, Zhejiang Normal University, Zhejiang, 321004, China

ARTICLE INFO

Keywords:
Deep learning
Limited training samples
Chinese food image dataset
Low-cost classification technology
Dietary management

ABSTRACT

We proposed FSF-ViT, a Vision Transformer (ViT)-based model integrating image augmentation and adaptive global-local feature fusion, for Few-Shot Food (FSF) classification. The proposed method focused on training with limited food images to reduce data collection and annotation costs. This approach achieved the highest classification accuracy of 95.1% on the test set. Compared to the ViT model, FSF-ViT improved average accuracy by 12.8%, 15.1%, 4.6%, and 8.3% on our constructed Food-30 and three benchmark datasets, respectively. Furthermore, this study visualized the classification results and verified the validity of FSF-ViT. This study provided low-cost and efficient technical support for rapid online dietary recording using smart devices, advancing the development of dietary management and health. (The Food-30 dataset and implementation code: https://github.com/HZAI-ZJNU/FSF-ViT; dataset DOI: 10.5281/zenodo.15619141).

1. Introduction

Dietary factors contribute to reducing risks of chronic diseases such as diabetes, cardiovascular disease, obesity, and cancer (Key et al., 2020). Dietary management serves as an important tool in managing weight and preventing weight gain or support weight loss. Food classification technology enables automatic food recognition and recording, which is an essential technology for dietary management applications (Konstantakopoulos et al., 2023; Nadeem et al., 2023; Xiao et al., 2025a). This technology has diverse applications, including food quality control (Das et al., 2025; Minho et al., 2025; Xiao et al., 2025), nutrition estimation (Kaushal et al., 2024; Shao et al., 2023), and food safety (Feng et al., 2023; Nath et al., 2024; Yang et al., 2025). For instance, it helps determine the market value of agricultural products like apples through automated quality assessment (Hu et al., 2021).

Deep learning enables advanced image processing through highly effective feature extraction (de Oliveira et al., 2023; Deng et al., 2024). Deep learning-based approaches substantially outperform traditional handcrafted feature-based methods in food classification (Konstantakopoulos et al., 2024; Liu, 2019; VijayaKumari et al., 2022). Min et al. (2023) proposed a deep progressive region enhancement network for food recognition, achieving 83.8% accuracy on Food2K, which is currently the largest food image dataset containing over one million images. Xiao et al. (2024) developed a method that integrates global features from Swin Transformer with local features from deep convolution modules, achieving 82.8% accuracy on UEC Food-256 dataset. Liu et al. (2024) proposed a convolution-enhanced dual-branch adaptive transformer, comprising a local fine-grained branch and a global coarse-grained branch, to explore local and global semantically-aware regions across different input images. Their method achieved classification accuracy of 92.4% and 91.6% on ETH FOOD-101 and Vireo Food-172 datasets, respectively. Gao et al. (2024) proposed AlsmViT, an improved Vision Transformer architecture that enhances food classification accuracy among visually similar but distinct categories through data augmentation and feature enhancement, achieving classification accuracy of 95.1% and 94.3% on ETH FOOD-101 and Vireo Food-172 datasets, respectively. Xiao et al. (2025a) introduced DiffAugment data augmentation technology and a local feature enhancement module to improve the model's feature representation capability, achieving validation accuracy of 85.7% and 94.1% on ChineseFoodNet and VireoFood-172 datasets, respectively.

Despite the excellent accuracy of deep learning-based food classification methods, they possess substantial limitations. First, these

* Correspondence to: School of Computer Science and TechnologyZhejiang Normal University, 688 Yingbin Avenue, Jinhua 321004, China.
E-mail address: xhy@zjnu.edu.cn (H. Xu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.foodchem.2025.145276
Received 11 February 2025; Received in revised form 10 June 2025; Accepted 21 June 2025
Available online 12 July 2025

methods rely on large-scale, labeled training samples, which refer to large amounts of categorized food images, requiring manual annotation of food images into their respective categories. These costly and time-intensive methods fail to meet the growing demand for efficient and economical identification techniques. Second, training data class imbalance limits the performance of these deep learning methods. Existing classification methods tend to overlook food categories with limited training samples while focusing primarily on well-represented ones. This bias leads to poor performance in classifying under-represented food categories. Therefore, these methods show limited scalability with imbalanced datasets. This data imbalance phenomenon is prevalent across mainstream food image datasets. Specifically, the large-scale Food2k dataset (Min et al., 2023) exhibits substantial class imbalance with image counts per category ranging from 153 to 1999, while the Vireo Food-172 dataset (Chen & Ngo, 2016) shows a disparity from 191 to 1061 images per category, and the ChineseFoodNet dataset (Chen et al., 2017) ranges from 41 to 1198 images per category. Developing low-cost, effective, and scalable food classification methods for online dietary recording is crucial. This approach promotes dietary management and health.

In this study, we proposed a Few-Shot Food image classification method based on Vision Transformer (ViT) (Dosovitskiy et al., 2021) with image augmentation and adaptive global-local feature fusion (FSF-ViT). The proposed method focused on training with limited food images to achieve accurate food classification, thereby substantially reducing data collection and annotation costs. Unlike existing food classification methods that treated augmented images as ordinary training samples, we explored the complementary relationship between features from augmented and original images to enhance feature learning. Specifically, the image augmentation module consisted of two components: CutCenter and CornerMix, where the CutCenter method extracted and magnified subtle local regions to enhance fine-grained feature learning, and the CornerMix method identified and erased irrelevant areas to reduce noise interference in food images. Based on these augmented images, we proposed an adaptive weighting method to learn the complementary relationship between features from augmented and original images. The learned weights then guided the feature fusion process to generate comprehensive representations that capture both global and fine-grained local features. In addition, we constructed a high-quality Chinese food dataset with limited samples to validate our method. The images were collected from real-world scenarios and accurately labeled with food categories. This research aimed to develop a low-cost, effective, and scalable food classification method to support online dietary recording, thereby facilitating dietary management and health improvement.

2. Materials and methods

2.1. Sample preparation

The samples in this study referred to food images. We constructed a small Chinese food dataset, Food-30, for few-shot food image classification. All food images were collected from real-world dining scenarios. We manually recognized the type of food displayed in these images and grouped them according to their types to create a high-quality dataset. This dataset comprised 30 Chinese food categories, with 110 images per category. Fig. 4 showed representative samples from each category. The Food-30 dataset was divided into the training set and test set with a ratio of 1:10, where each category contained 10 and 100 images in the training and test sets, respectively. To evaluate the generalization ability of the model, images in the training and test sets were collected from different restaurants.

To further validate the effectiveness of our method, we also conducted experiments on three other open-source food image datasets: ChineseFoodNet (Chen et al., 2017), Sushi-50 (Qiu et al., 2019), and Vireo Food-172 (Chen & Ngo, 2016). ChineseFoodNet is a benchmark Chinese food image dataset containing 208 categories of common Chinese cuisine across diverse culinary styles. The dataset presents challenges owing to high intra-class variation in appearance caused by varying cooking techniques. We sampled 2080 training images and approximately 17,000 test images from the original ChineseFoodNet dataset of 208 categories for our experiments. Sushi-50 consists of 50 sushi categories. The dataset contains approximately 4000 images, with 40–100 images per category, from which we sampled 500 training images and around 3000 test images. The Vireo Food-172 dataset contains 172 Chinese food categories. We selected 1720 training images and approximately 33,000 test images from this dataset.

2.2. Methods

We proposed a novel ViT-based few-shot food image classification method with image augmentation and adaptive global-local feature fusion. As shown in Fig. 1, it consisted of three modules: the image augmentation module, feature extractor module and global-local feature fusion module.

2.2.1. Image augmentation module

In food images, subtle differences among categories are typically reflected in subtle local regions, which play a crucial role in food classification. Furthermore, some original food images contain complex backgrounds, which often include dining environments and irrelevant food items that do not belong to the target food category. It is, hence, essential to eliminate background interference from original images, particularly those that contain irrelevant food items. As shown in Fig. 2, the central region of the image typically contains rich food details, whereas the four corner regions often contain distracting backgrounds.

Accordingly, we proposed the CutCenter and CornerMix methods, which were designed to extract and magnify local regions that contain rich food details and erase irrelevant areas, respectively, thereby giving more attention to subtle local regions and the principal regions of target food. Specifically, CutCenter selected a central rectangular region in the original image I and cropped it to a new augmented image I_{CC}, with dimensions half of those of the original image. The augmented image contains less backgrounds whereas food details are magnified, which facilitate the ability of the model to extract fine-grained local features. In addition, to reduce interfering information in the four corner regions of the original image, such as non-target food items and irrelevant objects, CornerMix randomly selected rectangle regions in these areas and erases them. The area S_r of each erasing rectangle I_{re} ranges from 0.02 to 0.1 times the original image area S_o. Following Zhong et al. (2020), the aspect ratio r of the erasing rectangle is randomly sampled from [0.3, 1/0.3]. The height H_p and width W_p of I_{re} are determined as follows:

$$S_r = random(0.02,\ 0.1) * S_o, \tag{1}$$

$$r = random(0.3,\ 1/0.3), \tag{2}$$

$$H_p = \sqrt{S_r * r}, \tag{3}$$

$$W_p = \sqrt{S_r / r}. \tag{4}$$

The augmented image I_{CM}, generated through CornerMix, had fewer distracting backgrounds than the original image, because of which the principal regions of the target food were more easily attended to and learned by the model. In summary, our method enhanced global-local feature representations at the image-level.

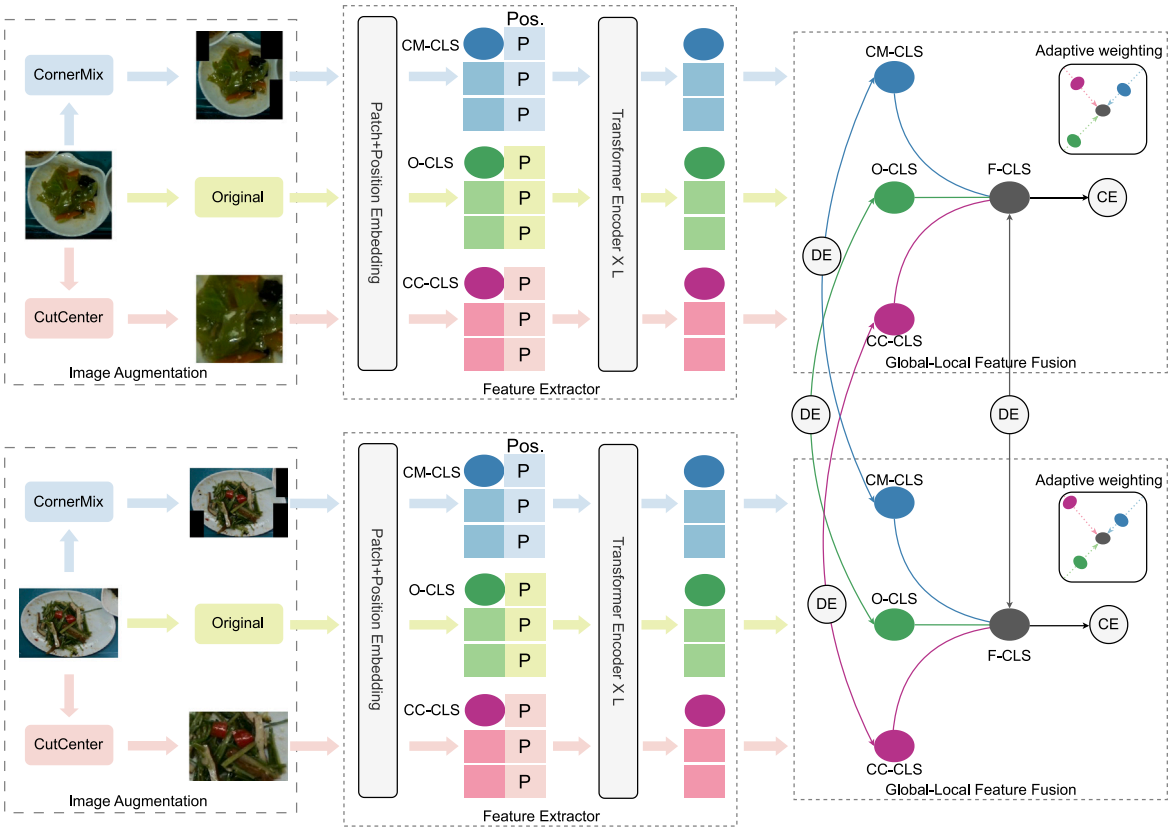

Fig. 1. The framework of FSF-ViT model. The FSF-ViT model is primarily composed of the image augmentation module, the feature extractor module, and the global-local feature fusion module.

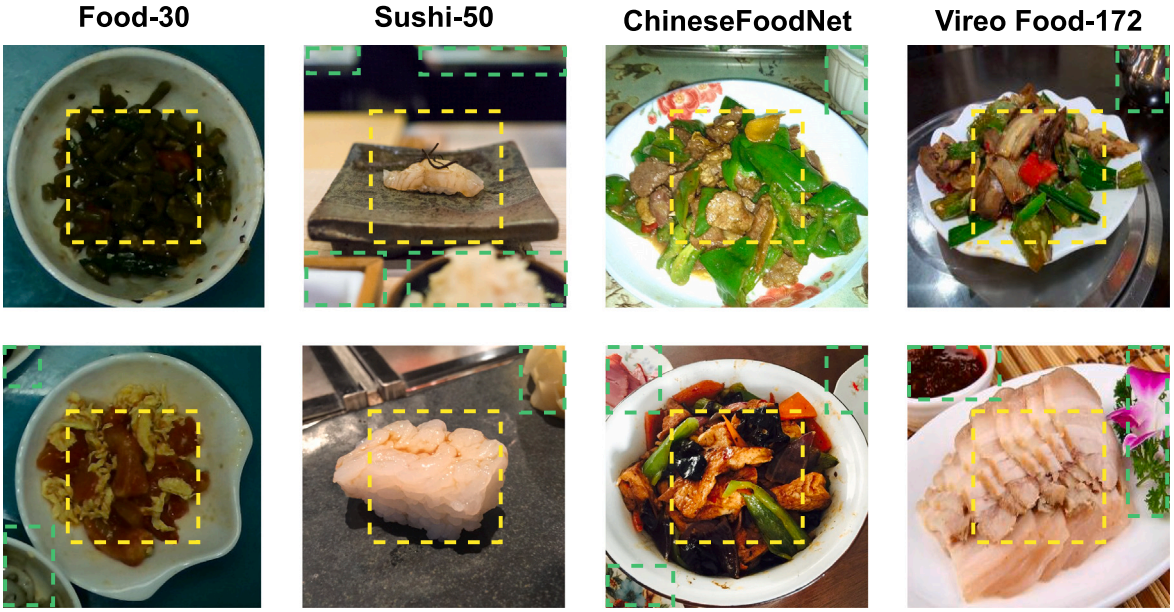

Fig. 2. Some examples of food images from Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172 datasets.

2.2.2. Feature extractor module

The ViT effectively captures relationships between arbitrary regions in an image, making it particularly suitable for addressing challenges of high class diversity and high shape similarity in food images. Consequently, we adopted ViT as our feature extraction network module. In particular, the image set (I, I_{CC}, I_{CM}) outputted by the image augmentation module served as input for this module. For an input image with size $W \times H \times 3$, we divided it into N non-overlapping patches $p_i \in \mathbb{R}^{P\times P\times 3}$, where $N = W \times H / P^2$. These patches were flattened and linearly projected to patch embeddings $V \in \mathbb{R}^{N\times d}$. A class token $V_{CLS} \in \mathbb{R}^d$ was added as the learnable parameter to gather global information, after which it was concatenated with V to obtain the expanded embeddings $V_P \in \mathbb{R}^{(N+1)\times d}$. In addition, to encode position information, we incorporated learnable embedding parameters $V_{Pos} \in$

$\mathbb{R}^{(N+1)\times d}$ by adding them to V_P. The combined representation that was fed into the transformer encoder is expressed as:

$$V_{TE} = V_P + V_{Pos} = [V_{CLS}; V] + V_{Pos}, \quad V_{TE} \in \mathbb{R}^{(N+1)\times d}. \tag{5}$$

The transformer encoder module consists of L stacked encoder blocks, each primarily consisting of layer normalization, multi-head attention, and mlp block. The input to the j-th encoder block, denoted by $S_{j-1} \in \mathbb{R}^{(N+1)\times d}$, where $j \in [1, L]$, is the output from the $(j-1)$-th encoder block. In addition, the dimension of the output of each encoder block remains consistent with that of its input. Finally, the transformer encoder produces features that assume the form $S_L = [S_L^{CLS}; S_L^1, \ldots, S_L^N]$, where $S_L^{CLS} \in \mathbb{R}^d$ is generated from the trainable class token V_{CLS}. Herein, an image set (I, I_{CC}, I_{CM}) was simultaneously input into the transformer encoder module, generating a set of class tokens, denoted by (C^O, C^{CC}, C^{CM}). By extracting features from both the original images and augmented images, we obtained multi-feature that captured both global and local characteristics. The relationships among these features were further explored in subsequent stages.

2.2.3. Global-local feature fusion module

The effective integration of features extracted from both original and augmented images can enhance feature performance, which is essential for food image classification. We selected four images from four different datasets and generated their corresponding augmented versions. Fig. 3 showed the attention maps of both original and augmented images generated via gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017). Using the ViT model, we separately trained on original and augmented datasets. Fig. 3 showed that while the activated regions (highlighted in warm colors) in original images only partially covered the regions within red dashed boxes, these areas were almost fully covered by the activated regions in augmented images. Augmented images were capable of capturing additional food details that complemented those in the original images, indicating that features from augmented and original images had a complementary relationship. Therefore, fusing these two types of features could enhance feature representation. For food image classification, some discriminative features should have been assigned higher weights owing to their crucial role in distinguishing similar categories. To address this issue, we proposed an adaptive weighting method with theoretical guarantee to learn the weight relationships between different features. Specifically, the fused feature $C^{FF} \in \mathbb{R}^d$, denoted as β for simplicity, was computed from multiple latent representations $C^1, \ldots, C^L$ by minimizing the following adaptive feature fusion loss:

$$\ell_F = \min_{a^l} \sum_{l=1}^{L} a^l \|\beta - C^l\|_F^2, \quad \beta = \sum_{l=1}^{L} a^l C^l, \quad s.t., \quad \sum_{l=1}^{L} a^l = 3, a^l > 0. \tag{6}$$

where $C = [C^O, C^{CC}, C^{CM}]$ and a^l denotes the weight of C^l, with $L = 3$. Considering the constraint on a^l, we solved this optimization problem using the Lagrange multiplier method. By introducing a Lagrange multiplier η, Eq. (6) can be re-formulated as:

$$\min \ell_F(a^l, \eta) = \sum_{l=1}^{L} (a^l)^p \|\beta - C^l\|_F^2 + \eta(\sum_{l=1}^{L} a^l - 3). \tag{7}$$

For simplicity, we define $e^l = \|\beta - C^l\|_F^2$. Given $L = 3$, the partial derivatives of $\ell_F(a^l, \eta)$ with respect to a^l and η are derived as:

$$\begin{cases} \frac{\partial \ell_F}{\partial a^l} = p(a^l)^{p-1} e^l + \eta \\ \frac{\partial \ell_F}{\partial \eta} = \sum_{l=1}^{L} a^l - 3 \end{cases} \Rightarrow \begin{cases} \frac{\partial \ell_F}{\partial a^1} = p(a^1)^{p-1} e^1 + \eta \\ \frac{\partial \ell_F}{\partial a^2} = p(a^2)^{p-1} e^2 + \eta \\ \frac{\partial \ell_F}{\partial a^3} = p(a^3)^{p-1} e^3 + \eta \\ \frac{\partial \ell_F}{\partial \eta} = a^1 + a^2 + a^3 - 3 \end{cases} \tag{8}$$

Therefore, we set the partial derivatives to zero, obtaining

$$\begin{cases} a^1 = (\frac{e^2}{e^1})^{\frac{1}{p-1}} a^2 \\ a^3 = (\frac{e^2}{e^3})^{\frac{1}{p-1}} a^2 \\ a^1 + a^2 + a^3 = 3 \end{cases} \tag{9}$$

where a^l is updated to the following form:

$$a^l = \frac{(e^l)^{\frac{1}{1-p}}}{\sum_{l=1}^{L} (e^l)^{\frac{1}{1-p}}} \tag{10}$$

Therefore, the fused features β can be represented as

$$\beta = a^1 C^O + a^2 C^{CC} + a^3 C^{CM}. \tag{11}$$

2.2.4. Multi-loss function

We designed a multi-loss function to guide the learning of features from original and augmented images, and their feature fusion.

Cross-entropy loss is widely utilized in image recognition, served as the loss function for the ViT model. For highly similar samples from different classes, the network was compelled to extract features with higher confidence to minimize the cross-entropy loss. This led to feature learning based on specific samples, resulting in a poor generalization capability. Owing to the high inter-class similarity among food images, using only the cross-entropy loss led to overfitting. The cross-entropy is defined as

$$L_{CE}(X, Y) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} y_{ij} \log f_j(x_i). \tag{12}$$

We considered a C-class classification problem. Let $S = \{s_1, s_2, \ldots, s_n\}$ denote a dataset. Let $X \in \mathbb{R}^{n\times d}$ denote the feature space, and let $Y = \{1, \ldots, C\}$ denote the label space. There exists a set $\{(s_i, x_i, y_i)\}_{i=1}^n$, where each $(s_i, x_i, y_i) \in (S \times X \times Y)$. The classifier was defined as a function $f : X \to \mathbb{R}^{n\times C}$ that mapped the feature space to the class probability space. Here, $f_j(x_i)$ represents the probability of feature x_i being classified as that of class j, and y_{ij} corresponds to the j-th element of the one-hot encoded label of sample s_i.

Pairwise confusion loss was introduced into the loss function as an additional regularization term. This function encouraged the model to learn more generalized features rather than sample-specific ones, thereby addressing the overfitting problem caused by solely relying on the cross-entropy loss alone. The pairwise confusion function is formulated as follows:

$$D_{EC}(x_i, x_k) = \|f(x_i) - f(x_k)\|_2^2, \tag{13}$$

where $f(x_i)$ denotes the class probability of the sample x_i, and $f(x_k)$ the class probability of the sample x_k, $(k \neq i)$.

$\mathbf{L_{VF}}$, combining cross-entropy loss and pairwise confusion loss, was proposed as our loss function to better supervise multi-feature learning. The proposed loss algorithm was developed in three sequential stages. First, we focused on the cross-entropy loss of the fused features C^{FF}. Let $X^F \in \mathbb{R}^{n\times d}$ denote the C^{FF} space. By substituting X^F for X in Eq. (12), the cross-entropy loss L_{ce}^{FF} of C^{FF} is formulated as follows:

$$L_{ce}^{FF} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} y_{ij} \log f_j(x_i^F). \tag{14}$$

Subsequently, we computed the pairwise confusion loss for the fused features C^{FF}. For each pair of features $(x_i^F, x_{i+[n/2]}^F)$ and their corresponding class probability $(f(x_i^F), f(x_{i+[n/2]}^F))$, the pairwise confusion loss L_{CFF} of C^{FF} was incorporated into the loss of our model as follows:

$$L_{CFF} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} y_{ij} \log f_j(x_i^F) + \sum_{i=1}^{[n/2]} \|f(x_i^F) - f(x_{i+[n/2]}^F)\|_2^2. \tag{15}$$

Finally, we incorporated the loss of features from both the original and augmented images to supervise feature learning, thereby obtaining better pre-fusion features, which in turn improved the fused feature representation. We computed the cross-entropy and pairwise confusion loss for four class tokens $(C^O, C^{CC}, C^{CM}, C^{FF})$. Let $X^O \in \mathbb{R}^{n\times d}$ denote the class token C^O space, $X^{CC} \in \mathbb{R}^{n\times d}$ denote the class token C^{CC} space, $X^{CM} \in \mathbb{R}^{n\times d}$ denote the class token C^{CM} space, and

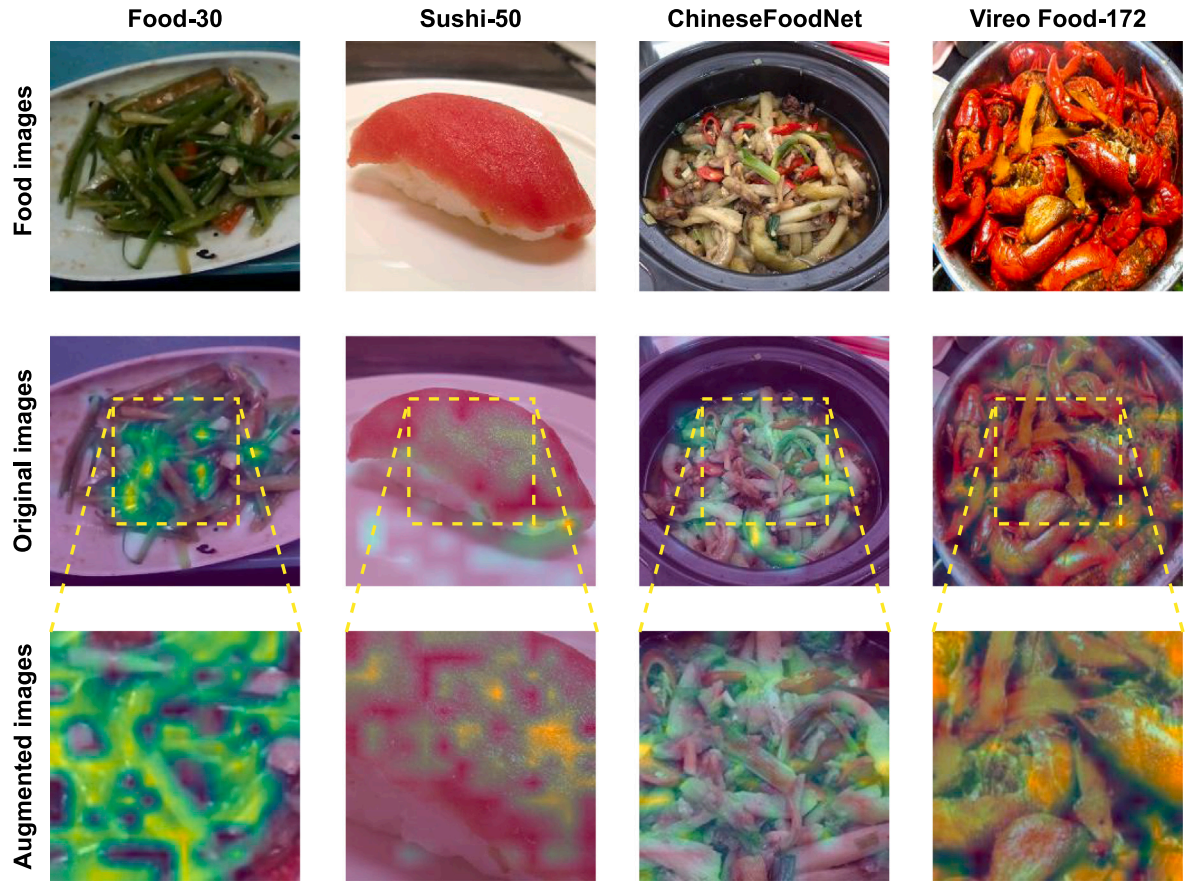

Fig. 3. Visualization results of ViT on original and augmented images (the warmer the color of the overlay image, the more discriminative that pixel is). All these augmented images are resized into the same fixed size. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$X^F \in \mathbb{R}^{n\times d}$ denote the fused features C^{FF} space. There exists a set $\{(s_i, x_i^F, x_i^O, x_i^{CC}, x_i^{CM}, y_i)\}_{i=1}^n$, where each $(s_i, x_i^F, x_i^O, x_i^{CC}, x_i^{CM}, y_i) \in (S\times X^F \times X^O \times X^{CC} \times X^{CM} \times Y)$. $f(x_i^O)$, $f(x_i^{CC})$, $f(x_i^{CM})$, and $f(x_i^F)$ separately represent the class probability of the feature $x_i^O, x_i^{CC}, x_i^{CM}$, and x_i^F. We defined T as (O, CC, CM, F). Our final loss function L_{VF} is formulated as

$$L_{VF} = \sum_{k=O}^{T}(-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} y_{ij}\log f_j(x_i^k)) + \sum_{k=O}^{T}\sum_{i=1}^{[n/2]} \|f(x_i^k) - f(x_{i+[n/2]}^k)\|_2^2. \tag{16}$$

The detailed model training procedure was outlined in Algorithm 1.

Algorithm 1 Training Algorithm of the FSF-ViT Model

Input: Data $D = \{(s_i,\ y_i)\}_{i=1}^n$
Initialize weights $a^l = 1$
for $epoch \in [0,\ epochs]$ **do**
  $\sum_{l=1}^{L} D^l \Leftarrow D$, through the image augmentation module
  $\sum_{l=1}^{L} C^l \Leftarrow \sum_{l=1}^{L} D^l$, through the ViT module
  Compute $\beta = \sum_{l=1}^{L} a^l C^l$
  Update $\mathbf{a^l} = \frac{(e^l)^{\frac{1}{1-p}}}{\sum_{l=1}^{L}(e^l)^{\frac{1}{1-p}}}$ and $\beta = \sum_{l=1}^{L} a^l C^l$
  $T = [\sum_{l=1}^{L} C^l,\ \beta]$
  Compute Loss: $L = \sum_{i=1}^{n}\sum_{t=C^1}^{T} L_{VF}(t_i)$
end for

2.3. Experimental setup

Few-shot image classification follows K-shot setting with K labeled images per class. Typically, K denotes a small number, e.g. K=1, 5, or 10 (Song et al., 2023; Xu et al., 2022). Achieving satisfactory accuracy by using a limited number of labeled samples is challenging, which requires strong feature extraction capabilities. We evaluated four datasets: Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172, where we selected k samples per category to form the training sets for k-shot experiments, with test sets of 3000, 3258, 16,867, and 33,154 images, respectively. The diverse scales of test sets strengthened the reliability of our experimental validation. All experiments were conducted on PyTorch with an NVIDIA RTX 3090 GPU. Top-1 accuracy (Top-1 Acc.) was adopted as the evaluation criterion. Images were resized to 224×224 pixels and were trained for 100 epochs with a batch size of 8. To address concerns about potential bias in training, we conducted experiments with multiple random seeds. Specifically, all reported results represented the average performance across three independent runs with different random initializations, and we included the standard deviations to demonstrate the robustness of our method.

3. Results

3.1. Quantitative evaluation

Conventional data augmentation techniques achieve enhancement at the data level by means of methods such as flipping and scaling. These approaches typically regard augmented images only as normal training samples, without considering the relationship between features from original and augmented images. However, our method

Table 1
Performance comparison on different methods (%). (1) The OG method only utilizes original images for training. (2) The OA method uses original and augmented images for training. (3) Building upon OA, our method combines features from both augmented and original images to enhance feature extraction capability.

Method	Food-30			Sushi-50			ChineseFoodNet			Vireo Food-172		
	10-shot	5-shot	1-shot	10-shot	5-shot	1-shot	10-shot	5-shot	1-shot	10-shot	5-shot	1-shot
OG	87.2 ± 0.3	83.2 ± 0.4	64.8 ± 0.4	65.2 ± 0.0	55.5 ± 0.7	25.2 ± 0.3	48.8 ± 0.1	41.3 ± 0.2	29.1 ± 0.1	61.8 ± 0.1	54.2 ± 0.1	33.5 ± 0.5
OA	88.7 ± 0.0	87.4 ± 0.2	73.3 ± 1.0	69.3 ± 0.1	62.1 ± 0.3	31.5 ± 0.2	53.2 ± 0.1	46.6 ± 0.1	30.4 ± 0.2	66.5 ± 0.1	59.8 ± 0.1	37.1 ± 0.2
FSF-ViT (Ours)	**95.1 ± 0.1**	**94.0 ± 0.0**	**84.6 ± 0.1**	**80.3 ± 0.1**	**73.1 ± 0.4**	**37.7 ± 0.7**	**54.0 ± 0.1**	**47.6 ± 0.2**	**31.3 ± 0.1**	**68.5 ± 0.0**	**63.8 ± 0.2**	**42.2 ± 0.0**

Table 2
Comparison of FSF-ViT and other models on Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172 datasets (%).

Method	Mixing type	Params (M)	Food-30			Sushi-50			ChineseFoodNet			Vireo Food-172		
			10-shot	5-shot	1-shot	10-shot	5-shot	1-shot	10-shot	5-shot	1-shot	10-shot	5-shot	1-shot
EfficientNet-L	Conv	117.27	95.2 ± 0.4	91.2 ± 0.4	68.2 ± 1.8	78.2 ± 0.4	67.6 ± 0.4	30.8 ± 1.2	48.4 ± 0.3	36.9 ± 0.2	18.0 ± 0.3	61.6 ± 0.4	51.7 ± 0.1	22.5 ± 0.6
ResNet-152	Conv	58.21	**95.9 ± 0.3**	92.1 ± 0.1	56.5 ± 2.3	76.8 ± 0.6	66.6 ± 0.3	25.2 ± 0.4	29.4 ± 0.6	23.8 ± 1.2	14.8 ± 0.2	42.9 ± 0.2	39.6 ± 5.3	20.4 ± 0.1
RegNetY-16G	Conv	80.66	91.2 ± 2.3	88.1 ± 1.4	61.9 ± 3.6	80.8 ± 0.1	68.0 ± 0.2	23.1 ± 1.3	46.0 ± 0.3	33.3 ± 0.2	12.7 ± 0.3	61.7 ± 0.1	49.4 ± 0.6	17.1 ± 0.8
ConvNeXt-B	Conv	87.54	88.9 ± 2.4	89.4 ± 2.3	68.9 ± 5.6	70.9 ± 1.7	58.2 ± 2.9	25.6 ± 2.6	34.3 ± 0.5	24.7 ± 0.6	12.8 ± 0.4	51.0 ± 0.0	41.7 ± 1.0	16.5 ± 1.3
FocalNet-B	Conv	87.94	83.6 ± 2.1	75.0 ± 4.0	13.7 ± 1.4	63.5 ± 1.0	45.7 ± 0.4	8.6 ± 0.6	39.4 ± 0.2	28.7 ± 0.0	8.6 ± 0.1	53.8 ± 0.2	43.0 ± 0.1	10.1 ± 0.3
InceptionNeXt-B	Conv	83.69	76.1 ± 0.9	73.4 ± 1.4	13.8 ± 1.1	55.4 ± 0.1	42.2 ± 1.7	8.6 ± 0.3	7.8 ± 0.3	7.2 ± 0.4	6.4 ± 0.3	22.5 ± 0.9	23.3 ± 1.3	8.4 ± 0.2
DeiT-B	Attn	86.57	93.9 ± 0.5	92.8 ± 1.9	56.5 ± 0.9	81.8 ± 0.0	69.8 ± 0.2	24.0 ± 0.4	51.0 ± 0.2	38.8 ± 0.0	16.5 ± 0.1	67.7 ± 0.1	57.6 ± 0.2	22.6 ± 0.1
PVT-Large	Attn	60.87	93.2 ± 0.9	91.4 ± 0.8	61.1 ± 0.1	**82.5 ± 0.2**	69.7 ± 0.1	24.8 ± 0.7	50.2 ± 0.5	37.5 ± 0.2	16.0 ± 0.1	64.4 ± 0.2	54.4 ± 0.2	21.8 ± 0.3
ViT-Base/16	Attn	85.67	87.2 ± 0.3	83.2 ± 0.4	64.8 ± 0.4	65.2 ± 0.0	55.5 ± 0.7	25.2 ± 0.3	48.8 ± 0.1	41.3 ± 0.2	29.1 ± 0.1	61.8 ± 0.1	54.2 ± 0.1	33.5 ± 0.5
Swim-B	Attn	86.71	88.8 ± 0.9	87.5 ± 2.2	73.9 ± 1.6	72.1 ± 0.4	60.3 ± 2.6	24.3 ± 1.3	39.0 ± 1.0	29.5 ± 0.7	13.1 ± 1.1	54.3 ± 0.9	42.2 ± 5.9	17.4 ± 3.1
T2T-ViT-24	Attn	63.50	88.5 ± 1.4	87.3 ± 0.9	46.0 ± 4.2	65.9 ± 0.2	52.4 ± 0.9	17.1 ± 0.5	37.2 ± 0.0	25.2 ± 0.2	11.0 ± 0.2	54.5 ± 0.2	41.5 ± 0.2	12.2 ± 2.2
TNT-B	Attn	64.81	86.2 ± 0.7	85.6 ± 0.2	58.3 ± 1.0	76.1 ± 0.2	62.0 ± 0.4	20.9 ± 0.1	47.1 ± 0.0	36.6 ± 0.2	14.6 ± 0.2	64.9 ± 0.2	54.0 ± 0.0	21.8 ± 0.1
BEiT-B	Attn	85.79	86.2 ± 0.3	88.1 ± 0.5	70.9 ± 0.7	70.5 ± 0.2	65.7 ± 0.7	29.9 ± 0.9	20.5 ± 0.3	22.2 ± 0.3	16.3 ± 0.1	36.2 ± 0.0	34.6 ± 0.1	25.2 ± 0.0
Hiera-B+	Attn	69.04	86.4 ± 0.5	80.4 ± 0.6	69.6 ± 0.5	57.6 ± 1.1	50.2 ± 1.0	28.6 ± 0.5	32.3 ± 0.6	28.7 ± 0.4	19.1 ± 0.0	46.9 ± 0.1	39.1 ± 1.0	23.6 ± 0.4
GC ViT-B	Attn	89.51	89.7 ± 1.3	83.9 ± 2.7	59.7 ± 6.4	66.8 ± 5.4	59.0 ± 1.6	24.3 ± 0.5	22.8 ± 1.9	27.0 ± 1.3	13.0 ± 3.6	36.6 ± 3.7	39.9 ± 3.0	19.2 ± 1.0
FSF-ViT (Ours)	Attn	85.67	95.1 ± 0.1	**94.0 ± 0.0**	**84.6 ± 0.1**	80.3 ± 0.1	**73.1 ± 0.4**	**37.7 ± 0.7**	**54.0 ± 0.1**	**47.6 ± 0.2**	**31.3 ± 0.1**	**68.5 ± 0.0**	**63.8 ± 0.2**	**42.2 ± 0.0**

explored and leveraged feature relationships between original and augmented images. We designed the following comparative experiments to evaluate the above-mentioned operations:

- OG: The baseline approach utilized only original images for training.
- OA: This method used original and augmented images for training.
- FSF-ViT: Building upon OA, our method combined features from both augmented and original images to enhance feature extraction capability.

Experimental evaluations were conducted on four datasets. Notably, augmented images used for OA were generated by our proposed CutCenter and CornerMix methods. As shown in Table 1, both OA and FSF-ViT outperformed the baseline OG, with FSF-ViT achieving the best performance across all datasets. Compared to OA, FSF-ViT achieved average accuracy improvements of 8.1%, 9.4%, 0.9%, and 3.7% on Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172, respectively. These experimental results validated the following: (1) the complementary relationship between features from original and those from augmented images, and (2) the effectiveness of FSF-ViT in feature fusion for improving the performance of few-shot food image classification.

3.2. Comparison with the state-of-the-art methods

We evaluated our method against state-of-the-art self-supervised classification approaches, which can be categorized into two types: CNNs and attention networks. The CNNs-based models include InceptionNeXt (Yu et al., 2024), ConvNeXt (Liu et al., 2022), FocalNet (Yang et al., 2022), EfficientNet (Tan & Le, 2021), RegNet (Radosavovic et al., 2020), and ResNet (He et al., 2016), while the attention networks comprise GC ViT (Hatamizadeh et al., 2023), Hiera (Ryali et al., 2023), BeiT (Bao et al., 2021), Swin Transformer (Liu et al., 2021), ViT (Dosovitskiy et al., 2021), DeiT (Touvron et al., 2021), T2T-ViT (Yuan et al., 2021), TNT (Han et al., 2021), and PVT (Wang et al., 2021). As shown in Table 2, our method achieved superior performance on all four datasets, with two exceptions. In the 10-shot experiment on the Food-30 dataset, ResNet-152 achieved the best Top-1 accuracy, while our method ranked third with only a 0.8% lower accuracy. In the 10-shot experiments on the Sushi-50 dataset, PVT achieved the highest accuracy, outperforming our method by 2.2%. Across all experiments, compared to the best results among other methods, our method demonstrated significant performance improvement, with an average Top-1 accuracy increase of 3.7%, 2.7%, 3.8%, and 5.2% on Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172, respectively. Notably, FSF-ViT demonstrated increasingly superior performance relative to other approaches as the number of training samples decreased. FSF-ViT showed remarkable performance on the Food-30 dataset, achieving a 10.7% improvement over the second-best method in the 1-shot setting, validating its effectiveness for few-shot food image classification.

3.3. Qualitative evaluation

We further validated the effectiveness of FSF-ViT by showing representative cases. Fig. 5 showed comparative examples for FSF-ViT and the baseline, ViT. The baseline misclassified some challenging samples, such as images with small-scale target objects or peripheral interference. However, FSF-ViT achieved accurate classification through effectively magnifying detailed areas and erasing corner regions. These results further validated that the image augmentation module was capable of enhancing feature representations, which consequently improved the generalization ability of FSF-ViT.

3.4. Ablation experiments

Ablation experiments are systematic studies to evaluate the contribution of each component in our proposed model. By selectively removing or replacing specific components while keeping others unchanged, we can quantitatively assess how each component affects the model's overall performance.

Effectiveness of the image augmentation module: We evaluated the effectiveness of CutCenter and CornerMix methods in the image augmentation module through ablation experiments. We conducted experiments across four datasets: Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172. The experimental results based on ViT were presented in Table A.1. Applied independently, the CutCenter method achieved average Top-1 accuracy improvements of 11.6%, 14.8%, 3.7%, and 6.9% on these datasets, while the CornerMix method yielded improvements of 9.5%, 11.2%, 2.3%, and 6.2%, respectively.

Fig. 4. Some samples of the Food-30 dataset, one sample is shown for each class.

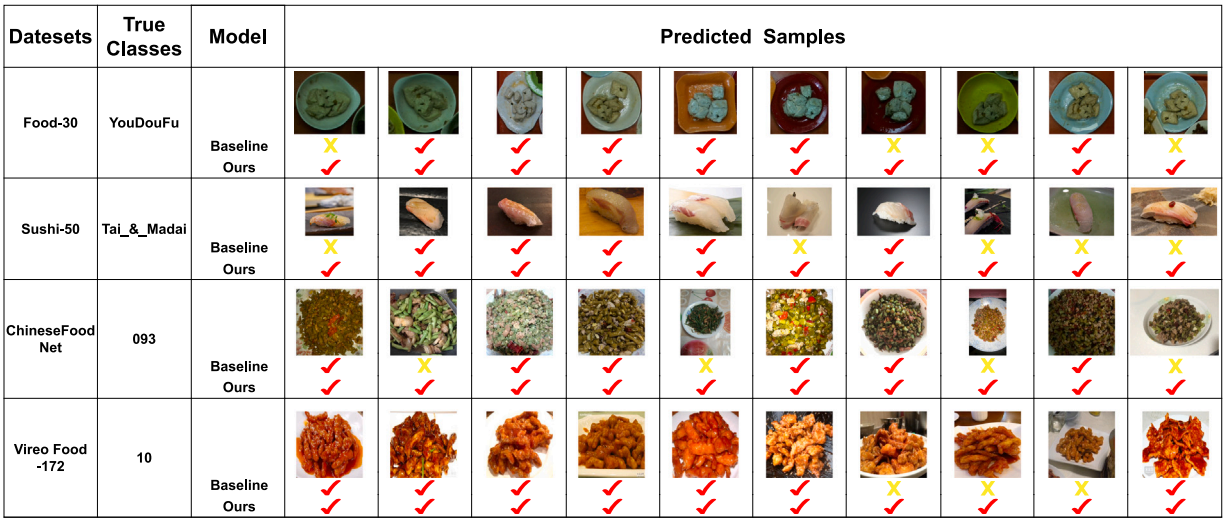

Fig. 5. Predicted results of the baseline and our proposed FSF-ViT from the Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172 datasets. We use the Vision Transformer as the baseline.

These significant performance improvements demonstrated that both methods enhanced discriminative feature learning through magnifying detailed regions and randomly erasing corner regions where non-target objects may appear. The remarkable ability of the CutCenter method validated its effectiveness in capturing crucial detail information for distinguishing subtle inter-class differences. By combining the two methods, our method achieved optimal performance across all datasets, yielding average Top-1 accuracy improvements of 12.8%, 15.1%, 4.6%, and 8.3%.

Effect of the different loss functions: Using FSF-ViT as the baseline, we conducted ablation studies on four datasets to analyze the effects of various loss functions. Fig. A.1 showed two loss functions: L_{CCF}, integrating the cross-entropy and pairwise confusion losses for fused features, and L_{VF}, which, based on L_{CCF}, adds both losses for pre-fusion features. Experimental results showed that L_{VF} improved the average Top-1 accuracy by 1.6%, 1.8%, 0.6%, and 0.8% on Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172, respectively. These experimental results demonstrated that L_{VF} could strengthen the final fused class feature representations by supervising pre-fusion feature learning.

Different weighting methods: We conducted ablation experiments to evaluate the effect of different weighting methods on the four datasets: Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172. The experimental results of FSF-ViT with different weighting approaches were presented in Table A.2. The Uniform Fixed Coefficient (UFC) method employed uniform weighting with fixed coefficients of 1. The Learnable Parameter Coefficient (LPC) method introduced three learnable parameters as weight coefficients, initialized to 1 with a sum constraint of 3. Table A.2 showed that our method yielded the best results in the majority of cases with only two instances of second-best performance. These experimental results demonstrated that our adaptive weighting method effectively learned complementary feature relationships, and thus enhanced the feature learning capability of FSF-ViT.

3.5. Parameter sensitivity and visualization

To further confirm the feature clustering capability of our method, we visualized the feature distributions of ViT and FSF-ViT with t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten & Hinton, 2008) on the four datasets. Compared with ViT, our method achieved more compact intra-class aggregations and more distinct inter-class separations in Fig. A.2. These results indicated that FSF-ViT could learn more subtle features, which formed better-separated clusters for different categories, facilitating the few-shot food image classification.

We conducted a three-dimensional statistical graph of Top-1 accuracy by adjusting the value of p, as described in Section 2.2.3. The statistical results obtained from the Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172 datasets were shown in Fig. A.3. The results clearly demonstrated that FSF-ViT achieved excellent performance with p in the range of [3,8]. Based on these experimental results, we selected $p = 5$ on these four datasets, which achieved the best Top-1 accuracy performance with the proposed FSF-ViT.

The attention maps of both the baseline and FSF-ViT were visualized using Grad-CAM. Fig. A.4(a) showed that FSF-ViT captured richer food information. FSF-ViT covered more comprehensive areas of food, whereas ViT only covered partial regions, neglecting some food details. Fig. A.4(b) demonstrated that FSF-ViT exhibited more prominent attention to discriminative characteristics. Compared with ViT, FSF-ViT emphasized the textural features of a cauliflower, facilitating differentiation between it and other categories with similar appearances. Fig. A.4(c) and Fig. A.4(d) showed that FSF-ViT effectively mined category-specific regions and mitigated the influence of interfering food items. However, ViT predominantly focused on the edge regions of the plate in Fig. A.4(c) and attended regions containing extraneous food in Fig. A.4(d). FSF-ViT maintained attention solely on regions containing the target food. In summary, FSF-ViT improved classification performance through more comprehensive and accurate attention coverage.

3.6. Experimental evaluation on other types of images

To further evaluate the robustness and versatility of our proposed method beyond food images, we conducted experiments on the Flower102 dataset (Nilsback & Zisserman, 2008). This dataset comprises 102 different flower species. We divided the dataset into 1020 training images and 6149 testing images. The images were captured under various resolutions, lighting conditions, and environments. The dataset presents unique challenges in distinguishing between flower species with subtle inter-class differences, such as similar petal arrangements, color patterns, and structural features, making it an excellent benchmark for evaluating feature extraction techniques and attention mechanisms.

As shown in Fig. A.5, our method outperformed other state-of-the-art approaches on the Flower102 dataset. Specifically, it achieved a peak accuracy of 99%, and maintained high performance (above 94%) across all experimental settings (1-shot, 5-shot, and 10-shot). These results demonstrated that our method was effective not only for food images but also for the classification of other types of images. The superior performance could be attributed to our method's enhanced ability in fine-grained feature extraction and representation learning.

4. Discussion

4.1. Challenges in few-shot food image classification

The challenges of few-shot food image classification are two-fold: First, food images belong to fine-grained visual data (Fu et al., 2017; Zhang et al., 2014). The subtle differences between food categories pose substantial classification challenges. Second, the distinguishing characteristics between food categories are primarily found in subtle local regions. However, global image representations often overlook these critical local features that are essential for food classification.

To address these challenges, we proposed a Vision Transformer-based few-shot food image classification method with image augmentation and adaptive global-local feature fusion. Specifically, the image augmentation method extracted and magnified discriminative local regions for enhanced fine-grained feature learning, while the adaptive weighting method facilitated the fusion of global-local representations.

4.2. Advantages of ViT in food image classification

Transformer leverages self-attention mechanisms to focus on salient regions instead of processing all features uniformly (Han et al., 2023). This mechanism directs attention to discriminative local regions, enhancing fine-grained feature extraction for food image classification. Experimental results demonstrated that Transformer-based approaches were generally superior to CNN-based methods, with FSF-ViT exhibiting superior performance among Transformer variants.

4.3. Limitations and future improvement directions

We proposed FSF-ViT for few-shot food image classification. Despite its superior classification performance, the model exhibited several limitations. Our method was primarily designed for Chinese cuisine, which may limit the applicability of the model to cuisines from other regions. Future work needs to incorporate diverse culinary categories to enhance model applicability across different regional cuisines. We aim to develop a comprehensive open-source food image dataset to facilitate future research in the deep learning-based food industry. Classification accuracy decreases with increasing category numbers, necessitating future research on advanced feature extraction techniques. For example, mining rich food-specific knowledge, especially ingredient information, improved classification performance (Liu et al., 2024; Luo et al., 2023). Serving as general and intermediate food attributes, ingredient-oriented features provided complementary information to category-oriented features, thereby enhancing feature learning (Jiang et al., 2020). Moreover, multi-scale algorithms can be integrated (Jing et al., 2023; Zhang et al., 2023). Compared to single-scale representation, multi-scale approaches enhance global feature representation, as larger scales with wider receptive fields provide richer information (Ren et al., 2024). In conclusion, future research directions focus on model optimization to enhance food recognition accuracy, thereby strengthening intelligent food applications. Such advancements can promote effective dietary management and health.

5. Conclusions

In this paper, we proposed a ViT-based deep learning method combining image augmentation and adaptive global-local feature fusion for few-shot food image classification. This method focused on training with limited samples, in which each category contained only 10, 5, or 1 food images, making our approach more economical in time and cost. We constructed a small dataset by collecting real-world images of 30 Chinese food categories. We performed experiments on our dataset and three benchmark food datasets to validate the effectiveness of the proposed method. The results demonstrated that FSF-ViT outperformed mainstream deep learning classification models and achieved the highest classification accuracy of 95.1% on the test set. Compared with ViT, FSF-ViT showed a significant performance improvement, with average accuracy improvements of 12.8%, 15.1%, 4.6%, and 8.3% on the Food-30, Sushi-50, ChineseFoodNet, and Vireo Food-172 datasets, respectively. In summary, we proposed a novel method for recognizing the category of food displayed in an image. This method provided low-cost and effective technical support for online dietary recording, facilitating dietary management and health.

CRediT authorship contribution statement

Jinhong Li: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Huiying Xu:** Writing – review & editing, Project administration, Methodology, Conceptualization. **Xinzhong Zhu:** Writing – review & editing, Validation, Software, Data curation. **Jiping Xiong:** Writing – review & editing, Supervision. **Xiaolei Zhang:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.foodchem.2025.145276.

Data availability

I have shared the link to my data in the Abstract.

References

Bao, H., Dong, L., Piao, S., & Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Chen, J., & Ngo, C. W. (2016). Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on multimedia* (pp. 32–41). http://dx.doi.org/10.1145/2964284.2964315.

Chen, X., Zhou, H., Zhu, Y., & Diao, L. (2017). ChineseFoodNet: A large-scale image dataset for Chinese food recognition. arXiv preprint arXiv:1705.02743.

Das, P., B., A. A., C., N. P., Katyal, M., Kesavan, R. k., Rustagi, S., Panda, S. K., Nayak, P. K., & Mohanta, Y. K. (2025). Recent advances on artificial intelligence-based approaches for food adulteration and fraud detection in the food industry: Challenges and opportunities. *Food Chemistry*, *468*, Article 142439. http://dx.doi.org/10.1016/j.foodchem.2024.142439.

de Oliveira, A. N., Bolognini, S. R. F., Navarro, L. C., Delafiori, J., Sales, G. M., de Oliveira, D. N., & Catharino, R. R. (2023). Tomato classification using mass spectrometry-machine learning technique: A food safety-enhancing platform. *Food Chemistry*, *398*, Article 133870. http://dx.doi.org/10.1016/j.foodchem.2022.133870.

Deng, Z., Fu, J., Yang, M., Zhang, W., Yun, Y. H., & Zhang, L. (2024). Geographical origin identification of hainan camellia oil based on fatty acid composition and near infrared spectroscopy combined with chemometrics. *Journal of Food Composition and Analysis*, *125*, Article 105730. http://dx.doi.org/10.1016/j.jfca.2023.105730.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International conference on learning representations*.

Feng, Y., Li, X., Zhang, Y., & Xie, T. (2023). Detection of atlantic salmon residues based on computer vision. *Journal of Food Engineering*, *358*, Article 111658.

Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Gao, X., Xiao, Z., & Deng, Z. (2024). High accuracy food image classification via vision transformer with data augmentation and feature augmentation. *Journal of Food Engineering*, *365*, Article 111833.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 87–110. http://dx.doi.org/10.1109/TPAMI.2022.3152247.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*, 15908–15919.

Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., & Molchanov, P. (2023). Global context vision transformers. In *2023 International conference on machine learning* (pp. 12633–12646).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, G., Zhang, E., Zhou, J., Zhao, J., Gao, Z., Sugirbay, A., Jin, H., Zhang, S., & Chen, J. (2021). Infield apple detection and grading based on multi-feature fusion. *Horticulturae*, *7*, 276.

Jiang, S., Min, W., Lyu, Y., & Liu, L. (2020). Few-shot food recognition via multi-view representation learning. *ACM Transactions on Multimedia Computing Communications and Application (TOMM)*, *16*, 1–20.

Jing, J., Gao, T., Zhang, W., Gao, Y., & Sun, C. (2023). Image feature information extraction for interest point detection: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 4694–4712. http://dx.doi.org/10.1109/TPAMI.2022.3201185.

Kaushal, S., Tammineni, D. K., Rana, P., Sharma, M., Sridhar, K., & Chen, H. H. (2024). Computer vision and deep learning-based approaches for detection of food nutrients/nutrition: New insights and advances. *Trends in Food Science & Technology*, *146*, Article 104408. http://dx.doi.org/10.1016/j.tifs.2024.104408.

Key, T. J., Bradbury, K. E., Perez-Cornago, A., Sinha, R., Tsilidis, K. K., & Tsugane, S. (2020). Diet, nutrition, and cancer risk: what do we know and what is the way forward? *BMJ*, *368*, http://dx.doi.org/10.1136/bmj.m511.

Konstantakopoulos, F. S., Georga, E. I., & Fotiadis, D. I. (2023). An automated image-based dietary assessment system for mediterranean foods. *IEEE Open Journal of Engineering in Medicine and Biology*, *4*, 45–54. http://dx.doi.org/10.1109/OJEMB.2023.3266135.

Konstantakopoulos, F. S., Georga, E. I., & Fotiadis, D. I. (2024). A review of image-based food recognition and volume estimation artificial intelligence systems. *IEEE Reviews in Biomedical Engineering*, *17*, 136–152. http://dx.doi.org/10.1109/RBME.2023.3283149.

Liu, P. (2019). Research on food image recognition based on ResNet. *Electronic Technology & Software Engineering*, *16*, 64–67.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 11966–11976).

Liu, Y., Min, W., Jiang, S., & Rui, Y. (2024). Convolution-enhanced bi-branch adaptive transformer with cross-task interaction for food category and ingredient recognition. *IEEE Transactions on Image Processing*, *33*, 2572–2586. http://dx.doi.org/10.1109/TIP.2024.3374211.

Luo, M., Min, W., Wang, Z., Song, J., & Jiang, S. (2023). Ingredient prediction via context learning network with class-adaptive asymmetric loss. *IEEE Transactions on Image Processing*, *32*, 5509–5523. http://dx.doi.org/10.1109/TIP.2023.3318958.

Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., & Jiang, S. (2023). Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 9932–9949. http://dx.doi.org/10.1109/TPAMI.2023.3237871.

Minho, L. A. C., de Lima Conceição, J., Barboza, O. M., de Freitas Santos Junior, A., & dos Santos, W. N. L. (2025). Robust DEEP heterogeneous ensemble and META-learning for honey authentication. *Food Chemistry*, *482*, Article 144001. http://dx.doi.org/10.1016/j.foodchem.2025.144001.

Nadeem, M., Shen, H., Choy, L., & Barakat, J. M. H. (2023). Smart diet diary: Real-time mobile application for food recognition. *Applied System Innovation*, *6*, http://dx.doi.org/10.3390/asi6020053.

Nath, P. C., Mishra, A. K., Sharma, R., Bhunia, B., Mishra, B., Tiwari, A., Nayak, P. K., Sharma, M., Bhuyan, T., Kaushal, S., Mohanta, Y. K., & Sridhar, K. (2024). Recent advances in artificial intelligence towards the sustainable future of agri-food industry. *Food Chemistry*, *447*, Article 138945.

Nilsback, M. E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing* (pp. 722–729). http://dx.doi.org/10.1109/ICVGIP.2008.47.

Qiu, J., Lo, F. P. W., Sun, Y., Wang, S., & Lo, B. (2019). Mining discriminative food regions for accurate food recognition. In *British machine vision conference*.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 10425–10433).

Ren, J., Li, C., An, Y., Zhang, W., & Sun, C. (2024). Few-shot fine-grained image classification: A comprehensive review 5. (pp. 405–425). http://dx.doi.org/10.3390/ai5010020.

Ryali, C., Hu, Y. T., Bolya, D., Wei, C., Fan, H., Huang, P. Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., & Feichtenhofer, C. (2023). Hiera: A hierarchical vision transformer without the bells-and-whistles. In *2023 international conference on machine learning* (pp. 29441–29454).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE international conference on computer vision* (pp. 618–626).

Shao, W., Min, W., Hou, S., Luo, M., Li, T., Zheng, Y., & Jiang, S. (2023). Vision-based food nutrition estimation via RGB-d fusion network. *Food Chemistry*, *424*, Article 136309.

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, *55*, 40. http://dx.doi.org/10.1145/3582688.

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *2021 International conference on machine learning* (pp. 10096–10106).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *2021 International conference on machine learning* (pp. 10347–10357).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

VijayaKumari, G., Vutkur, P., & P., V. (2022). Food classification using transfer learning technique. *Global Transitions Proceedings*, *3*, 225–229. http://dx.doi.org/10.1016/j.gltp.2022.03.027.

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF international conference on computer vision* (pp. 548–558).

Xiao, Z., Diao, G., & Deng, Z. (2024). Fine grained food image recognition based on swin transformer. *Journal of Food Engineering*, *380*, Article 112134. http://dx.doi.org/10.1016/j.jfoodeng.2024.112134.

Xiao, Z., Ling, R., & Deng, Z. (2025). FoodCSWin: A high-accuracy food image recognition model for dietary assessment. *Journal of Food Composition and Analysis*, *139*, Article 107110. http://dx.doi.org/10.1016/j.jfca.2024.107110.

Xiao, T., Xie, C., Yang, L., He, X., Wang, L., Zhang, D., Cui, T., Zhang, K., Li, H., & Dong, J. (2025). A general deep learning model for predicting and classifying pea protein content via visible and near-infrared spectroscopy. *Food Chemistry*, *478*, Article 143617. http://dx.doi.org/10.1016/j.foodchem.2025.143617.

Xu, S. L., Zhang, F., Wei, X. S., & Wang, J. (2022). Dual attention networks for few-shot fine-grained recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 2911–2919. http://dx.doi.org/10.1609/aaai.v36i3.20196.

Yang, X., Ho, C. T., Gao, X., Chen, N., Chen, F., Zhu, Y., & Zhang, X. (2025). Machine learning: An effective tool for monitoring and ensuring food safety, quality, and nutrition. *Food Chemistry*, *477*, Article 143391. http://dx.doi.org/10.1016/j.foodchem.2025.143391.

Yang, J., Li, C., Dai, X., & Gao, J. (2022). Focal modulation networks. *Advances in Neural Information Processing Systems*, *35*, 4203–4217.

Yu, W., Zhou, P., Yan, S., & Wang, X. (2024). Inceptionnext: When inception meets convnext. In *2024 IEEE/CVF conference on computer vision and pattern recognition* (pp. 5672–5683).

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., Tay, F. E. H., Feng, J., & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *2021 IEEE/CVF international conference on computer vision* (pp. 538–547).

Zhang, N., Donahue, J., Girshick, R., & Darrell, T. (2014). Part-based R-CNNs for fine-grained category detection. In *Computer vision – ECCV 2014* (pp. 834–849).

Zhang, W., Sun, C., & Gao, Y. (2023). Image intensity variation information for interest point detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 9883–9894. http://dx.doi.org/10.1109/TPAMI.2023.3240129.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 34*, (pp. 13001–13008).