DPFA-Net: A Lightweight Hybrid Neural Network with Dual Path Feature Aggregation for Food Image Recognition

Xiangyi Zhu[1†], Wenli Zhang[1†], Yingnan Sheng[1†], Congrui Lv[1†], Guorui Sheng[1*], Weiqing Min[2,3], Shuqiang Jiang[2,3]

[1*]School of Computer and Artificial Intelligence, Ludong University, Yantai, 264025, Shandong, China.
[2]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China.
[3]University of Chinese Academy of Sciences, Beijing, 100049, China.

*Corresponding author(s). E-mail(s): shengguorui@ldu.edu.cn;
Contributing authors: zhuxiangyi@m.ldu.edu.cn;
zhangwenli@m.ldu.edu.cn; 13964313827@163.com;
lvcongrui@m.ldu.edu.cn; minweiqing@ict.ac.cn; sqjiang@ict.ac.cn;
†These authors contributed equally to this work.

Abstract

Food image recognition holds significant application potential in the field of computer vision. However, due to performance constraints on mobile devices, the scale and computational overhead of models face notable limitations, making effective deployment on mobile platforms challenging. To address this issue, this paper proposes a lightweight Dual-Path Feature Aggregation Network (DPFA-Net), designed to enhance the performance of food recognition tasks through efficient local and global feature extraction strategies. Specifically, the DPFA architecture comprises two core modules: the GhostBottleneck module for local feature encoding and the Position Mamba Vision Transformer (PM-ViT) module for global modeling. In this work, the GhostBottleneck module is utilized to extract local features from images. Furthermore, by integrating the Mamba structure with the Separable Self-Attention (SSA) structure, we construct the Mamba Attention (MA) module, which replaces the traditional Attention mechanism in Vision Transformers to build the PM-ViT module, enabling the capture

of global features in food images. The redesigned DPFA-Net effectively fuses local and global information, achieving efficient food image recognition. The experiments were conducted on the ETHZ Food-101, Vireo Food-172, and UEC Food-256 datasets. The results show that, while reducing the number of parameters, DPFA-Net achieved Top-1 accuracies of 91.46%, 91.59%, and 75.33%, respectively, representing a 1.50%–3.9% improvement over MobileViTv2. Compared to MobileViTv2, DPFA-Net improves performance by 1.50%-3.9%, fully validating the effectiveness and superiority of the DPFA architecture.

Keywords: Food Image Recognition, Food Computing, Lightweight, Vision Transformer

1 Introduction

With the rapid advancement of digital technologies, food computing has been increasingly applied in food science and nutritional health, gradually reshaping the entire industry ecosystem [1]. Its applications span the entire food chain, from production and processing to end-user consumption, offering new technical means and service models for dietary health management [2–5]. Among these developments, food image recognition has emerged as a core technology [6–8], playing a pivotal role in advancing intelligent food systems. It is widely utilized in scenarios such as precise dietary health monitoring, intelligent upgrades and quality control in the food industry, as well as collaborative supply chain management and risk prevention.As food and nutrition domains continue to evolve, extending the application boundaries of food recognition technologies becomes increasingly important. Mobile devices, due to their portability and widespread adoption, have become a key platform for deploying food image recognition applications. Migrating high-accuracy food recognition models to mobile platforms is expected to facilitate personal dietary management, industrial digital transformation, and supply chain optimization.However, most state-of-the-art food recognition models are typically characterized by large model sizes and high computational demands, making them unsuitable for resource-constrained mobile environments. Therefore, achieving a balance between model accuracy and lightweight efficiency has become a pressing research challenge. To address this issue, we propose a novel lightweight Dual-Path Feature Aggregation Network (DPFA-Net), specifically designed for efficient deployment of food image recognition tasks on mobile devices with limited resources.

While considerable progress has been made in lightweight food image recognition, existing methods still struggle to meet the practical demands of mobile deployment. As illustrated in Fig. 1, food images present a unique challenge: they often exhibit high intra-class variability and low inter-class discrimination. Specifically, instances within the same food category can differ significantly due to variations in ingredients, shapes, or cooking methods (see the left side of Fig. 1). For example, within the category of 'waffle', different samples use various ingredients and display various shapes, making key distinguishing features more dependent on the global context rather than local

regions. However, different food categories can share similar textures or components (see the right side of Fig. 1). For example, dishes such as sea intestine fried rice, stir-fried sea intestine, and minced pork with sea intestine belong to different categories, but exhibit substantial visual similarities in texture and ingredients. These characteristics further complicate the recognition task, requiring the model to capture long-range dependencies and cross-category similarities effectively.However, traditional convolutional neural networks (CNNs), constrained by limited receptive fields, often fail to model such long-range interactions. Deepening the network to overcome this issue typically leads to increased parameter counts and computational complexity, which contradicts the goal of lightweight design.Recently, Vision Transformers (ViTs), leveraging attention mechanisms, have shown strong potential in global feature modeling. However, their computational complexity grows quadratically with input resolution, making them less suitable for lightweight scenarios due to training difficulties and inference costs. The emerging Vision Mamba (Vim) architecture provides a promising alternative by achieving superior global modeling capability with significantly lower computational overhead, offering new possibilities for efficient visual recognition.In contrast, while CNNs are less effective at capturing fine-grained dependencies and long-range interactions, they excel at extracting local features. Therefore, an open challenge in the recognition of lightweight food is how to efficiently integrate local and global features within a resource-constrained framework.

Lightweight food image recognition faces two fundamental challenges. First, food images often exhibit fine-grained characteristics [9] due to the subtle visual differences among diverse ingredients. These intra-image variations are difficult to distinguish using coarse appearance cues, requiring models to possess refined visual discrimination capabilities. Second, the key discriminative information in food images is frequently scattered across multiple non-contiguous regions, demanding strong global modeling abilities.Traditional convolutional neural networks (CNNs) [10–13] extract low-level visual features through local receptive fields, but they rely on deep stacking or dilated convolutions to expand the receptive field—approaches that are inherently constrained in lightweight designs. On one hand, deeper network architectures [14, 15] significantly increase the model's parameter count and computational overhead. On the other hand, commonly used strategies for reducing resource consumption, such as channel pruning and spatial downsampling, often result in the loss of long-range dependencies.This conflict between the need for global feature modeling and the constraints of lightweight deployment has emerged as a core bottleneck in improving recognition performance using CNN-based architectures. Although Vision Transformers (ViTs) [16] demonstrate strong capability in capturing long-range relationships between pixels, their self-attention mechanism incurs quadratic computational complexity with respect to input resolution, making them heavily dependent on large-scale annotated datasets and high-performance hardware during training. Moreover, the high-dimensional matrix operations involved in ViTs conflict with the low power and limited computing capacities of edge devices, hindering the balance between accuracy and efficiency in lightweight deployment.Recently, Vision Mamba (Vim) has achieved promising performance on several visual tasks, often surpassing ViT. However, practical applications remain limited. As noted by Weihao Yu, Dongchen Han, and others

[17], the state space model (SSM) at the core of the Mamba block may not be inherently suitable for vision tasks; rather, the architectural design itself is the key contributor to its effectiveness.Thus, a pressing issue in food image recognition is how to construct a lightweight architecture that simultaneously supports efficient training on server-side platforms and fast inference on mobile devices, all while maintaining high recognition accuracy.

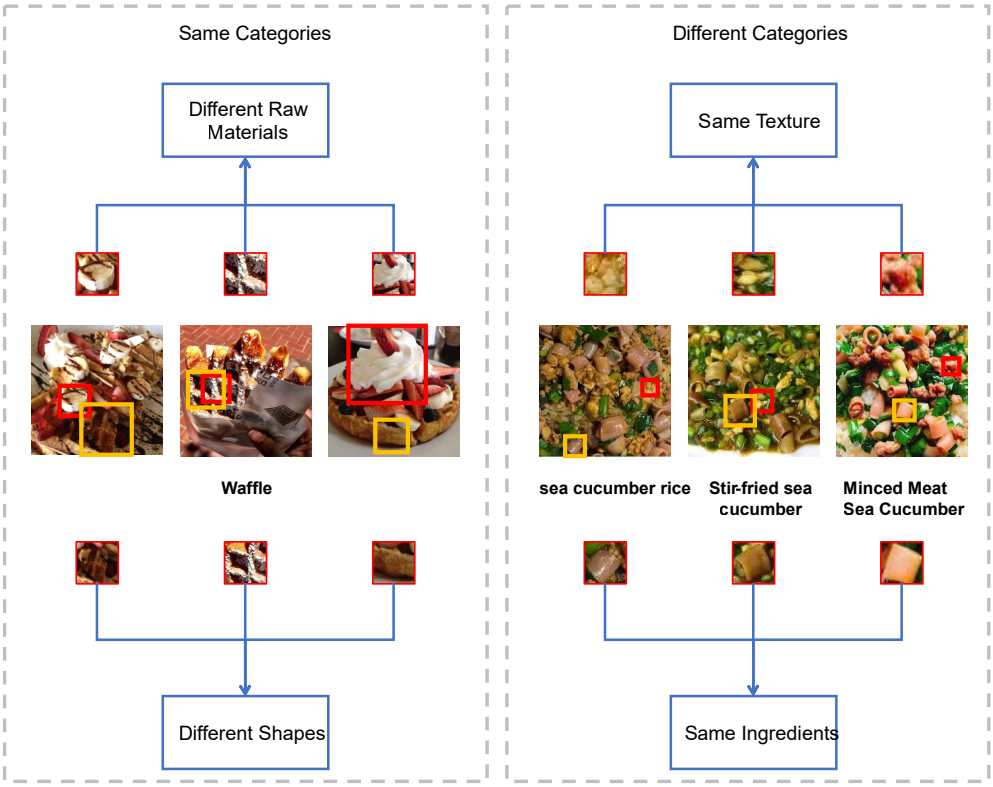

Fig. 1 Example images from the ETHZ Food-101 and Vireo Food-172 datasets. The left side illustrates the high intra-class variability within the same category, while the right side highlights the low inter-class distinction across different categories.

This paper aims to address two major challenges in lightweight food image recognition: the difficulty of jointly modeling local and global features, and the limitations of model deployment under constrained computational resources. Food images typically exhibit large intra-class variations and small inter-class differences- foods of thesame category may vary significantly in appearance due to differences in cooking style, shape, and viewing angle, while foods from different categories often share highly similar colors and textures. These characteristics impose higher demands on

the model's representation capability, requiring it to be both sensitive to fine-grained local textures and capable of capturing global semantic structures.

To this end, we propose DPFA-Net, which achieves efficient fusion of local and global information while maintaining low parameter count and computational cost, thereby significantly improving recognition performance in resource-constrained environments such as mobile devices. The main contributions of this work are summarized as follows:

1. The proposed DPFA-Net achieves significant reductions in both parameter count and computational cost, enabling efficient deployment on mobile devices with limited resources. To maintain strong global feature modeling capabilities while ensuring lightweight design, we redesigned the Position Mamba Vision Transformer (PM-ViT) module and connected it in series with the GhostBottleneck to form the Dual-Path Feature Aggregation (DPFA) Block. This block employs a sequential feature propagation mechanism, allowing the PM-ViT to further aggregate global information on top of the local features extracted by the convolutional branch. As a result, it naturally fuses fine-grained local textures with global semantic information. This design not only enhances the model's discriminative power and robustness but also achieves an excellent balance between lightweight efficiency and recognition performance.
2. In terms of local feature extraction, we introduce the GhostBottleneck module to capture detailed texture information in food images. By employing lightweight convolutional operations, this module efficiently extracts spatially localized features, highlighting critical structural details while reducing redundant computations. This provides a strong local representation foundation for subsequent global modeling.
3. For global feature modeling, we combine Structured Separable Self-Attention (SSA) with the Mamba architecture to construct the Mamba Attention (MA) module. Traditional SSA computes correlations only among pixels at the same relative positions within each patch, which limits its ability to model long-distance dependencies across patches. To overcome this limitation, we design a Local-Relation (LR) Block that computes correlations not only within individual patches but also across different patches, thereby enhancing contextual interaction and significantly improving global modeling capacity.
4. Extensive experiments are conducted on three representative public food image datasets. The results demonstrate that our proposed DPFA-Net not only achieves lower parameter count and computational cost but also surpasses current state of-the-art lightweight and hybrid models in recognition accuracy, validating its superiority and deployability in practical food recognition applications.

2 Related work

2.1 Lightweight CNNs, ViTs, Mamba and Hybrid Models

Among numerous well-known CNNs, ResNet [18] is undoubtedly one of the most acclaimed architectures. However, higher-accuracy CNN models often come with large parameter counts and high FLOPs. To address this challenge, a series of

lightweight CNN models, such as ShuffleNetV2 [19], ESPNetV2 [20], EfficientNet [21], MobileNetV3 [22], and MobileNetV4 [23], have emerged, significantly reducing parameter counts and computational complexity while maintaining competitive performance. Specifically designed for resource-constrained environments, MobileNetV3 is optimized primarily for mobile CPU, while MobileNetV4 further extends optimization to mobile CPU, DSP, GPU, as well as Apple's Neural Engine and Google's Pixel EdgeTPU. Nevertheless, lightweight CNN-based architectures typically struggle to effectively capture global features.

To address this limitation, ViTs draw inspiration from the successful application of transformers in natural language processing, introducing them to image recognition tasks to develop a novel approach for capturing global information. However, the high parameter count and computational demands of ViTs have shifted research focus toward optimizing self-attention mechanisms to improve efficiency. Notable efforts include Swin Transformer [24], EfficientFormer [25], LightViT [26], EfficientViT [27], MiniViT [28], and TinyViT [29]. Nevertheless, these transformer-based lightweight models face challenges such as complex training processes and high computational costs due to the quadratic growth of token interactions. HAFormer [30] proposes an Efficient Transformer module that simplifies the quadratic computations of traditional transformers through spatial reduction, linear projection, and segmentation strategies. While this module reduces computational overhead to some extent, it still incurs significant resource consumption and time costs when processing high-resolution or large-scale images.

Recent research trends indicate that constructing lightweight hybrid systems by combining the strengths of CNNs and Vision Transformers can enhance prediction accuracy and training stability in mobile vision tasks. Representative works in this area include MobileFormer [31], CMT [32], CvT [33], BoTNet [34], Next-ViT [35], EdgeViTs [36], as well as MobileViT v1 [37] and v2 [38]. These hybrid architectures successfully integrate the advantages of local and global information processing, although they still face challenges related to large model sizes in certain scenarios. Additionally, models based on State Space Models (SSMs) have been proposed, such as Vision Mamba [39], VMamba [40], and MambaVision [41], which partially address the computational and memory efficiency issues of ViTs when handling high-resolution images. The emergence of MambaOut [42] and MILA [43] further demonstrates that the effectiveness of Mamba-series models lies in their unique architectural structure rather than the SSM itself.

2.2 Lightweight Food Recognition

In recent years, with the deepening of research in the field of food computing, a series of deep learning-based food recognition methods have emerged. Due to the growing demand for lightweight food image recognition in practical application scenarios, researchers have begun to explore ways to reduce model complexity and computational costs while maintaining performance. Sheng et al. [44] conducted an in-depth investigation into the critical issue of making models lightweight while efficiently extracting features. They designed a model named LP-ViT, which retains positional information during the extraction of global features. By integrating an inverted residual structure,

LP-ViT achieves efficient fusion of global and local features, improving model accuracy while maintaining a lightweight design.

To overcome the limitations of traditional CNNs in capturing global information, Sheng et al. [45] proposed GSNet, which employs a global shuffle operation to enable convolutions to extract global features from images, thereby improving food image recognition accuracy while maintaining a lightweight model. Yang et al. [46] introduced AFNet, which utilizes aggregation blocks for global feature encoding and integrates residual models to effectively capture both global and local features. This approach enhances recognition accuracy while reducing parameter count and computational load. However, compared to ViTs, convolutional operations still fall short in comprehensively capturing global information.

3 Method

3.1 Overview of DPFA-Net

DPFA-Net is a hybrid architecture that integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Food images typically exhibit highly complex and diverse visual characteristics: significant intra-class variations may exist within the same food category, while visually similar appearances are often observed across different food categories. Therefore, effectively capturing both fine-grained local features and global contextual information is crucial for enhancing the model's representation capability.

To address this challenge, DPFA-Net adopts a dual-path structure within each fundamental block. This design aims to achieve collaborative modeling of local textures and global semantics in food images. A single-path architecture often introduces representational bias: convolution-based branches tend to overlook global dependencies, while attention-based mechanisms may weaken local details. The dual-path structure integrates the local branch and the global branch in a complementary manner, enabling the model to maintain lightweight efficiency while balancing local discriminability and global consistency. The network utilizes the GhostBottleneck module to efficiently extract fine-grained local features from food images. This module applies lightweight convolutional operations that significantly reduce computational costs while preserving critical spatial information. Concurrently, we introduce the Position Mamba Vision Transformer (PM-ViT) module to capture long-range spatial dependencies and global structure within the image. The PM-ViT is built upon the Mamba backbone, incorporating structured state-space modeling with efficient attention computation.

By sequentially stacking multiple DPFA Blocks—each composed of a GhostBottleneck followed by a PM-ViT—DPFA-Net achieves progressive fusion of local and global features across all network stages, as illustrated in Fig. 2. The DPFA Block connects the GhostBottleneck and PM-ViT modules in a serial rather than parallel manner. This design facilitates the progressive fusion of local features and global information, enhancing the interaction and representation between the two types of features. In a parallel configuration, the convolutional branch first performs convolution operations on the input feature to extract local texture information, while the

PM-ViT branch simultaneously computes global correlations on the same input, effectively introducing global dependencies directly on the original feature map. Since these two branches operate in different feature spaces, the local features produced by the convolutional branch and the global representations generated by the PM-ViT branch exhibit significant heterogeneity. As a result, their fusion may lead to inconsistent feature distributions, thereby weakening the model's discriminative ability and increasing both parameter count and computational cost.In contrast, the serial connection allows the PM-ViT module to model global dependencies on top of the features refined by the GhostBottleneck—aggregating global information based on already extracted local representations. This sequential feature propagation enables a more natural integration of local and global information while effectively avoiding redundant computations inherent in parallel structures. Consequently, it achieves an optimal balance between lightweight efficiency and recognition performance. In the DPFA Block, The GhostBottleneck focuses on modeling intra-class diversity by emphasizing fine-grained textures, thereby improving robustness to variations within the same food category. In contrast, the PM-ViT captures long-distance dependencies to better distinguish visually similar but semantically distinct food classes.

This dual-path design facilitates effective integration of fine-grained local features and semantic global representations at each depth level. Such coordinated feature fusion significantly improves the network's ability to handle food recognition challenges characterized by large intra-class variations and small inter-class differences. Additionally, the lightweight design of both modules ensures the overall computational efficiency of the model, making DPFA-Net highly suitable for deployment on resource-constrained mobile devices.

3.2 Position Mamba Vision Transformer

The structure of the Position Mamba Vision Transformer is illustrated in Fig. 2(c).We first construct the Mamba Attention (MA) Block by integrating the Mamba structure with Separable Self-Attention(SSA), replacing the traditional Self-Attention mechanism in ViT to reduce parameter count and computational complexity while maintaining model accuracy. Since SSA only computes correlations for pixels at the same relative positions within each patch, we introduce a Local Representation(LR) Block before and after the MA module to capture correlations between different patches. The structure is illustrated in Fig. 2. After processing by the GhostBottleneck module, which captures local information, the resulting feature map is fed into the PM-ViT module. Initially, the first LR Block captures correlations between different patches. Subsequently, the MA Block models long-range pixel correlations. This is followed by another LR Block to further enhance inter-patch correlations. Finally, the output is obtained through a Feed-Forward Network (FFN). The specific design of our MA Block is depicted in Fig. 2(d). Upon entering the MA Block, the input is split into two branches. One branch sequentially undergoes a 1×1 standard convolution, a 3×3 Depth Wise Convolution (DWConv), an activation layer, and SSA. The other branch passes through a 1×1 standard convolution and an activation layer to produce its output. The outputs of the two branches are combined via a Hadamard

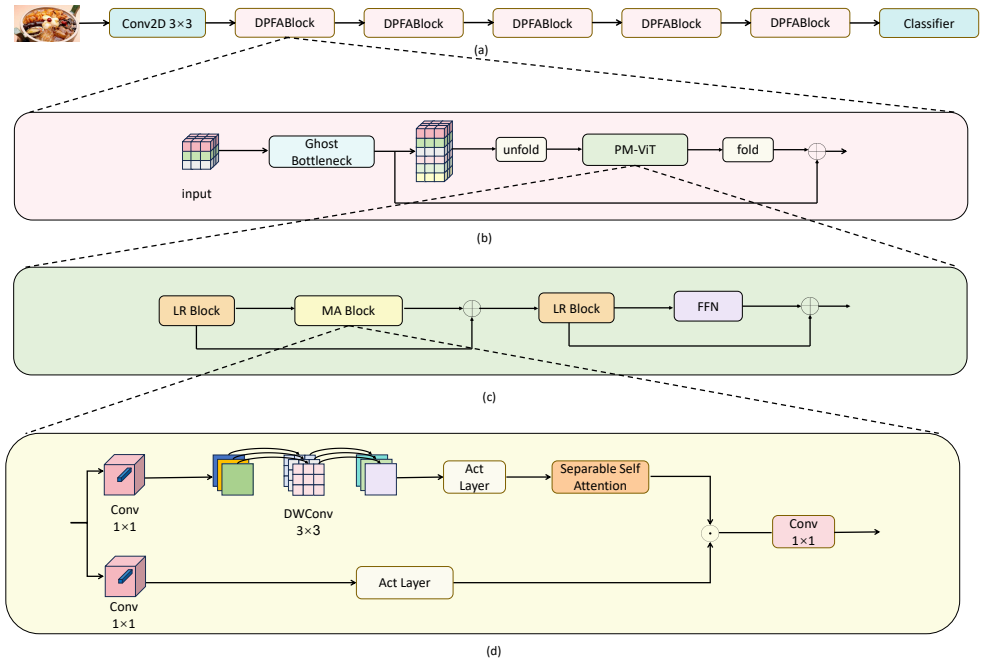

Fig. 2 Overall architecture of DPFA-Net: (a) The backbone network of DPFA-Net. (b) The architecture of the DPFA Block. (c) The architecture of the PM-ViT Block. (d) The architecture of the MA Block.

product, effectively fusing the feature information from both branches. The merged features then undergo another 1×1 convolution to produce the final output.

We propose the Position Mamba Vision Transformer (PM-ViT) module, which integrates the efficiency of the Mamba architecture in state-space processing with the global feature extraction capability of Separable Self-Attention. By introducing the Mamba Attention (MA) module to replace the conventional Self-Attention in ViTs, the model significantly reduces parameter count and computational complexity while enhancing its ability to capture global information.

In the PM-ViT structure, considering that Separable Self-Attention calculates correlations only among pixels with the same relative position within each patch, we incorporate Local Representation Blocks (LR Blocks) before and after the MA module to strengthen feature interactions across different patches.

We adopt unfold and fold operations to replace the traditional patch embedding and positional embedding used in ViTs, which improves the model's efficiency. Compared with traditional self-attention mechanisms, the Separable Self-Attention module decomposes the attention operation into two independent branches—spatial and channel—and leverages efficient operators such as group convolution. The input channels are divided into several groups, each independently undergoing convolution

and attention computation. This strategy substantially reduces parameters and computational cost, achieving a lightweight attention mechanism that balances efficiency with expressive power.

After feature extraction by the GhostBottleneck, the input features first pass through the first LR Block to enhance inter-patch correlations, then enter the MA Block to capture long-range dependencies. Following the MA Block, the features pass through the second LR Block to further strengthen information exchange among patches. The LR Block enhances information exchange between patches by computing correlations both among different patches and among pixels within each patch, providing crucial support for the MA Block in global information modeling, thereby improving the model's ability to capture global dependencies in food images.

Finally, the features are passed through a Feed-Forward Network (FFN) to produce the output. Within the MA Block, the input features are divided into two branches: one branch sequentially applies a 1×1 convolution, a 3×3 depthwise separable convolution (DWConv), an activation layer, and a Separable Self-Attention module; the other branch passes through a 1×1 convolution and an activation layer. The outputs of the two branches are then merged via a Hadamard Product (element-wise multiplication), enabling fine-grained interaction and fusion of corresponding features. A subsequent 1×1 convolution generates the final output. This overall design facilitates efficient collaboration between local and global feature representations. To achieve this functionality, the LR Block employs a 3×3 depthwise convolution to effectively capture local contextual information, enabling more natural and coherent feature interactions across patches while maintaining computational efficiency.

In the PM-ViT module, the design of the MA Block is inspired by the Mamba-Like Linear Attention (MLLA) structure [43], which demonstrates superior performance in efficiently modeling long-range dependencies while reducing computational complexity. We adopt a lightweight Separable Self-Attention (SSA) to replace the conventional Linear Self-Attention, thereby reducing parameter count while maintaining strong global feature modeling capability. However, since SSA only computes correlations between pixels at the same relative positions within each patch, it has limitations in capturing long-range dependencies across patches. To address this issue, LR Blocks are introduced before and after the MA Block to further enhance feature interaction and contextual dependency. This mechanism enables the model to more effectively capture global semantic relationships and contextual information, thereby improving overall representation capability and recognition robustness.

Specifically, assume the input is denoted as $X \in \mathbb{R}^{C\times H\times W}$. First, the unfold operation is applied to obtain $X_p \in \mathbb{R}^{C\times P\times N}$.Subsequently, the LR Block processes X_p to produce $X_l \in \mathbb{R}^{C\times P\times N}$, where C represents the number of channels, H is the height, W is the width, P denotes the patch size, and N indicates the number of patches , as shown in Equations (1)–(2).

$$X_p = \text{unfold}(X), X_p \in R^{C\times P\times N} \tag{1}$$

$$X_l = LR(X_p), X_l \in R^{C\times P\times N} \tag{2}$$

Subsequently, the X_l enters the MA Block. The left branch processes the input through a 1×1 convolution and an activation layer to obtain a value map $V^{C\times P\times N}$. The right branch applies a 1×1 convolution followed by a 3×3 depth-wise convolution (DWConv), and then passes through Separable Self-Attention (SSA) to generate an attention map $M^{C\times P\times N}$. Finally, the value map and attention map are multiplied to produce the final output $O_a^{C\times P\times N}$, as shown in Equations (3)–(5):

$$V = \text{ActLayer}(\text{Conv}(X_l)), V \in R^{C\times P\times N} \tag{3}$$

$$M = SSA(DWConv(Conv(X_l))), M \in R^{C\times P\times N} \tag{4}$$

$$O_a = V \cdot M, O_a \in R^{C\times P\times N} \tag{5}$$

The output from the integration of the two branches is processed through a 1×1 convolution and then combined via a residual connection to produce the output $O_m^{C\times P\times N}$, as shown in Equation (6):

$$O_m = \text{Conv}(O_a) + X_p, O_m \in R^{C\times P\times N} \tag{6}$$

Subsequently, the output is sequentially processed through the LR Block and the Feed-Forward Network (FFN) to obtain $O_f^{C\times P\times N}$. This is then added to $O_m^{C\times P\times N}$ to produce the final output of the PM-ViT, which is transformed via a fold operation to obtain $X_{out} \in R^{C\times H\times W}$, as shown in Equations (7)–(9):

$$O_f = \text{FFN}(LR(O_m)), O_f \in R^{C\times P\times N} \tag{7}$$

$$O_g = O_f + O_m, O_g \in R^{C\times P\times N} \tag{8}$$

$$X_{out} = fold(O_g), X_{out} \in R^{C\times H\times W} \tag{9}$$

3.3 Dual Path Feature Aggregation Block

As illustrated in Fig. 2(b), the convolutional component of the DPFA Block adopts the GhostBottleneck proposed in GhostNetV1[53], while the transformer component utilizes the PM-ViT module introduced in this paper. Each DPFA Block is constructed by serially connecting a GhostBottleneck and a PM-ViT module.

The input features are first passed through the GhostBottleneck, which effectively extracts fine-grained local features from food images while significantly reducing computational cost. Specifically, the GhostBottleneck module generates a portion of the primary features using standard convolutions and then produces additional "ghost" features through inexpensive operations, simulating a richer feature representation at a lower computational cost. The residual connection within the module further facilitates efficient information propagation and enhances feature expressiveness. This design achieves lightweight yet accurate feature extraction with substantially fewer parameters and reduced computational overhead.

Next, the features are fed into the PM-ViT module to further capture global spatial information and long-range dependencies. Within each DPFA Block, a skip connection fuses the outputs of the GhostBottleneck and the PM-ViT, enabling efficient complementation between local and global features.

By stacking multiple DPFA Blocks, DPFA-Net is able to maintain lightweight characteristics while extracting complex features from food images, significantly improving the model's representational capacity. The DPFA Block achieves efficient integration of local and global features by chaining the GhostBottleneck with the PM-ViT. The GhostBottleneck enhances the model's sensitivity to fine-grained intra-class variations, improving its adaptability to diverse instances within the same food category. Meanwhile, the PM-ViT strengthens the model's ability to discriminate between different components or regions via global dependency modeling. The skip connection further promotes multi-level feature fusion, enabling the model to effectively distinguish visually similar but semantically different food items in complex scenes, thereby improving inter-class separability and overall recognition performance.

3.4 Structure of Dual Path Feature Aggregation Network

To maintain high recognition accuracy while reducing model complexity, the hierarchical structure of DPFA-Net is redesigned based on the following considerations:

First, food image recognition is a fine-grained classification task that is highly sensitive to local details. Therefore, more GhostBottleneck modules are introduced in the early stages of the network to enhance the extraction of local features such as edge and texture information.

Second, unlike traditional deep convolutional networks that typically capture global features in high-level layers, DPFA-Net incorporates PM-ViT modules into the middle layers. By adopting a multi-branch parallel structure, it effectively models spatial contextual relationships, enabling the network to gain global perception capabilities at earlier stages. As a result, the dependence on deep stacking of modules in later stages is significantly reduced.

Third, considering that feature extraction focuses vary across different layers, the shallow layers primarily utilize lightweight convolutional structures for detailed feature extraction, while the middle layers integrate more PM-ViT blocks to model global semantic information. The deeper layers maintain moderate depth to facilitate convergence for semantic discrimination.

This overall architectural design ensures that DPFA-Net achieves strong feature representation capability while remaining lightweight, making it particularly suitable for food image recognition tasks in resource-constrained environments. The detailed network architecture is shown in Table 1.

4 Results

4.1 Datasets

To validate the effectiveness of the proposed model, we selected three publicly available food image datasets for evaluation: ETHZ Food-101[47], Vireo Food-172[48], and

Table 1 Network specifications of DPFA-Net. $\alpha \in [0.5, 2.0]$ represents the width multiplier used to generate DPFA-Net models with different levels of complexity.

Component	Input	Operator	Patch Size	Output Channel	Stride
Head	256×256	Conv2D 3×3	-	32α	2
Block Group 1	128×128	GhostBottleneck	-	32α	2
	128×128	GhostBottleneck	-	32α	1
	128×128	PM-ViT	2×2	32α	-
Block Group 2	128×128	GhostBottleneck	-	64α	2
	64×64	GhostBottleneck	-	64α	1
	32×32	GhostBottleneck	-	64α	1
	32×32	PM-ViT	2×2	64α	-
	32×32	PM-ViT	2×2	64α	-
Block Group 3	64×64	GhostBottleneck	-	96α	2
	32×32	GhostBottleneck	-	96α	1
	32×32	GhostBottleneck	-	96α	1
	32×32	PM-ViT	2×2	96α	-
	32×32	PM-ViT	2×2	96α	-
	32×32	PM-ViT	2×2	96α	-
Block Group 4	32×32	GhostBottleneck	-	160α	2
	16×16	GhostBottleneck	-	160α	1
	16×16	GhostBottleneck	-	160α	1
	16×16	PM-ViT	2×2	160α	-
	16×16	PM-ViT	2×2	160α	-
Block Group 5	16×16	GhostBottleneck	-	320α	2
	8×8	PM-ViT	2×2	320α	-

UEC Food-256[49]. These datasets are representative in terms of scale, category complexity, and relevance to real-world applications, making them suitable for assessing the robustness and effectiveness of the model. The ETHZ Food-101 dataset comprises 101 categories with a total of 101,000 images, of which 75,750 are used for training and 25,250 for validation. The Vireo Food-172 dataset includes 172 categories, with a total of 110,241 images, split into 66,071 for training and 44,170 for validation. The UEC Food-256 dataset contains 256 categories, with 31,395 images, of which 22,095 are used for training and 9,300 for validation.

4.2 Implementation details

We adopt an input image resolution of 256×256, set the batch size to 32, and use the AdamW optimizer for model training. In the initial training phase, we implement a linear warm-up strategy with an initial learning rate of 1×10^{-6} for 20,000 iterations. Subsequently, we employ a cosine annealing strategy to adjust the learning rate from 2×10^{-3} to 2×10^{-5}.

4.3 Results on ETHZ Food-101

Table 2, 3 presents a performance comparison of various models on the ETHZ Food-101 dataset, with models grouped by parameter count in ascending order. We organize the experimental results by parameter scale, and our model outperforms current

mainstream lightweight models across all seven parameter ranges. In the parameter range of 0.5M to 1M, DPFA-Net-0.5 achieves an accuracy of 87.1%, surpassing AFNet-1.0 (86.4%) and ShuffleNetV2-0.5 (74.3%). Compared with the above two models, although the computational overhead slightly increases, the ratio of performance to resource consumption remains advantageous under the trend of continuously improving device computational capabilities. In the 1M–2M range, with similar or fewer parameters, DPFA-Net-0.75 (88.9%) significantly outperforms GhostNetV2-0.5 (81.2%) and MobileNetV3-0.5 (82.4%), with Top-1 accuracy improvements of 7.7%, 6.5%, and 2.0%, respectively. In the 2M–4M model group, DPFA-Net-1.0 and DPFA-Net-1.25 (89.2%/89.9%) maintain competitiveness, demonstrating clear improvements over ViT-based MobileNetV3-0.75 (85.5%) and MobileViTv2-1.0 (87.6%). For larger parameter models, DPFA-Net-1.5 to DPFA-Net-2.0 achieve Top-1 accuracies exceeding 90%, with DPFA-Net-2.0 (91.0%, 3497.0M FLOPs) outperforming MobileViTv2-1.75 (88.9%, 5493M FLOPs) by 2.1% in accuracy while reducing computational load by nearly 57%. Compared with Mamba-based models, VMamba-t (86.1%) has 30.0M parameters, whereas DPFA-Net-2.0 has only 9.2M and achieves 91.0%, an improvement of 4.9%, far surpassing VMamba-t. Similarly, VMamba-s (87.2%) has 49.5M parameters, while DPFA-Net-1.75 has just 7.2M (less than one-sixth of the former) yet reaches 90.6%. Against the latest lightweight Transformer, EdgeViT-xxs (62.2%, 3.8M parameters), DPFA-Net-1.0 (89.2%, 2.3M parameters) reduces parameters by nearly 40% while achieving a much higher accuracy, showing a clear lead. Overall, the comparison demonstrates that the DPFA architecture consistently delivers higher recognition accuracy with comparable or even fewer parameters, exhibiting exceptional parameter efficiency and significant performance advantages.

4.4 Results on Vireo Food-172

Table 4, 5 presents the performance comparison of models on the Vireo Food-172 dataset, which includes complex Chinese dishes with diverse visual features and significant background interference. With a similar number of parameters, DPFA-Net 0.75 (1.5M,89.7%) outperforms ShuffleNetV2-1.0 (1.M,81.0%) by 8.7%; DPFA-Net-1.0 (2.4M,90.1%) surpasses ShuffleNetV2-1.5 (2.7M82.4%) by 7.7%; and DPFA-Net-1.5 (5.3M,91.2%) achieves a 7.4% higher accuracy than ShuffleNetV2-2.0 (5.7M,83.8%). Compared to EfficientNetB0 (4.8M,83.6%), DPFA-Net-1.25 (3.8M,91.7%) improves accuracy by 8.1% with fewer parameters. Computational efficiency evaluations show that DPFA-Net maintains moderate resource consumption while achieving high recognition rates. In scenarios with both higher parameter counts and accuracy, DPFA-Net-1.75 (2925.6M FLOPs) reduces computational load by 46.7% compared to MobileViTv2-1.75 (5493.4M FLOPs). Across all parameter ranges, DPFA-Net achieves higher accuracy with fewer parameters, surpassing MobileNetV3 by 5% and MobileViTv2 by approximately 3%. When compared with the latest lightweight Transformers, DPFA-Net shows remarkable advantages. For example, EfficientViT-M4 (52.3%) has 8.5M parameters, whereas DPFA-Net-1.5, with only 5.346M parameters (about 37% less than the former), achieves an accuracy of 91.2%, leading by a large margin. Similarly, compared to EfficientViT-M5 (50.7%, 12.2M parameters), DPFA

Table 2 Performance comparison of CNN-based models on the ETHZ Food-101 dataset. DPFA-Net-x: where x denotes the width multiplier of the base model.

Method	Top-1 Acc	#Params	#FLOPs
AFNet-0.5 [50]	82.9%	0.3M	60.1M
ShuffleNetV2-0.5 [19]	74.3%	0.5M	41.6M
AFNet-0.75 [50]	85.8%	0.5M	41.6M
AFNet-1.0 [50]	86.4%	0.6M	164.0M
DPFA-Net-0.5	**87.1%**	**0.7M**	**319.5M**
AFNet-1.25 [50]	87.3%	0.9M	265.0M
AFNet-1.5 [50]	87.8%	1.3M	363.0M
ShuffleNetV2-1.0 [19]	78.0%	1.4M	148.8M
DPFA-Net-0.75	**88.9%**	**1.5M**	**773.1M**
MobileNetV3-0.5 [22]	82.4%	1.5M	73.3M
GhostNetV2-0.5 [51]	81.2%	1.7M	54.0M
AFNet-1.75 [50]	88.4%	1.7M	474.0M
AFNet-2.0 [50]	88.3%	2.1M	576.0M
DPFA-Net-1.0	**89.2%**	**2.3M**	**918.5M**
ShuffleNetV2-1.5 [19]	80.3%	2.6M	303.6M
MobileNetV4-Conv-S [23]	82.0%	2.6M	409.6M
MobileNetV3-0.75 [22]	85.5%	2.8M	161.9M
DPFA-Net-1.25	**89.9%**	**3.7M**	**1621.0M**
MobileNetV3-1.0 [22]	86.2%	4.3M	218.9M
EfficientNeT-B0 [21]	85.2%	4.7M	566.9M
GhostNetV2-1.0 [51]	83.6%	5.0M	176.9M
DPFA-Net-1.5	**90.4%**	**5.3M**	**2718.0M**
ShuffleNetV2-2.0 [19]	82.0%	5.6M	596.4M
MobileNetV3-1.25 [22]	86.2%	6.4M	366.8M
DPFA-Net-1.75	**90.6%**	**7.2M**	**2925.6M**
GhostNetV2-1.3 [51]	84.8%	7.8M	282.5M
MobileNetV3-1.5 [22]	86.5%	8.6M	500.4M
MobileNetV4-Conv-M [23]	83.6%	8.6M	1740.8M
DPFA-Net-2.0	**91.0%**	**9.2M**	**3497.0M**
MobileNetV4-Hybrid-M [23]	84.6%	9.9M	1945.6M
GhostNetV2-1.6 [51]	85.5%	11.2M	415.0M
GhostNetV2-1.9 [51]	85.7%	15.3M	572.8M
MobileNetV4-Conv-L [23]	85.4%	31.4M	4403.2M
MobileNetV4-Hybrid-L [23]	86.3%	36.6M	5120.1M

Net-1.75 has 7.263M parameters, about 40% fewer than the former, yet achieves 91.16% accuracy, significantly surpassing it and demonstrating superior performance.

4.5 Results on UEC Food-256

Table 6, 7 presents a performance comparison of models on the UEC Food-256 dataset, which encompasses 256 diverse food categories and poses significant recognition challenges. DPFA-Net demonstrates remarkable advancements over various mainstream lightweight networks. Compared to efficient CNN architectures such as MobileNetV3, DPFA-Net achieves a performance improvement of 5%-9%, with DPFA-Net-1.25 (72.1%) significantly outperforming MobileNetV3-1.25 (65.7%). When compared to the AFNet model, DPFA-Net-1.0 (71.9%) surpasses AFNet-2.0 (70.7%)

Table 3 Performance comparison of ViT-based and Mamba-based models on the ETHZ Food-101 dataset. DPFA-Net-x: where x denotes the width multiplier of the base model.

Method	Top-1 Acc	#Params	#FLOPs
DPFA-Net-0.5	**87.1%**	**0.7M**	**319.5M**
EHFR-Net-0.5 [44]	88.2%	0.8M	412.2M
MobileViTv2-0.5 [38]	86.9%	1.1M	465.9M
DPFA-Net-0.75	**88.9%**	**1.5M**	**773.1M**
EHFR-Net-0.75 [44]	89.2%	1.8M	955.9M
EfficientViT-M0 [27]	35.7%	2.1M	205.1M
DPFA-Net-1.0	**89.2%**	**2.3M**	**918.5M**
MobileViTv2-0.75 [38]	87.3%	2.5M	1029.9M
EfficientViT-M1 [27]	39.9%	2.8M	307.3M
EHFR-Net-1.0 [44]	89.5%	2.8M	1238.5M
DPFA-Net-1.25	**89.9%**	**3.7M**	**1621.0M**
EdgeViT-xxs [36]	62.2%	3.8M	1126.4M
EfficientViT-M2 [27]	43.5%	4.0M	409.6M
MobileViTv2-1.0 [38]	87.6%	4.4M	1814.7M
EHFR-Net-1.25 [44]	89.9%	4.5M	2066.5M
TinyViT-5M [29]	78.3%	5.1M	2457.6M
DPFA-Net-1.5	**90.4%**	**5.3M**	**2718.0M**
EdgeViT-xs [36]	66.3%	6.4M	2252.8M
EHFR-Net-1.5 [44]	90.0%	6.4M	2941.2M
EfficientViT-M3 [27]	42.4%	6.6M	512.1M
MobileViTv2-1.25 [38]	88.3%	6.9M	2820.2M
Vim-t [39]	59.9%	7.0M	102.4M
DPFA-Net-1.75	**90.6%**	**7.2M**	**2925.6M**
EfficientViT-M4 [27]	46.2%	8.5M	614.4M
EHFR-Net-1.75 [44]	90.0%	8.7M	3840.1M
DPFA-Net-2.0	**91.0%**	**9.2M**	**3497.0M**
MobileViTv2-1.5 [38]	88.6%	9.9M	4046.4M
TinyViT-11M [29]	82.4%	10.6M	3891.2M
EHFR-Net-2.0 [44]	90.0%	11.1M	4731.0M
EfficientViT-M5 [27]	44.8%	12.1M	1126.4M
EdgeViT-s [36]	61.7%	12.8M	3891.2M
MobileViTv2-1.75 [38]	88.9%	13.4M	5493.3M
MobileViTv2-2.0 [38]	89.5%	17.5M	7161.0M
TinyViT-21M [29]	85.7%	20.7M	8396.8M
Vim-s [39]	74.0%	25.5M	103.1M
VMamba-t [40]	86.1%	30.0M	4976.6M
VMamba-s [40]	87.2%	49.5M	8939.5M
VMamba-b [40]	87.7%	87.6M	15728.6M
Vim-b [39]	78.7%	96.9M	204.9M

while reducing the number of parameters by approximately 4%. In comparisons with more efficient architectures, DPFA-Net-1.25 (72.1%) outperforms EHFR-Net-1.25 (71.4%), which has a larger parameter count, and achieves an 8.1% higher accuracy than EfficientNetB0 (64.0%) with fewer parameters. In comparison to state-of-the-art Transformer-based models, DPFA-Net-1.0 (71.9%) achieves similar Top-1 accuracy with a 10.7% reduction in FLOPs compared to MobileViTv2-0.75 (69.8%). Against Mamba-based models, VMamba-s has 49.6M parameters and 62.1% accuracy, while DPFA-Net-1.75 has only 7.31M parameters (about one-seventh) yet achieves 72.9%,

Table 4 Performance comparison of CNN-based models on the Vireo Food-172 dataset. DPFA-Net-x: where x denotes the width multiplier of the base model.

Method	Top-1 Acc	#Params	#FLOPs
AFNet-0.5 [50]	83.7%	0.4M	69.0M
ShuffleNetV2-0.5 [19]	74.3%	0.5M	41.6M
AFNet-0.75 [50]	86.5%	0.6M	134.0M
DPFA-Net-0.5	**87.5%**	**0.7M**	**319.5M**
AFNet-1.0 [50]	87.1%	0.7M	165.0M
AFNet-1.25 [50]	88.0%	1.0M	265.0M
AFNet-1.5 [50]	87.8%	1.3M	363.0M
ShuffleNetV2-1.0 [19]	81.0%	1.4M	148.8M
DPFA-Net-0.75	**89.7%**	**1.5M**	**773.1M**
MobileNetV3-0.5 [22]	83.0%	1.6M	73.4M
GhostNetV2-0.5 [51]	81.8%	1.8M	54.1M
AFNet-1.75 [50]	88.9%	1.9M	474.0M
AFNet-2.0 [50]	89.0%	2.3M	576.0M
DPFA-Net-1.0	**90.1%**	**2.4M**	**918.5M**
ShuffleNetV2-1.5 [19]	82.4%	2.7M	303.7M
MobileNetV4-Conv-S [23]	84.5%	2.7M	409.7M
MobileNetV3-0.75 [22]	85.9%	2.9M	162.0M
DPFA-Net-1.25	**90.7%**	**3.8M**	**1621.0M**
MobileNetV3-1.0 [22]	86.7%	4.4M	219.0M
EfficientNeT-B0 [21]	83.6%	4.8M	567.0M
GhostNetV2-1.0 [51]	84.7%	5.1M	117.0M
DPFA-Net-1.5	**91.2%**	**5.3M**	**2178.0M**
ShuffleNetV2-2.0 [19]	83.8%	5.7M	596.6M
MobileNetV3-1.25 [22]	86.9%	6.5M	366.9M
DPFA-Net-1.75	**91.2%**	**7.3M**	**2925.6M**
GhostNetV2-1.3 [51]	85.7%	7.9M	282.5M
MobileNetV3-1.5 [22]	86.5%	8.7M	500.4M
MobileNetV4-Conv-M [23]	85.7%	8.7M	1740.9M
DPFA-Net-2.0	**91.2%**	**9.2M**	**3497.0M**
MobileNetV4-Hybrid-M [23]	86.3%	10.0M	1945.7M
GhostNetV2-1.6 [51]	86.2%	11.3M	415.1M
GhostNetV2-1.9 [51]	85.7%	15.3M	572.8M
MobileNetV4-Conv-L [23]	87.3%	31.5M	4403.3M
MobileNetV4-Hybrid-L [23]	87.8%	36.7M	5120.2M

far surpassing VMamba-s.Compared with the MobileNetV4 series, MobileNetV4-Hybrid-L (36.8M parameters, 66.6% accuracy) is outperformed by DPFA-Net-2.0 (9.261M parameters, 73.3% accuracy), which achieves higher accuracy with far fewer parameters. Against TinyViT, TinyViT-11M (10.7M parameters, 53.1% accuracy) is outperformed by DPFA-Net-1.75 (7.31M parameters) with about 32% fewer parameters and a much higher 72.9% accuracy. The experimental results indicate that our model is capable of efficiently handling complex datasets like UEC Food-256. While maintaining its lightweight characteristics, DPFA-Net achieves leading recognition performance, providing a practical solution for large-scale food recognition applications in resource-constrained environments.

Table 5 Performance comparison of ViT-based and Mamba-based models on the Vireo Food-172 dataset. DPFA-Net-x: where x denotes the width multiplier of the base model.

Method	Top-1 Acc	#Params	#FLOPs
DPFA-Net-0.5	**87.5%**	**0.7M**	**319.5M**
EHFR-Net-0.5 [44]	89.2%	0.8M	412.3M
MobileViTv2-0.5 [38]	87.3%	1.2M	465.9M
DPFA-Net-0.75	**89.7%**	**1.5M**	**773.1M**
EHFR-Net-0.75 [44]	89.9%	1.8M	956.0M
EfficientViT-M0 [27]	44.2%	2.2M	205.2M
DPFA-Net-1.0	**90.1%**	**2.4M**	**918.5M**
MobileViTv2-0.75 [38]	88.0%	2.5M	1030.0M
EfficientViT-M1 [27]	47.8%	2.8M	307.4M
EHFR-Net-1.0 [44]	90.3%	2.8M	1210.3M
DPFA-Net-1.25	**90.7%**	**3.8M**	**1621.0M**
EdgeViT-xxs [36]	67.2%	3.8M	1126.5M
EfficientViT-M2 [27]	49.8%	4.0M	409.7M
MobileViTv2-1.0 [38]	88.2%	4.5M	1814.7M
EHFR-Net-1.25 [44]	90.6%	4.5M	2066.5M
TinyViT-5M [29]	81.6%	5.1M	2457.7M
DPFA-Net-1.5	**91.2%**	**5.3M**	**2178.0M**
EHFR-Net-1.5 [44]	90.7%	6.5M	2941.2M
EdgeViT-xs [36]	67.5%	6.5M	2252.9
EfficientViT-M3 [27]	50.8%	6.6M	512.2M
MobileViTv2-1.25 [38]	88.0%	6.9M	2820.2M
Vim-t [39]	75.1%	7.0M	102.5M
DPFA-Net-1.75	**91.2%**	**7.3M**	**2925.6M**
EfficientViT-M4 [27]	52.3%	8.5M	614.5M
EHFR-Net-1.75 [44]	90.7%	8.8M	3840.2M
DPFA-Net-2.0	**91.2%**	**9.2M**	**3497.0M**
MobileViTv2-1.5 [38]	88.7%	10.0M	4046.4M
TinyViT-11M [29]	84.9%	10.6M	3891.3M
EHFR-Net-2.0 [44]	90.8%	11.1M	4731.1M
EfficientViT-M5 [27]	50.7%	12.2M	1126.5M
EdgeViT-s [36]	65.5%	12.8M	3891.3M
MobileViTv2-1.75 [38]	89.1%	13.5M	5493.4M
MobileViTv2-2.0 [38]	89.4%	17.6M	7161.0M
TinyViT-21M [29]	87.6%	20.7M	8396.9M
Vim-s [39]	77.3%	25.5M	103.2M
VMamba-t [40]	88.1%	30.1M	4976.8M
VMamba-s [40]	88.8%	49.5M	8939.6M
VMamba-b [40]	89.3%	87.7M	15728.8M
Vim-b [39]	80.3%	97.0M	205.0M

4.6 Analysis of Memory and Latency

As shown in Table 8, DPFA-Net-0.5 maintains a moderate memory footprint of 2379.0M while achieving a Top-1 accuracy of 87.1%, which is substantially higher than other lightweight baselines such as EHFR-Net-0.5 and MobileNetV4-Conv-S. For example, although AFNet-0.5, EfficientViT-M0, and EdgeViT-xxs exhibit lower memory usage, their Top-1 accuracies are only 82.9%, 35.7%, and 62.2%, respectively—indicating a clear performance gap compared to DPFA-Net-0.5. This

Table 6 Performance comparison of CNN-based models on the UEC Food-256 dataset. DPFA-Net-x: where x denotes the width multiplier of the base model.

Method	Top-1 Acc	#Params	#FLOPs
AFNet-0.5 [50]	65.5%	0.5M	69.0M
ShuffleNetV2-0.5 [19]	74.3%	0.5M	41.6M
DPFA-Net-0.5	**71.7%**	**0.7M**	**319.5M**
AFNet-0.75 [50]	86.5%	0.6M	134.0M
AFNet-1.0 [50]	70.0%	1.2M	266.0M
AFNet-1.25 [50]	70.0%	1.2M	266.0M
DPFA-Net-0.75	**71.9%**	**1.5M**	**773.1M**
ShuffleNetV2-1.0 [19]	55.2%	1.5M	149.0M
AFNet-1.5 [50]	70.4%	1.6M	73.5M
MobileNetV3-0.5 [22]	62.1%	1.7M	363.0M
GhostNetV2-0.5 [51]	61.1%	1.9M	54.2M
AFNet-1.75 [50]	71.4%	2.1M	474.0M
DPFA-Net-1.0	**71.9%**	**2.4M**	**919.6M**
AFNet-2.0 [50]	70.7%	2.5M	576.0M
ShuffleNetV2-1.5 [19]	57.5%	2.7M	303.7M
MobileNetV4-Conv-S [23]	62.3%	2.78M	409.8M
DPFA-Net-1.25	**72.1%**	**3.8M**	**1621.0M**
MobileNetV3-1.0 [22]	65.5%	4.5M	219.0M
EfficientNeT-B0 [21]	64.0%	4.9M	567.1M
GhostNetV2-1.0 [51]	63.9%	5.2M	177.1M
DPFA-Net-1.5	**72.8%**	**5.4M**	**2178.0M**
ShuffleNetV2-2.0 [19]	60.1%	5.9M	596.7M
MobileNetV3-1.25 [22]	65.7%	6.6M	367.0M
DPFA-Net-1.75	**72.9%**	**7.3M**	**2925.6M**
GhostNetV2-1.3 [51]	65.0%	8.0M	282.7M
MobileNetV3-1.5 [22]	67.1%	8.8M	500.5M
MobileNetV4-Conv-M [23]	65.9%	8.8M	1741.0M
DPFA-Net-2.0	**73.3%**	**9.3M**	**3497.0M**
MobileNetV4-Hybrid-M [23]	66.1%	10.1M	1945.8M
GhostNetV2-1.6 [51]	65.5%	11.4M	415.2M
GhostNetV2-1.9 [51]	66.1%	15.5M	573.0M
MobileNetV4-Conv-L [23]	66.2%	31.6M	4403.4M
MobileNetV4-Hybrid-L [23]	66.6%	36.8M	5120.3M

demonstrates that DPFA-Net-0.5 leverages a slightly higher memory cost to achieve significantly better accuracy and feature modeling capabilities.In contrast, models such as MobileNetV4-Conv-S, EHFR-Net-0.5, VMamba-t, and MobileViTv2-0.5 not only require substantially more memory—3015.0M, 3188.0M, 4529.0M, and 5291.0M, respectively—but also fail to deliver notable accuracy improvements; in most cases, their accuracies are even lower than that of DPFA-Net-0.5. Moreover, latency experiments conducted on an NVIDIA A800 GPU further validate the efficiency of the proposed model. DPFA-Net-0.5 achieves an inference latency of 14.7 ms, which is comparable to MobileViTv2-0.5 (14.8 ms) and VMamba-t (15.7 ms), yet significantly lower than EHFR-Net-0.5 (37.0 ms) and Vim-t (20.8 ms), while maintaining superior accuracy.This balance between computational delay and recognition performance demonstrates that DPFA-Net-0.5 achieves an optimal trade-off among accuracy, memory consumption, and inference speed. Overall, these results highlight the superior

Table 7 Performance comparison of CNN-based models on the UEC Food-256 dataset. DPFA-Net-x: where x denotes the width multiplier of the base model.

Method	Top-1 Acc	#Params	#FLOPs
DPFA-Net-0.5	**71.7%**	**0.7M**	**319.5M**
EHFR-Net-0.5 [44]	69.3%	0.8M	412.3M
MobileViTv2-0.5 [38]	69.1%	1.2M	465.9M
DPFA-Net-0.75	**71.9%**	**1.5M**	**773.1M**
EHFR-Net-0.75 [44]	70.2%	1.8M	956.0M
EfficientViT-M0 [27]	27.7%	2.2M	205.3M
DPFA-Net-1.0	**71.9%**	**2.4M**	**919.6M**
MobileViTv2-0.75 [38]	69.8%	2.6M	1030.0M
EfficientViT-M1 [27]	34.2%	2.8M	307.5M
EHFR-Net-1.0 [44]	70.3%	2.9M	1210.4M
DPFA-Net-1.25	**72.1%**	**3.8M**	**1621.0M**
EdgeViT-xxs [36]	30.3%	3.9M	1126.6M
EfficientViT-M2 [27]	35.5%	4.0M	409.8M
MobileViTv2-1.0 [38]	70.0%	4.5M	1814.8M
EHFR-Net-1.25 [44]	71.4%	4.6M	2066.5M
TinyViT-5M [29]	47.6%	5.2M	2457.8M
DPFA-Net-1.5	**72.8%**	**5.4M**	**2178.0M**
EHFR-Net-1.5 [44]	71.4%	6.5M	2941.3M
EdgeViT-xs [36]	30.0%	6.5M	2253.0M
EfficientViT-M3 [27]	36.9%	6.7M	512.3M
Vim-t [39]	34.1%	7.0M	102.6M
MobileViTv2-1.25 [38]	71.2%	7.0M	2820.3M
DPFA-Net-1.75	**72.9%**	**7.3M**	**2925.6M**
EfficientViT-M4 [27]	36.7%	8.5M	614.6M
EHFR-Net-1.75 [44]	71.9%	8.8M	3840.2M
DPFA-Net-2.0	**73.3%**	**9.3M**	**3497.0M**
MobileViTv2-1.5 [38]	71.2%	10.0M	4046.5M
TinyViT-11M [29]	53.1%	10.7M	3891.4M
EHFR-Net-2.0 [44]	72.1%	11.2M	4731.2M
EfficientViT-M5 [27]	38.9%	12.2M	1126.6M
EdgeViT-s [36]	27.1%	12.8M	3891.4M
MobileViTv2-1.75 [38]	71.4%	13.6M	5493.4M
MobileViTv2-2.0 [38]	71.5%	17.7M	7161.1M
TinyViT-21M [29]	61.0%	20.8M	8397.0M
Vim-s [39]	45.1%	25.5M	103.3M
VMamba-t [40]	62.0%	30.1M	4976.8M
VMamba-s [40]	62.1%	49.6M	8939.7M
VMamba-b [40]	63.9%	87.8M	15728.8M
Vim-b [39]	51.9%	97.0M	205.1M

efficiency of DPFA-Net-0.5, which attains high recognition accuracy, low memory cost, and fast inference speed, making it highly suitable for real-time deployment on resource-constrained mobile or embedded devices.

4.7 Qualitative Analysis and Visualization

DPFA-Net overcomes the limitations of traditional CNN models, which primarily focus on local feature extraction, demonstrating superior performance in capturing global

Table 8 Comparative Analysis of Memory and Latency among Different Models

Method	Top-1 Acc	#Params	#FLOPs	#Memory	Latency
AFNet-0.5 [50]	82.9%	0.3M	60.1M	269.0M	10.4ms
DPFA-Net-0.5	**87.1%**	**0.7M**	**319.5M**	**2379.0M**	**14.7ms**
EHFR-Net-0.5 [44]	88.2%	0.8M	412.2M	3188.0M	37.0ms
MobileViTv2-0.5 [38]	86.8%	1.1M	469.5M	5291.0M	14.8ms
EfficientViT-M0 [27]	35.7%	2.1M	205.1M	474.0M	9.3ms
MobileNetV4-Conv-S [23]	82.0%	2.6M	409.6M	3015.0M	7ms
EdgeViT-xxs [36]	62.2%	3.8M	1126.4M	943.0M	12.4ms
TinyViT-5M [29]	78.3%	5.1M	2457.6M	2031.0M	9.4ms
Vim-t [39]	59.9%	7.0M	102.4M	2330.0M	20.8ms
VMamba-t [40]	86.1%	30.0M	4976.6M	4529.0M	15.7ms

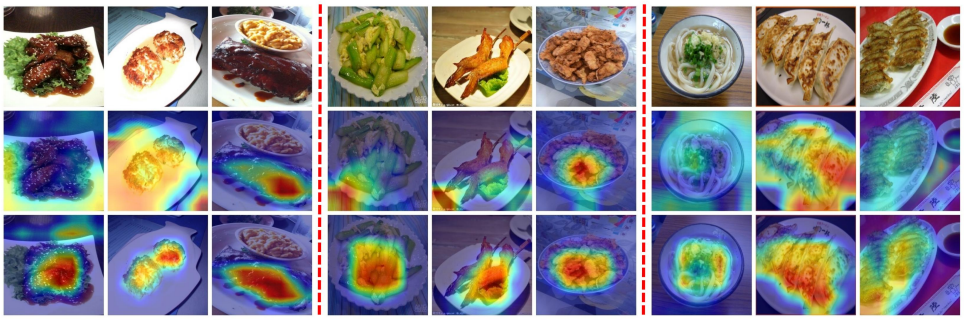

Fig. 3 Visualization of experimental results

features. Fig. 3 illustrates a comparison using the Grad-CAM[52] method. The visualization results are derived from models based solely on CNNs. In Fig. 3, the first row displays original food images from the ETHZ Food-101 dataset (columns 1–3), the Vireo Food-172 dataset (columns 4–6), and the UEC Food-256 dataset (columns 7–9). The second row shows heatmaps generated by the CNN-based network, while the third row presents heatmaps generated by DPFA-Net. From Fig. 3, the following observations can be made: (1) For food images with diverse ingredients and complex backgrounds, CNN-based networks tend to focus on irrelevant regions, with scattered attention and susceptibility to background interference, leading to misclassification of food items. (2) DPFA-Net accurately localizes critical ingredient regions, with precision focused attention, demonstrating that DPFA-Net effectively integrates local and global features to overcome the limitations of single-feature representations. Its attention mechanism adaptively filters discriminative information, suppresses visual noise, and enhances the model's robustness in recognizing food items in complex scenes.

4.8 Ablation Investigations

In this section, we conduct an ablation analysis of the key design elements in the proposed model through image classification on three datasets. The results are summarized in Table 9.

Effectiveness of DPFABlock: To investigate the efficacy of local feature extraction (GhostBottleneck module) and global feature extraction (PM-ViT) within DPFABlock, we integrated local feature analysis with global information processing by constructing DPFABlock as the core component. Experiments demonstrate that, compared to single-feature extraction strategies, the fused architecture DPFA-Net-1.0 achieves superior performance across datasets. On ETHZ Food-101, DPFA-Net-1.0 attains a Top-1 accuracy of 89.20%, surpassing models using only local modules (86.94%) and global modules (86.89%). On Vireo Food-172, it achieves a Top-1 accuracy of 90.10%, outperforming local models (84.82%) and global models (88.12%). On UEC Food-256, the Top-1 accuracy reaches 71.89%, outperforming the local model (70.84%) and the global model (69.58%). These results validate the effectiveness of DPFABlock in food image analysis, as it simultaneously captures texture details and holistic features, thereby enhancing recognition accuracy. Moreover, this fused architecture maintains high accuracy while ensuring a reasonable parameter scale and computational load, demonstrating its deployment potential and resource efficiency in practical applications.

Effectiveness of PM-ViT: To evaluate the role of PM-ViT in the DPFA-Net architecture, we conducted an ablation study by replacing the PM-ViT module with a standard Transformer module. The results indicate that the model using the PM-ViT Block outperforms the one employing the standard Transformer across all datasets: an accuracy improvement of 0.35% on ETHZ Food-101 (89.20% vs. 88.85%), 0.38% on Vireo Food-172 (90.10% vs. 89.72%), and 2.58% on UEC Food-256 (71.89% vs. 69.31%). Meanwhile, the FLOPs are reduced to 919.552M, accounting for only 23% of those of the standard Transformer (3939.33M). These experiments confirm that PM-ViT optimizes the attention mechanism, significantly reducing both the number of parameters and computational complexity while maintaining high recognition accuracy. This enables more efficient extraction of key features in food images, offering a viable solution for food recognition applications in resource-constrained environments. As an indispensable component of the DPFA-Net architecture, PM-ViT provides critical support for enhancing model performance.

Effectiveness of the LR Block: To evaluate the contribution of the LR Block, we conducted an ablation study by removing it (w/o LR). The results show a clear performance drop across all datasets: the Top-1 accuracy decreases to 86.4% on ETHZ Food-101, 87.8% on Vireo Food-172, and 66.1% on UEC Food-256. Meanwhile, the number of parameters slightly increases (by approximately 0.1M–0.2M). These results indicate that the LR Block effectively enhances the model's ability to capture local textures and critical features. It improves food image recognition accuracy while maintaining a lightweight design, demonstrating its indispensable role in the DPFA-Net architecture.

Effectiveness of the MA Module: To assess the impact of the MA module, we replaced it with SSA (MA $\rightarrow$ SSA), which resulted in a significant decrease in

Table 9 Ablation study results of the proposed method on the ETHZ Food-101, Vireo Food-172 and UEC Food-256 dataset.

Dataset	Ablation	Top-1 Acc.	#Params	#FLOPs
ETHZ Food-101	DPFA-Net-1.0	89.2%	2.3M	918.5M
	w/o LR	86.4%	2.4M	733.1M
	w/o local	86.9%	1.1M	491.5M
	w/o global	86.8%	2.4M	750.5M
	MA→SSA	83.8%	1.9M	490.5M
	PM-ViT→ViT	88.8%	1.9M	3939.3M
Vireo Food-172	DPFA-Net-1.0	90.1%	2.3M	918.5M
	w/o LR	87.8%	2.5M	733.1M
	w/o local	84.8%	1.1M	491.5M
	w/o global	88.1%	2.5M	750.5M
	MA→SSA	85.7%	1.9M	490.5M
	PM-ViT→ViT	89.7%	1.9M	3939.3M
UEC Food-256	DPFA-Net-1.0	71.8%	2.3M	919.5M
	w/o LR	66.1%	2.5M	733.1M
	w/o local	70.8%	1.1M	491.5M
	w/o global	69.5%	2.5M	750.5M
	MA→SSA	64.7%	2.0M	490.5M
	PM-ViT→ViT	69.3%	2.3M	3939.3M

Top-1 accuracy across the three datasets: from 89.2% to 83.8% on ETHZ Food-101, from 90.1% to 85.7% on Vireo Food-172, and from 71.8% to 64.7% on UEC Food-256. This demonstrates that the MAmodule plays a critical role in modeling global features within DPFA-Net. MA adaptively captures key features across channels and spatial locations, effectively integrating local and global information, thereby enhancing fine-grained food recognition. Although SSA slightly reduces the computational cost, the substantial performance drop further confirms that the MA module is indispensable for maintaining high recognition accuracy and robustness in complex scenarios.

Our proposed DPFA-Net model integrates local feature extraction, global information capture, and an optimized transformer mechanism, achieving exceptional recognition accuracy while maintaining a lightweight design. Experiments validate the multifaceted advantages of this architecture: on ETHZ Food-101, DPFA-Net-1.0 (2.341M parameters, 89.20% accuracy) significantly outperforms single-feature models and the standard transformer; on Vireo Food-172, its performance (2.364M parameters, 90.10% accuracy) also leads the field; and on UEC Food-256 (2.391M parameters, 71.89% accuracy), it demonstrates robust feature extraction capabilities on complex datasets. Notably, DPFA-Net-1.0 achieves a breakthrough in computational efficiency, with an operational load of approximately 919M FLOPs, only 23% of that required by the standard transformer (3939M FLOPs), making it highly suitable for resource-constrained environments. This balanced design philosophy optimizes the model across three dimensions—accuracy, parameter count, and computational complexity—providing an efficient and practical solution for the field of food recognition.

5 Conclusion

In this paper, we design and implement a novel lightweight architecture, DPFA-Net, which effectively integrates local and global representations through a dual-path mechanism. DPFA-Net is built upon two complementary modules: on one hand, it employs the computationally efficient GhostBottleneck structure to precisely capture local texture features; on the other hand, it introduces the PM-ViT module, which integrates Vision Mamba with separable self-attention, significantly enhancing feature extraction efficiency and multi-scale contextual integration capabilities in complex visual tasks. Experimental results demonstrate that our approach achieves Top-1 recognition accuracies of 91.02%, 91.21%, and 73.27% on the ETHZ Food-101, Vireo Food-172, and UEC Food-256 benchmark datasets, respectively, confirming the superior performance of this architecture in food image classification tasks.

For future research, we plan to:

1. Design more efficient feature fusion and interaction algorithms to further reduce computational complexity while enhancing model representation capabilities;
2. Deeply investigate the complementary fusion mechanisms between the Mamba sequence modeling paradigm and efficient convolutional structures to improve the model's ability to parse complex textures and spatial structures in food images;
3. Extend the DPFA framework to practical application scenarios, such as developing mobile-based intelligent food recognition and nutritional analysis systems to provide personalized dietary recommendations.

These research directions not only hold significant theoretical value but also exhibit substantial application potential and societal benefits in the context of the widespread adoption of smart devices.

6 Declaration

6.1 Funding Declaration

The authors did not receive support from any organization for the submitted work.

References

[1] Kawano Y, Yanai K (2015) Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools Appl* 74(14):5263–5287. https://doi.org/10.1007/s11042-015-2685-2

[2] Ishino A, Yamakata Y, Karasawa H, Aizawa K (2021) RecipeLog: Recipe authoring app for accurate food recording. In: Proceedings of the 29th ACM International Conference on Multimedia, pp 2798–2800. https://doi.org/10.1145/3474085.3478563

[3] Min W, Wang Z, Liu Y, Luo M, Kang L, Wei X, Wei X, Jiang S (2023) Large scale visual food recognition. *IEEE Trans Pattern Anal Mach Intell* 45(8):9932–9949.

https://doi.org/10.1109/TPAMI.2023.3267030

[4] Rostami A, Nagesh N, Rahmani A, Jain R (2022) World food atlas for food navigation. In: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, pp 39–47. https://doi.org/10.1145/3552484.3555748

[5] Rostami A, Pandey V, Nag N, Wang V, Jain R (2020) Personal food model. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 4416–4424. https://doi.org/10.1145/3394171.3413769

[6] Nakamoto K, Amano S, Karasawa H, Yamakata Y, Aizawa K (2022) Prediction of mental state from food images. In: Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and related APPlications, pp 21–28. https://doi.org/10.1145/3552485.3554937

[7] Lo FPW, Sun Y, Qiu J, Lo B (2020) Image-based food classification and volume estimation for dietary assessment: A review. *IEEE J Biomed Health Inform* 24(7):1926–1939. https://doi.org/10.1109/JBHI.2020.2972068 .

[8] Yamakata Y, Ishino A, Sunto A, Amano S, Aizawa K (2022) Recipe-oriented food logging for nutritional management. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 6898–6904. https://doi.org/10.1145/3503161.3547957

[9] Ródenas J, Nagarajan B, Bolaños M, Radeva P (2022) Learning multi-subset of classes for fine-grained food recognition. In: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, pp 17–26. https://doi.org/10.1145/3552484.3555754

[10] Kawano Y, Yanai K (2013) Real-time mobile food recognition system. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1–7. https://doi.org/10.1109/CVPRW.2013.5

[11] Kawano Y, Yanai K (2015) Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools Appl* 74(14):5263–5287. https://doi.org/10.1007/s11042-014-2000-8

[12] Pouladzadeh P, Shirmohammadi S (2017) Mobile multi-food recognition using deep learning. *ACM Trans Multimedia Comput Commun Appl* 13(3s):1–21. https://doi.org/10.1145/3052432

[13] Nie W, Liu C (2023) Assessing food safety risks based on a geospatial analysis: toward a cross-regional food safety management. *J Sci Food Agric* 103(13):6654–6663. https://doi.org/10.1002/jsfa.12574

[14] Jiang S, Yan C, Tang X et al (2019) Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Trans Image Process* 29:265–276. https://doi.

org/10.1109/TIP.2019.2942344

[15] Kagaya H, Aizawa K, Ogawa M (2014) Food detection and recognition using convolutional neural network. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp 1085–1088. https://doi.org/10.1145/2647868.2654932

[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://arxiv.org/abs/2010.11929

[17] Gu A, Dao T (2023) Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752

[18] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90

[19] Ma N, Zhang X, Zheng H-T, Sun J (2018) Shufflenet v2: Practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 116–131. https://doi.org/10.1007/978-3-030-01246-5_10

[20] Mehta S, Rastegari M, Shapiro L, Hajishirzi H (2019) ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9190–9200. https://doi.org/10.1109/CVPR.2019.00938

[21] Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, pp 6105–6114. PMLR. http://proceedings.mlr.press/v97/tan19a.html

[22] Saeta B, Shabalin D (2021) Swift for TensorFlow: A portable, flexible platform for deep learning. *Proc Mach Learn Syst* 3:240–254. https://doi.org/10.1145/3504999

[23] Qin D, Leichner C, Delakis M, Fornoni M, Luo S, Yang F, Wang W, Banbury C, Ye C, Akin B et al (2024) MobileNetV4: Universal models for the mobile ecosystem. In: Proceedings of the European Conference on Computer Vision, pp 78–96. Springer. https://doi.org/10.1007/978-3-031-30000-1_5

[24] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00984

[25] Li Y, Yuan G, Wen Y, Hu J, Evangelidis G, Tulyakov S, Wang Y, Ren J (2022) Efficientformer: Vision transformers at MobileNet speed. *Adv Neural Inf Process Syst* 35:12934–12949

[26] Huang T, Huang L, You S, Wang F, Qian C, Xu C (2022) LightVit: Towards light-weight convolution-free vision transformers. arXiv preprint arXiv:2207.05557

[27] Liu X, Peng H, Zheng N, Yang Y, Hu H, Yuan Y (2023) EfficientViT: Memory efficient vision transformer with cascaded group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14420–14430. https://doi.org/10.1109/CVPR52688.2023.01419

[28] Zhang J, Peng H, Wu K, Liu M, Xiao B, Fu J, Yuan L (2022) MiniViT: Compressing vision transformers with weight multiplexing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12145–12154. https://doi.org/10.1109/CVPR52688.2022.01179

[29] Wu K, Zhang J, Peng H, Liu M, Xiao B, Fu J, Yuan L (2022) TinyViT: Fast pretraining distillation for small vision transformers. In: Proceedings of the European Conference on Computer Vision, pp 68–85. Springer. https://doi.org/10.1007/978-3-031-19774-9_5

[30] Xu G, Jia W, Wu T, Chen L, Gao G (2024) HaFormer: Unleashing the power of hierarchy-aware features for lightweight semantic segmentation. *IEEE Trans Image Process*. https://doi.org/10.1109/TIP.2024.1234567

[31] Chen Y, Dai X, Chen D, Liu M, Dong X, Yuan L, Liu Z (2022) Mobile-Former: Bridging MobileNet and Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5270–5279. https://doi.org/10.1109/CVPR52688.2022.00519

[32] Guo J, Han K, Wu H, Tang Y, Chen X, Wang Y, Xu C (2022) CMT: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12175–12185. https://doi.org/10.1109/CVPR52688.2022.01210

[33] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L (2021) CvT: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 22–31. https://doi.org/10.1109/ICCV48922.2021.00010

[34] Srinivas A, Lin T-Y, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16519–16529. https://doi.org/10.1109/CVPR46437.2021.01637

[35] Li J, Xia X, Li W, Li H, Wang X, Xiao X, Wang R, Zheng M, Pan X (2022) Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios. arXiv preprint arXiv:2207.05501

[36] Pan J, Bulat A, Tan F, Zhu X, Dudziak L, Li H, Tzimiropoulos G, Martinez B (2022) EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers. In: Proceedings of the European Conference on Computer Vision, pp 294–311. Springer. https://doi.org/10.1007/978-3-031-19775-6_17

[37] Mehta S, Rastegari M (2021) MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178. https://doi.org/10.48550/arXiv.2110.02178

[38] Mehta S, Rastegari M (2022) Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680. https://doi.org/10.48550/arXiv.2206.02680

[39] Liu X, Zhang C, Zhang L (2024) Vision Mamba: A comprehensive survey and taxonomy. arXiv preprint arXiv:2405.04404. https://doi.org/10.48550/arXiv.2405.04404

[40] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y. (2024). Vmamba: Visual state space model. *Advances in Neural Information Processing Systems*, 37, 103031–103063.

[41] Hatamizadeh A, Kautz J (2025) MambaVision: A hybrid Mamba-Transformer vision backbone. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 25261–25270. https://doi.org/10.48550/arXiv.2407.08083

[42] Yu W, Wang X (2025) MambaOut: Do we really need Mamba for vision? In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 4484–4496. https://doi.org/10.1109/CVPR52525.2025.00448

[43] Han D, Wang Z, Xia Z, Han Y, Pu Y, Ge C, Song J, Song S, Zheng B, Huang G (2024) Demystify Mamba in vision: A linear attention perspective. Adv Neural Inf Process Syst 37:127181–127203. https://doi.org/10.48550/arXiv.2403.09374

[44] Sheng G, Min W, Zhu X, Xu L, Sun Q, Yang Y, Wang L, Jiang S (2024) A lightweight hybrid model with location-preserving ViT for efficient food recognition. Nutrients 16(2):200. https://doi.org/10.3390/nu16020200

[45] Sheng G, Sun S, Liu C, Yang Y (2022) Food recognition via an efficient neural network with transformer grouping. Int J Intell Syst 37(12):11465–11481. https://doi.org/10.1002/int.22995

[46] Yang Y, Min W, Song J, Sheng G, Wang L, Jiang S (2024) Lightweight food recognition via aggregation block and feature encoding. ACM Trans Multimedia Comput Commun Appl 20(10):1–25. https://doi.org/10.1145/3609266

[47] Bossard L, Guillaumin M, Van Gool L (2014) Food-101—mining discriminative components with random forests. In: European Conference on Computer Vision,

pp 446–461. https://doi.org/10.1007/978-3-319-10593-2_28

[48] Klasson M, Zhang C, Kjellström H (2019) A hierarchical grocery store image dataset with visual and semantic labels. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 491–500. https://doi.org/10.1109/WACV.2019.00057

[49] Kawano Y, Yanai K (2014) Foodcam-256: a large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp 761–762. https://doi.org/10.1145/2647868.2654943

[50] Liu R, Mi L, Chen Z (2020) AFNet: adaptive fusion network for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens 59(9):7871–7886. https://doi.org/10.1109/TGRS.2020.2982740

[51] Tang Y, Han K, Guo J, Xu C, Xu C, Wang Y (2022) GhostNetv2: enhance cheap operation with long-range attention. Adv Neural Inf Process Syst 35:9969–9982. https://doi.org/10.48550/arXiv.2207.01797

[52] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 618–626. https://doi.org/10.1109/ICCV.2017.74

[53] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.