Lightweight Food Image Recognition With Global Shuffle Convolution

Guorui Sheng, Weiqing Min, *Senior Member, IEEE*, Tao Yao, Jingru Song, Yancun Yang, Lili Wang, and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Consumer behaviors and habits in food choices impact their physical health and have implications for climate change and global warming. Efficient food image recognition can assist individuals in making more environmentally friendly and healthier dietary choices using end devices, such as smartphones. Simultaneously, it can enhance the efficiency of server-side training, thereby reducing carbon emissions. We propose a lightweight deep neural network named Global Shuffle Net (GSNet) that can efficiently recognize food images. In GSNet, we develop a novel convolution method called global shuffle convolution, which captures the dependence between long-range pixels. Merging global shuffle convolution with classic local convolution yields a framework that works as the backbone of GSNet. Through GSNet's ability to capture the dependence between long-range pixels at the start of the network, by restricting the number of layers in the middle and rear, the parameters and floating operation operations (FLOPs) can be minimized without compromising the performance, thus permitting a lightweight goal to be achieved. Experimental results on four popular food recognition datasets demonstrate that our approach achieves state-of-the-art performance with higher accuracy and fewer FLOPs and parameters. For example, in comparison to the current state-of-the-art model of MobileViTv2, GSNet achieved 87.9% accuracy of the top-1 level on the Eidgenössische Technische Hochschule Zürich (ETHZ) Food-101 dataset with 28% reduction in the parameters, 37% reduction in the FLOPs, but a 0.7% more accuracy.**

***Index Terms*—Climate change and global warming, deep learning, food recognition, global shuffle convolution, lightweight, long-range dependence.**

I. Introduction

Climate change and global warming have exhibited an alarming escalation in recent years, prompting growing awareness of the impact of dietary choices on the environment among the global population of 7.7 billion people [1], [2]. Increasing numbers of consumers recognize that adopting eco-friendly and sustainable food options can contribute significantly to mitigating these issues at an individual level. Moreover, such choices drive producers and supply chains to embrace more environmentally friendly practices. For instance, reducing meat consumption can lower the greenhouse gas emissions associated with livestock farming, while prioritizing local and seasonal foods helps minimize carbon emissions during transportation. Efficient food image recognition plays a pivotal role as the initial step in empowering individuals to make such sustainable choices. Accurate dietary recommendations derived from this recognition not only assist consumers in selecting environmentally friendly foods but also aid in choosing those that promote personal health. This capability can readily be harnessed through the smartphones that people carry with them daily. However, given the constraints in power consumption and memory of such end devices, it is imperative to optimize the neural network utilized for food recognition.

Food image recognition occupies a pivotal position within the rapidly evolving interdisciplinary realm of food computing [3], playing an indispensable role across various domains, such as dietary analysis, healthcare, and the food industry [4], [5], [6], [7], [8], [9]. The proliferation of diverse cuisines and culinary techniques has led to a surge in food image datasets, posing challenges for sustainable expansion of server-side food image recognition. Moreover, the substantial carbon footprint resulting from large-scale training of artificial intelligence on server infrastructure has emerged as a pressing concern. Furthermore, food image recognition entails intricate fine-grained analysis, offering valuable insights for refining similar models in the domain of fine-grained recognition [10]. Despite the widespread adoption of deep learning methods in current approaches, characterized by their high parameter count and extensive training and inference durations [11], [12], this article focuses on developing lightweight deep neural network models tailored specifically for food image recognition.

The rapid integration of artificial intelligence, particularly deep learning, has permeated various sectors, including food and agriculture [13], [14], [15], [16], [58]. However, research on lightweight approaches for food image recognition remains relatively sparse. Early endeavors primarily relied on lightweight convolutional neural network (CNN)-based methods for food image analysis. However, the inherent challenge lay in extracting long-range information from images due to the dispersed nature of ingredients. As illustrated in Fig. 1, the discriminative factors in food identification often lie within the scattered arrangement of ingredients, compounded by variations in size, shape, and

Manuscript received 30 December 2023; revised 20 February 2024; accepted 6 April 2024. This article was recommended by Associate Editor C. Josephson. *(Corresponding author: Yancun Yang.)*

Guorui Sheng, Tao Yao, Jingru Song, Yancun Yang, and Lili Wang are with the Department of Information and Electrical Engineering, Ludong University, Yantai 264025, China (e-mail: shengguorui@ldu.edu.cn; yaotao@ldu.edu.cn; songjingru@m.ldu.edu.cn; Harryyang@ldu.edu.cn; wanglili@ldu.edu.cn).

Weiqing Min and Shuqiang Jiang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: minweiqing@ict.ac.cn; sqjiang@ict.ac.cn).

Digital Object Identifier 10.1109/TAFE.2024.3386713

Fig. 1. Some samples from ETHZ Food-101 [41] and Vireo Food-172 [51]. Ingredients are scattered throughout the food image.

distribution arising from different cooking methods. Capturing these long-range relationships amidst scattered food images is crucial for accurate dish recognition.

While vision transformer (ViT) excels in capturing global information by leveraging attention mechanisms, its computational demands and training complexity pose significant hurdles [17]. To reconcile this, efforts, such as those by Sheng et al. [18], have attempted to amalgamate ViT's global representation capabilities with CNN's local feature extraction prowess. Nonetheless, the resultant models still entail considerable parameter counts and computational overheads.

The challenges in lightweight food image recognition are twofold. First, the scattered distribution of ingredients necessitates a nuanced understanding of long-range pixel correlations crucial for accurate recognition. However, conventional CNN architectures excel at capturing local features, requiring increasingly complex networks to model distant pixel relationships, thus contravening lightweight design principles. Second, while ViT offers a promising avenue for extracting long-range correlations, the quadratic increase in token interactions necessitates extensive computational resources and data for training, making adherence to lightweight constraints challenging.

Our work has addressed key challenges in lightweight food recognition, namely, the limited expression of long-range information by CNNs and the complexity of training ViT models. We employ global shuffle convolution to capture dispersed food ingredients' long-range information within food images, facilitating comprehensive global expression alongside local convolution. This parallel block serves as the foundational structure of Global Shuffle Net (GSNet), markedly enhancing food image recognition accuracy. In addition, recognizing GSNet's emphasis on extracting long-range features in the early stages, we significantly reduce network layers in the intermediate and posterior sections to minimize parameter count and computational complexity. We design GSNet and conduct extensive experiments across various prominent food image databases, demonstrating superior recognition performance compared with existing CNN-based, ViT-based, and hybrid lightweight networks. As illustrated in Fig. 2, GSNet surpasses several widely used lightweight CNN and ViT models renowned for their state-of-the-art (SOTA) performance, such as MobileNetV2 [19], MobileNetV3 [20], and MobileViTv2 [21]. Notably, GSNet achieves an 88.4% top-1 accuracy with only 3.1 M parameters, significantly outperforming MobileNetV3 (86.2%) despite having fewer parameters (4.3 M).

We summarize our contributions as follows.

1) We design a simple, effective, and easy-to-implement pure convolutional model to capture the dependencies between remote pixels on the food image plane to effectively handle the dispersed distribution of ingredients in food images. Simultaneously extracting short-range features and long-range features through a parallel structure effectively improves the accuracy of food image recognition.
2) Based on the fact that the model is dedicated to capturing dependence between long-range pixels at the front of the network, we redesigned a new lightweight neural network that adapts to this feature and effectively reduces the number of parameters and calculations.
3) We conducted extensive and comprehensive experiments on four major food image datasets, and the results indicate that our approach achieves SOTA performance with higher accuracy and fewer floating operation operations (FLOPs) and parameters, outperforming SOTA CNN-based, ViT-based, and hybrid lightweight models.

II. Related Works

A. Lightweight CNNs, ViTs, and Hybrid Models

ResNet [22] is one of the most successful CNN architectures. However, the best-performing CNN models are usually high in parameters and FLOPs. Lightweight CNNs that achieve competitive performance with fewer parameters and FLOPs include ShuffleNetV2 [23], ESPNetV2 [24], EfficientNet [25], MobileNetV2, [19] and MobileNetV3 [20]. MobileNetV3 [20] belongs to the category of models developed specifically for resource-constrained environments, such as mobile devices. The basic blocks of MobileNetV3 [3] include the MobileNetV2 [19] block and the squeeze-and-excite network [26]. The common problem of CNN-based lightweight models is their weak ability to extract global information.

In order to extract global information more efficiently, ViT brings transformer models for natural language processing tasks to the vision domain, especially image recognition. The extensive use of ViT in the field of machine vision has also attracted some research on its lightweight. Most efforts have been focused on improving the self-attention process to increase efficiency, such as SwinT [27], EfficientFormer [28], LightViT [29], EfficientViT [30], MiniViT [31], and TinyViT [32].

The common problems of ViT-based lightweight models are the difficulty of training and the high computational cost due to the quadratic number of interactions between tokens. Recently, some researchers have tried to construct compact hybrid models that integrate CNN and ViT for mobile vision tasks, which shows that combining convolution and transformer achieves improvement in prediction accuracy as well as training stability. Subsequently, there have been a large number of lightweight works on these models, such as MobileFormer [33], CMT [34], CvT [35], BoTNet [36], Next-ViT [38], EdgeViTs [38], MobileViTv1 [39], and MobileViTv2 [21]. The hybrid lightweight model based on CNN and ViT has done a good fusion in extracting global information and local information, but there is still the problem of large model size.

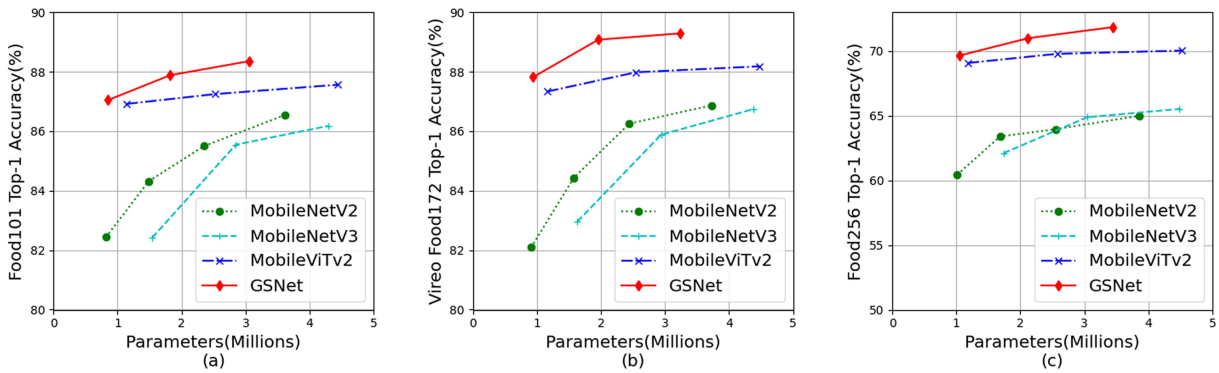

Fig. 2. Comparison with SOTA CNN-based (MobileNetV2 [19] & V3 [20]) and Hybrid (MobileViTv2 [21]) lightweight models across different datasets. (a): ETHZ Food-101 [41]. (b): Vireo Food-172 [51]; (c): UEC Food256 [52].

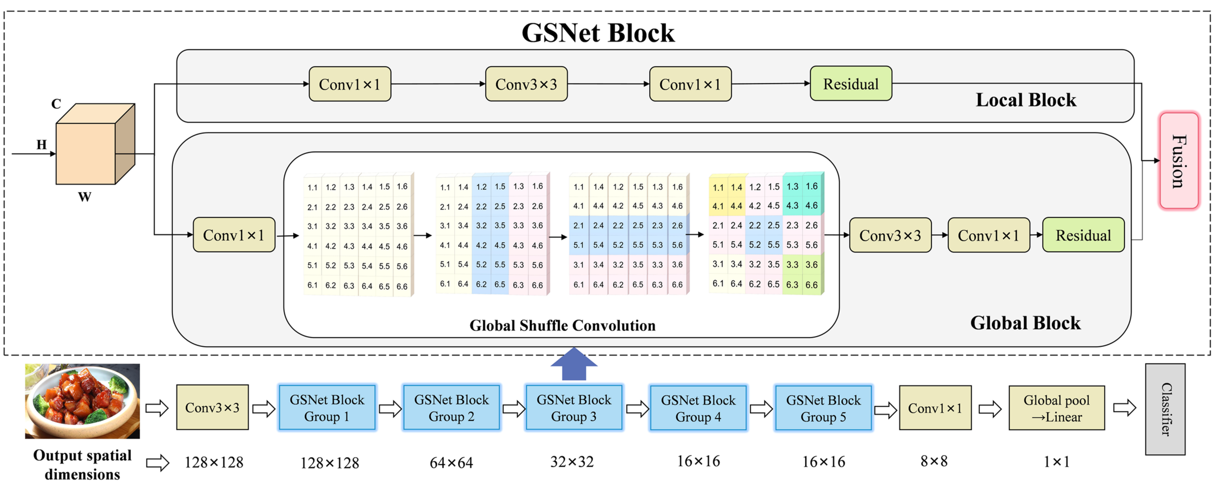

Fig. 3. GSNet. Here, Conv $n \times n$ in the GSNet represents a standard $n \times n$ convolution. In global shuffle convolution block, to illustrate the implementation process, assume that both H and W are 6.

B. Lightweight Food Recognition

Recently, Min et al. [3] gave a survey on food computing including food recognition. In earlier years, various handcrafted features were utilized for recognition [40], [41]. For example, Mehta and Rastegari [39] utilized random forests to mine discriminative image patches as a visual representation. Due to the rise of deep learning technology, many recognition methods based on deep learning have emerged [11], [12], [42], [43], [44], [45].

Given the necessity of lightweight food image recognition, a lot of related research work has been proposed. Early researchers used the lightweight CNN method for food image recognition [46], [47], [48], [49]. Tan et al. [49] recently proposed a novel lightweight neural architecture search (LNAS) model to self-generate a thin CNN that can be executed on mobile devices, achieving nearly 76% recognition accuracy on the Eidgenössische Technische Hochschule Zürich (ETHZ) Food-101 dataset. The recognition accuracy of these CNN-based lightweight food recognition is generally low. ViT provides a new option for extracting global features of food images, Sheng et al. [18] tried to extract global and local features with a parallel structure composed of the ViT group and CNN and obtained the SOTA performance. However, due to the multihead attention mechanism of the ViT, the model size is still large.

In contrast to the works that use ViT, we propose a simple yet effective pure convolution network, which is based upon the characteristics of food images and allows for better control over parameters and calculations. In this architecture, a global shuffle convolution is utilized to identify global features and a parallel network structure along with CNN is fashioned to draw out local features, resulting in SOTA performance.

III. Method

A. Brief Review of GSNet

Our objective is to propose a network model that can not only effectively deal with the dispersion and diversity of food image features, but also realize lightweight so that it can be better extended on the server side and deployed on edge and end devices.

The proposed GSNet is shown in Fig. 3. We use global shuffle convolution to capture the long-range information of food ingredients scattered in food images to enhance the model's expressiveness, and then form a parallel block with local convolution. This parallel block is used as the basic structure of GSNet,

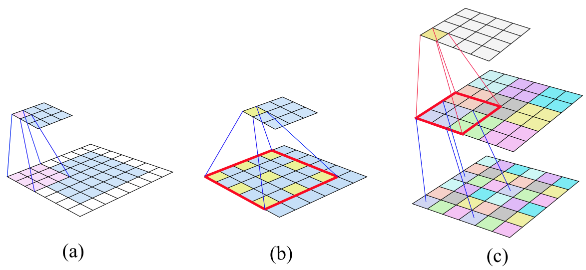

Fig. 4. (a) Local convolution. (b) Dilated convolution. (c) Global shuffle convolution.

which effectively improves the food image recognition accuracy. Based on the fact that GSNet focuses on capturing long-range dependence among different spatial pixels in the front part of the model, we reduce the number of network layers in the middle and rear parts, and correspondingly the effective reduction of the number of parameters and FLOPs is obtained. The experimental results show that this strategy can effectively reduce the number of parameters and FLOPs on the premise of ensuring recognition accuracy.

B. Global Shuffle Convolution

The global shuffle convolution method divides the image into several patches first, and then in each convolution operation, corresponding position pixels are taken out from each patch to participate in the convolution. Since patches cover the entire image, this convolution operation is to extract scattered correlation information. No matter how far the same ingredient is in the dish, it can be captured by the global shuffle convolution operation. As shown in Figs. 3 and 4(c), by first resetting the row and then resetting the column, the distant pixels are concentrated into 2×2 patches, and then a normal convolution with a kernel of 3×3 size is performed, so that not only four elements are involved in the calculation of correlation pixel, but five more elements from a greater distance. Through this, the correlation information between long-range pixels is quickly obtained. Here, the convolution kernel size is set to 3×3 and stride set to 1. In Fig. 4(c), the middle image is the intermediate result after the rows and columns of the bottom image are reset, and the spatial variation law can be seen through pixels of the same color. The top plane is the result of a 3×3 convolution of the middle plane.

Compared with global shuffle convolution, local convolution [see Fig. 4(a)] extracts the local correlation in the image through the convolution operation on the local area, and then translates it with a certain step size and performs the convolution operation multiple times to achieve full coverage of local information. It can build deeper and more nonlinear networks but ignores the correlation between pixel vectors in the global scope, leading to information loss compared with the fully connected model. Dilated convolution is a variant of local convolution that expands the receptive field compared with local convolution. As shown in Fig. 4(b), under the same kernel size, the dilated convolution skips some pixel positions to perform convolution operations, thus having a larger local field of view [50]. Dilated convolution can express a broader range of local correlations, but also due to the operation of ignoring certain pixels, some information is lost. On the other hand, although dilated convolution can extract long-range related information at different distances by adjusting the dilated rate, ultra-long-distance related information needs to be obtained by stacking more layers, so it is not as efficient as global shuffle convolution to extract global features. Summarily, global shuffle convolution is more appropriate for image recognition of food because of its excellent ability to capture comprehensive long-range correlation information.

C. Approach

Suppose the input size of a convolution layer is $[N, C^{\text{in}}, IH^{\text{in}}, IW^{\text{in}}]$, the output size is $[N, C^{\text{out}}, IH^{\text{out}}, IW^{\text{out}}]$, and convolution kernel size is $[C^{\text{out}}, C^{\text{in}}, K^H, K^W]$, where N denotes the batch size, C denotes the number of channels, IH and IW denotes the height and width of input images, respectively. During local convolution calculation, the value of a specific feature point $X^{(t+1)}$ of a specific channel of the output feature map with input $X^{(t)}$ is calculated as follows:

$$\begin{aligned} X^{(t+1)}_{N_i, C^{\text{out}}_j, IH^{\text{out}}_k, IW^{\text{out}}_l} &= B_{C^{\text{out}}_j} \\ &+ \sum_h \sum_w \sum_{0 \le c < C^{in}} W_{C^{\text{out}}_j, c, h, w} \times X^{(t)}_{N_i, c, h, w} \end{aligned} \tag{1}$$

where B denotes bias parameters with size C^{out}, W denotes weight parameters with size $[C^{\text{out}}, C^{\text{in}}, K^H, K^W]$, $h \in [IH^{\text{out}}_k, IH^{\text{out}}_k + K^H - 1]$, and $l \in [IW^{\text{out}}_l, IW^{\text{out}}_l + K^W - 1]$.

The complete formula of the global shuffle convolution calculation method is more complicated. For simplicity, in an image plane, let the number of groups in the row direction be equal to K^H, the number of groups in the column direction is K^W, and $H^{\text{in}}/K^H = H^{\text{out}}, W^{\text{in}}/K^W = W^{\text{out}}$, that is, the number of groups is consistent with the kernel size, and the size of each group is the same as the output plane, and the stride of the convolution is taken as the group size. Then, the calculation of the value of the specified feature point Y of the output feature map with input X is similar to formula (1) but

$$\begin{aligned} &h \in \left\{ IH^{\text{out}}_k + k * IH^{\text{out}} \,\middle|\, k = 0, \ldots, K^H - 1 \right\} \\ &w \in \left\{ IW^{\text{out}}_l + k * IW^{\text{out}} \,\middle|\, k = 0, \ldots, K^W - 1 \right\}. \end{aligned} \tag{2}$$

Then, fold the 4-D matrix into a 2-D matrix for $X^{(t)}$, combine bias parameters into weight parameters, and add a constant row to $X^{(t)}$, the output of the local convolutional network

$$\begin{aligned} &f(X^{(0)}) = f^{(T-1)}(\cdots f^{(1)} \\ &(f^{(0)}(X^{(0)}W^{(0)})W^{(1)}) \cdots W^{(T-1)}) \end{aligned} \tag{3}$$

where $X^{(t)}(t = 0, \ldots, T-1)$ is the 2-D matrix of the input or 2-D matrix of the output layer, T is the number of layers, $W^{(t)}$ is the parameter matrix of each layer, $f^{(t)}$ is the nonlinear activation function used by each layer. When nonlinear activation functions are not used: $f(X^{(0)}) = X^{(0)}W^{(0)}W^{(1)} \cdots W^{(T-1)}$.

For global shuffle convolution, the output of the network

$$f(X^{(0)}) = f^{(T-1)}(\cdots f^{(1)}(f^{(0)} (X^{(0)}M^{(0)}W_g^{(0)})M^{(1)}W_g^{(1)})\cdots M^{(T-1)}W_g^{(T-1)}) \quad (4)$$

where $M^{(t)}$ is a linearity transformation matrix and $W_g^{(t)}$ is the parameter matrix. Likewise, when not using a nonlinear activation function

$$f(X^{(0)}) = X^{(0)}M^{(0)}W_g^{(0)}M^{(1)}W_g^{(1)}\cdots M^{(T-1)}W_g^{T-1}. \quad (5)$$

In linear mode, the difference between global shuffle convolution and local convolution can be understood from two perspectives: 1) In the global shuffle convolutional net, during the operation of each layer, the input matrix is first column-transformed, i.e., $X^{(t)} \cdot M^{(t)}$, and then multiply it with the parameter matrix $W_g^{(t)}$; 2) the parameter matrix of the global shuffle convolution correspond to the parameter matrix of the local convolution, namely

$$M^{(t)} \cdot W_g^{(t)} \leftrightarrow W^{(t)} \quad (6)$$

that is, in linear mode, global shuffle convolution and local convolution are equivalent, but their parameter positions are adjusted. However, neural networks are nonlinear, only (1) is true, i.e., the global shuffle convolution is a series of images whose plane pixels are misaligned (the misalignment pattern is fixed). Using only global shuffle convolutions in the network is generally ineffective unless the dislocation results in clustered color patches similar to normal images.

In our work, the parallel network structure of global shuffle convolution and local convolution are both used, the local convolution represents the most features of the image, and the global shuffle convolution assists in the collection of food information scattered around the image, which ultimately improves the accuracy of recognition.

D. Implementation Details

The essence of the global shuffle convolution method is to calculate the correlation between several pixels at any distance on the image plane, that is, to convolve several pixel vectors selected at different positions in the entire image, which is equivalent to adjusting these several pixel vectors to a local region, and then perform local convolution on this region.

When implementing the global shuffle convolution calculation, our actual practice is to first perform the relocation in the column direction of the image plane, then perform the rearrangement in the row direction, and finally perform the local convolution. As shown in Fig. 3, after the rearrangement in column and row directions, several pixels (2, 2), (2, 5), (5, 2), and (5, 5) are adjusted to be adjacent to each other, these pixels come from the scattered positions of the entire image plane, and local convolution on them is equivalent to global shuffle convolution.

This implementation achieves certain flexibility: the number of groups does not have to be the same as the size of the convolution kernel, and the stride of the convolution does not have to be the same as the size of the group so that correlations between more complex plane pixel vectors at different positions can be represented.

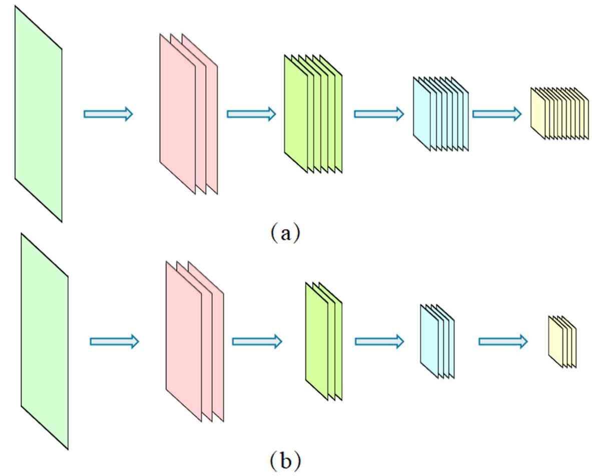

Fig. 5. Hierarchical network layout. (a) Hierarchical layout of traditional neural networks. (b) By drastically reducing the number of layers in the back of the network, the hierarchical network layout adopted by GSNet effectively reduces the number of parameters and computation.

E. Network Architecture

This section introduces the basic parallel block, the hierarchical network layout, and the detailed network architecture of GSNet.

Parallel block: The basic block used in our network is parallel, one branch uses local convolution and the other uses global shuffle convolution, the outputs of the two branches are concatenated and then propagated along the neural network. The local convolution branch is the inverse residual model derived from MobileNetV2, the other branch replaces the depth-wise convolution part of it with our proposed global shuffle convolution. In parallel block, the local convolution branch is responsible for extracting the local features at the pixel level, and the global shuffle convolution branch is responsible for capturing the long-range dependence between pixels in different spatial locations. The local convolution branch is the main bearer since most of the image features are revealed through local correlations. The global shuffle convolution branch provides correlation features between pixels from entire image plane and is used to add long-range feature to improve the expressiveness of image features.

Adjusted network layout: As shown in Fig. 5, we use a network structure that differs from traditional models. In the traditional network, since the local convolution can only represent the local correlation of the image, the long-range features are obtained after the multilayer local convolution. The characteristic of this network structure is that there are fewer layers in the front part of the network and more layers in the back part. Due to the large number of channels in the middle and rear, the network is heavily parameterized. By using the parallel block structure, GSNet obtains the long-range correlation information at the beginning of the network without relying on the shrinking part

TABLE I
NETWORK SPECIFICATION

Component	Input	Operator	Exp Ratio	GR×GC	Out	Stride
Head	$256 \times 256 \times 3$	Conv2D 3×3	-	-	16	2
Block group 1	$128 \times 128 \times 16$	PB 3×3	1	16×16	16	1
Block group 2	$128 \times 128 \times 16$	PB 3×3	1	16×16	24	2
	$64 \times 64 \times 24$	PB 3×3	3	8×8	24	1
Block group 3	$64 \times 64 \times 24$	PB 3×3	3	8×8	40	2
	$32 \times 32 \times 40$	PB 3×3	3	8×8	40	1
	$32 \times 32 \times 40$	PB 3×3	3	8×8	40	1
Block group 4	$32 \times 32 \times 40$	PB 3×3	6	8×8	80	2
	$16 \times 16 \times 80$	PB 3×3	2.5	4×4	80	1
	$16 \times 16 \times 80$	PB 3×3	2.3	4×4	80	1
Block group 5	$16 \times 16 \times 80$	PB 3×3	6	4×4	160	2
Tail	$8 \times 8 \times 160$	Conv2D 1×1	-	-	1280	1
Classifier	$8 \times 8 \times 1280$	Global Avg Pool	-	-	1280	1
	$1 \times 1 \times 1280$	Dropout (0.2)	-	-	1280	-
	$1 \times 1 \times 1280$	Linear	-	-	n classes	-

PB: Parallel Block of GSNet; Exp Ratio: Expansion Ratio in MobileNetV2 [19] block; GR×GC: GR and GC means the number of groups in the row direction and column direction of the feature map when doing global shuffle convolution.

in the rear of the network with more layers. Therefore, in this work, we reduce the number of layers in the back of the network drastically, effectively reduce the number of parameters and computation, and then develop a lightweight food recognition network. The following experimental results show that this strategy is effective, reducing the number of parameters and computations while achieving higher accuracy.

Network specification: The detailed network specification is given in Table I. The network first obtains a image plane through a local convolution and then passes through a series of parallel block groups. In each parallel block group, the group number is set according to the size of the current image resolution. At the tail of the network, the number of channels is expanded by convolution, then global pooling and dropout are performed to obtain and adjust the single-pixel output, and finally, a fully connected layer is used to map to the number of classes.

IV. EXPERIMENTS

A. Datasets

To evaluate the proposed model, we conduct experiments on four food datasets: ETHZ Food-101 [41], Vireo Food-172 [51], UEC Food-256 [52], and ISIA Food-500 [53]. ETHZ Food-101 has 101 categories, we use 75 750 images for training and 25 250 for validation. Vireo Food-172 provides 172 categories, we use 66 071 images for training and 44 170 images for validation. UEC Food-256 has 256 categories where 22 095 images are used for training and 9300 images are used for validation. ISIA Food-500 is a comprehensive food dataset composed of 500 food types from Wikipedia, we use 239 378 images for training and 120 142 images for validation.

B. Training Settings

We train our models using an input image resolution 256×256, a batch size of 256, and SGD optimizer with 0.9 momentum [54]. We use the initial learning rate of 0.1 for first 3000 iterations of linear warm-up and then a cosine schedule with the learning rate ranging from 0.0004 to 0.8. Furthermore, we use the same data augmentation method as MobileViTv2 for image preprocessing.

TABLE II
PERFORMANCE COMPARISON ON ETHZ FOOD-101[41]

Method	Top-1 Acc.	#Params	#FLOPs
MobileNetV2 -1.25 [19]	86.5%	3.6M	496.5M
MobileNetV3 -1.0 [20]	86.2%	4.3M	**218.9M**
MobileViTv2 -1.0 [21]	87.6%	4.4M	1843.4M
GSNet -2.0	**88.4%**	**3.1M**	1051.2M
MobileNetV2 -1.0 [19]	85.5%	2.4M	313.0M
MobileNetV3 -0.75 [20]	85.5%	2.8M	**161.9M**
MobileViTv2 -0.75 [21]	87.2%	2.5M	1051.4M
GSNet -1.5	**87.9%**	**1.8M**	665.3M
MobileNetV2 -0.5 [19]	82.4%	**0.8M**	112.9M
MobileNetV3 -0.5 [20]	82.4%	1.5M	**73.3M**
MobileViTv2 -0.5 [21]	86.9%	1.1M	480.2M
GSNet -1.0	**87.0%**	0.9M	295.0M
LNAS-NET [49]	75.9%	1.8M	-
LTBDNN(TD-192) [18]	76.8%	12.2M	-
GSNet -1.0	**87.0%**	**0.9M**	**295.0M**

GSNet -x: x denotes width multiplier on the base model.

The bolded terms in each column represent: highest accuracy, minimal parameter count, and minimal computational workload within each group, respectively.

C. Experiment Results

Results on ETHZ Food-101: Table II presents results on ETHZ Food-101. The results are grouped according to similar numbers of parameters. Our model surpasses all other models in three parameter ranges. Among all models with around 1 M parameters, our model achieves 87.0% top-1 accuracy, which is 0.1%, 4.6%, and 4.6% higher than MobileViTv2, MobileNetV3, and MobileNetV2, respectively. In around 2–3 M parameter budget models, our model's top-1 accuracy is 87.9%, which is 0.7% higher than MobileViTv2, and 2.4% higher than MobileNetV3 and MobileNetV2. Our model also achieves the highest top-1 accuracy of 88.4% in the parameter range of 3–5 M, surpassing MobileViTv2, MobileNetV3, and MobileNetV2 by 0.8%, 2.2%, and 1.9%, respectively. We also compare with recent lightweight

TABLE III
PERFORMANCE COMPARISON ON VIREOFOOD-172[51]

Method	Top-1 Acc.	#Params	#FLOPs
MobileNetV2 -1.25 [19]	86.9%	3.7M	496.7M
MobileNetV3 -1.0 [20]	86.7%	4.4M	**219.0M**
MobileViTv2 -1.0 [19]	88.2%	4.5M	1843.4M
GSNet -2.0	**89.3%**	**3.2M**	1051.4M
MobileNetV2 -1.0 [19]	86.3%	2.4M	313.1M
MobileNetV3 -0.75 [20]	85.9%	3.0M	**162.0M**
MobileViTv2 -0.75 [21]	88.0%	2.5M	1051.5M
GSNet -1.5	**89.1%**	**2.0M**	665.5M
MobileNetV2 -0.5 [19]	82.1%	**0.9M**	113.0M
MobileNetV3 -0.5 [20]	83.0%	1.6M	**73.4M**
MobileViTv2 -0.5 [21]	87.3%	1.2M	480.2M
GSNet -1.0	**87.8%**	**0.9M**	295.0M

GSNet-x: x denotes width multiplier on the base model.
The bolded terms in each column represent: highest accuracy, minimal parameter count, and minimal computational workload within each group, respectively.

TABLE IV
PERFORMANCE COMPARISON ON UEC FOOD256[52]

Method	Top-1 Acc.	#Params	#FLOPs
MobileNetV2 -1.25 [19]	65.0%	3.9M	496.8M
MobileNetV3 -1.0 [20]	65.5%	4.5M	**219.1M**
MobileViTv2 -1.0 [21]	70.0%	4.5M	1843.4M
GSNet -2.0	**71.9%**	**3.5M**	1051.6M
MobileNetV2 -1.0 [19]	64.0%	2.6M	313.2M
MobileNetV3 -0.75 [20]	64.9%	3.0M	**162.1M**
MobileViTv2 -0.75 [21]	69.8%	2.6M	1051.5M
GSNet -1.5	**71.0%**	**2.1M**	665.6M
MobileNetV2 -0.5 [19]	60.4%	**1.0M**	113.1M
MobileNetV3 -0.5 [20]	62.1%	1.7M	**73.5M**
MobileViTv2 -0.5 [21]	69.1%	1.2M	466.0M
GSNet -1.0	**69.6%**	1.1M	295.1M

GSNet -x: x denotes width multiplier on the base model.
The bolded terms in each column represent: highest accuracy, minimal parameter count, and minimal computational workload within each group, respectively.

food recognition networks; the results show that the recognition accuracy of our network (87.0%) is much higher than that of LNAS-NET (75.9%) and LTBDNN (TD-192) (76.8%) in the case of much fewer parameters.

Results on Vireo Food-172: Table III presents results on VireoFood-172. Compared with MobileViTv2 in every parameter range, our model achieves better top-1 accuracy of 87.8% versus 87.3%, 89.1% versus 88.0%, and 89.3% versus 88.2% with much lower FLOPs of 295 M versus 480 M, 665 M versus 1,052 M, and 1,051 M versus 1,843 M. Although MobileNetV3 and MobileNetV2 have much lower FLOPs, they lag in accuracy by a margin of more than 2% with our models.

Results on UEC Food256: As seen in Table IV, the results are similar to the other two datasets. Our models achieve the highest top-1 accuracy in every parameter range. Compared with MobileViTv2, our model has fewer parameters and FLOPs. Compared with MobileNetV3 and MobileNetV2, our model achieves much higher top-1 accuracy with fewer parameters but slightly more FLOPs.

TABLE V
PERFORMANCE COMPARISON ON ISIA FOOD-500[53]

Method	Top-1 Acc.	#Params	#FLOPs
MobileNetV2 -1.25 [19]	63.0%	4.2M	497.2M
MobileNetV3 -1.0 [20]	63.8%	4.8M	**219.4M**
MobileViTv2 -1.0 [21]	65.2%	4.6M	1843.6M
GSNet -2.0	**64.9%**	**4.1M**	1052.2M
MobileNetV2 -1.0 [19]	62.7%	2.9M	313.5M
MobileNetV3 -0.75 [20]	60.5%	3.4M	**162.4M**
MobileViTv2 -0.75 [21]	64.6%	2.7M	1051.6M
GSNet -1.5	**64.3%**	**2.6M**	666.1M
MobileNetV2 -0.5 [19]	57.9%	1.3M	113.4M
MobileNetV3 -0.5 [20]	58.5%	2.1M	**73.8M**
MobileViTv2 -0.5 [21]	**63.0%**	**1.2M**	480.3M
GSNet -1.0	62.0%	1.4M	295.4M

GSNet -x: x denotes width multiplier on the base model.
The bolded terms in each column represent: highest accuracy, minimal parameter count, and minimal computational workload within each group, respectively.

Results on ISIA Food-500: Table V presents experimental results on dataset ISIA Food-500. Because of its wide range, large scale, and offering of both Chinese and western food, it is harder for food recognition in Food-500. Even so, our proposed GSNet still achieves competitive results: compared with SOTA ViT-based lightweight network MobileViTv2, the FLOPs are greatly reduced with almost the same recognition rate. Compared with the SOTA CNN-based lightweight network MobileNetV2 and V3, our model has significantly better performance with similar parameters: GSNet -1.5/-2.0 obtain 64.3%/64.9% top-1 accuracy, which is +1.6%/1.1% higher than that of MobileNetv2/v3 (63.8%/62.7%) with a similar number of parameters.

The experimental results demonstrate the effectiveness and the generalization of our design. With the proposed parallel block, although we reduce the number of layers in the middle and rear parts of the network, the proposed network provides reasonable accuracy gains over the general network architecture. Considering experiments on four different food datasets with consistent results, the proposed model should be effective and efficient for general food vision tasks.

Comparison and Analysis with Results Based on Lightweight Networks using ViT: Experimental results reveal that compared with the SOTA lightweight model based on ViT, MobileViTv2 [21], GSNet achieves comparable or superior recognition accuracy while requiring fewer parameters and significantly less computational load. We believe this is based on the following reasons: ViT possesses powerful capabilities for extracting global information. However, the common challenges of ViT-based lightweight models include the difficulty of training and the high computational cost stemming from the quadratic number of interactions between tokens. When modeling the global context, ViT also incorporates positional information of patches, further increasing parameter quantity and computational load. A key differentiating feature of food images lies in the correlated characteristics among the same type of ingredients dispersed throughout the image. The distant correlations among the dispersed identical ingredients do not require consideration of specific patch positional information. Our designed GSNet is

TABLE VI
PERFORMANCE COMPARISON ON IMAGENET

Method	Top-1 Acc.	#Params	#FLOPs
MobileNetV1 -1.0 [56]	70.6%	4.2M	575M
MobileNetV2 -1.0 [19]	72.8%	4.2M	575M
MobileNetV3 -1.0 [20]	75.2%	5.4M	219M
ShuffleNetV2 -1.5 [23]	72.6%	**3.5M**	299M
GhostNetV1 -1.0 [57]	73.9%	5.2M	**141M**
MobileViTv1 -S [39]	**78.4%**	5.6M	2000M
MobileViTv2 -1.0 [21]	78.1%	4.9M	1800M
GSNet-1.0	75.3%	5.3M	1054M

GSNet-x: x denotes width multiplier on the base model.
The bolded terms in each column represent: highest accuracy, minimal parameter count, and minimal computational workload within each group, respectively.

precisely tailored to exploit this characteristic of food images. Consequently, it achieves recognition performance using CNN volume parameters and computational load that match or exceed those of ViT models.

Results on ImageNet: Table VI presents results on ImageNet-1 K. The results are grouped according to CNN-based method and ViT-based method, all with similar numbers of parameters. In the comparison with the CNN-based lightweight method, it can be found that GSNet has the same accuracy as the newly released MobileNetV3 when the number of parameters is roughly the same, but the FLOPs of GSNet are higher, which is due to the parallel structure including global shuffle convolution. Compared with ViT-based methods, taking mobileViTv2 as an example, our method has lower accuracy (75.3% versus 78.1%), but also lower FLOPs (1054 M versus 1800 M). It is found that the performance of our method on ImageNet is comparable to the SOTA CNN-based method, worse than the SOTA ViT-based model, and the overall performance is not as good as the experimental results on the food dataset. We believe this is because global shuffle convolution is more specific in dealing with the dispersed distribution of ingredients in food images since it can effectively extract correlated features between long-range pixels.

D. Qualitative Analysis and Visualization

Different from the image recognition mechanism of the traditional local convolution, the network including the global shuffle convolution tends to collect similar color patch information globally in the image plane. Fig. 6 shows the comparison by the method provided by Grad-CAM [55]: results are obtained using only local convolution and using both global shuffle convolution and local convolution. In Fig. 6, the first row is the original image, the second row is heat maps generated by using only local convolution, and the third row is heat maps generated by using local and global convolution. The following can be seen from Fig. 6.

1) Using only local convolution tends to identify locally clustered patches, which can be well-focused when they appear in food images. When the background is relatively monotonous and contains similar color blocks, the local convolution will also focus on the background incorrectly and cause recognition failure.
2) Local and global shuffle convolution tends to collect similar color patches globally, and its focal area tends to be wider than local convolution, covering multiple color patches at the same time.
3) Both models are affected if there are distinct color blocks in the background, but the local and global shuffle model is significantly less affected.

In summary above results show that the local and global shuffle model is more suitable to the scattered-color features of food images and can achieve better recognition results.

Fig. 7 illustrates cases of misrecognition by our method on the ETHZ Food-101 and Vireo Food-172 datasets. Based on the visual results reflected in the heatmaps, we analyze the reasons for recognition failures as follows. Whether employing local convolution or global convolution, both tend to extract features from prominent color blocks present in the image. Global convolution, however, can gather information on the correlation among dispersed but related color blocks in the image, thereby generating global features. Nevertheless, a characteristic of convolutional operations is their susceptibility to being drawn towards color blocks with strong color consistency, making them prone to being misled by the background and failing to focus on the target object. While global convolution may mitigate this issue to some extent by collecting information on the correlation among related color blocks globally, the impact is more significant when using local convolution alone, leading to a higher likelihood of recognition failures.

E. Ablation Study

In this section, we ablate important design elements in the proposed model using image classifications on four datasets.

Effectiveness of global shuffle convolution: Ablations of the global shuffle convolution effect on four datasets are reported in Table VII. The models with global shuffle convolution blocks obtain more higher top-1 accuracy: 69.6% (Food-256), 87.0% (Food-101), and 87.8% (Food-172) compared with models without global shuffle convolution blocks: 69.1% (Food-256), 85.5% (Food-101), and 86.5% (Food-172). That indicates the global shuffle convolution block is effective in improving models' accuracy by gathering long-range features. We also exclude local convolution blocks and train the models with only global shuffle convolution blocks. Surprisingly, they achieve top-1 accuracy: 56.0% (Food-256), 73.7% (Food-101), and 76.8% (Food-172). The results confirm that global shuffle convolution can indeed extract fairly discriminative features for food images.

Activation function: Compared with traditional networks, we make a significant reduction in parameters and computation by the strategy of reducing the number of layers. Considering that the more radical activation function could be effective in expanding the searching domain for the simpler model architecture, we use HardSwish as the activation function of all nonlinear layers. Here, we compared the effectiveness of two typical activation functions HardSwish and rectified linear unit (ReLU). Compared with ReLU with gradient values of 0 and 1, HardSwish is featured with a steep

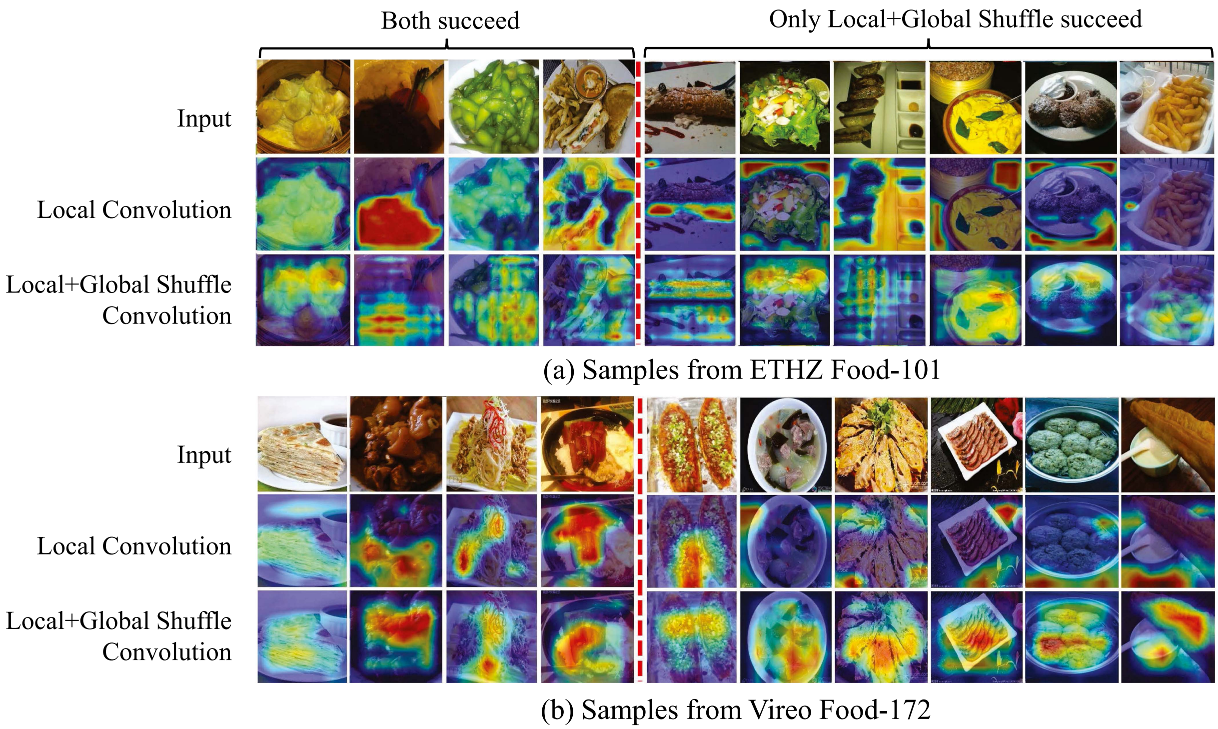

Fig. 6. Visualization of experimental results comparison. (a) Examples from dataset ETHZ Food-101. (b) Examples from dataset Vireo Food-172; Left 4 columns are cases where both local convolution and local+global shuffle convolution can correctly identified; Right 6 columns are cases where local convolution fails but local+global shuffle convolution succeed. The first row is the original image, the second row is the heat maps generated by using only local convolution, and the third row is the heat maps generated by using local+global shuffle convolution. (a) Samples from ETHZ Food-101. (b) Samples from Vireo Food-172.

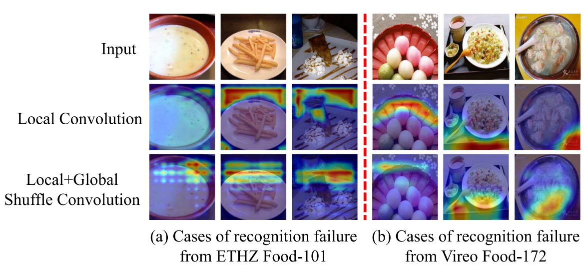

Fig. 7. Visualization of recognition failure cases. (a) Examples from dataset ETHZ Food-101. (b) Examples from dataset Vireo Food-172. The first row is the original image, the second row is the heat maps generated by using only local convolution, and the third row is the heat maps generated by using local+global shuffle convolution. (a) Cases of recognition failure from ETHZ Food-101. (b) Cases of recognition failure from Vireo Food-172.

TABLE VII
ABLATION STUDY

Dataset	Method	Top-1 Acc.	#Params	#FLOPs
Food-101	GS-1.0	**87.0%**	0.9M	295.0M
	w/o GSC	85.5%	0.6M	158.0M
	w/o LC	73.7%	0.6M	158.0M
	w/o HS	86.8%	0.9M	295.0M
Food-172	GS-1.0	**87.8%**	0.9M	295.0M
	w/o GSC	86.5%	0.7M	158.0M
	w/o LC	76.8%	0.7M	158.0M
	w/o HS	87.4%	0.9M	295.0M
Food256	GS-1.0	**69.6%**	1.1M	295.1M
	w/o GSC	69.1%	0.6M	158.0M
	w/o LC	55.6%	0.8M	158.2M
	w/o HS	68.9%	1.1M	295.1M
Food-500	GS-1.0	**62.0%**	1.4M	295.4M
	w/o GSC	59.8%	1.1M	158.5M
	w/o LC	49.3%	1.1M	158.5M
	w/o HS	61.4%	1.4M	295.4M

GS-1.0: GSNet-1.0; GSC: Global Shuffle Convolution; LC: Local Convolution; HS: HardSwish activation function.
The bolded terms in each column represent: highest accuracy, minimal parameter count, and minimal computational workload within each group, respectively.

curve and wider gradient values ranging from $(-1/2, 3/2)$. As given in Table VII, the models using HardSwish achieve higher top-1 accuracy: 69.6% (Food-256), 87.0% (Food-101), and 87.8% (Food-172), compared with models using ReLU: 68.9% (Food 256), 86.8% (Food-101), and 87.4% (Food-172). The results show that the HardSwish activation function helps to find better solutions.

V. Conclusion and Future Work

Focusing on the specific attributes of food images, we introduce a lightweight and efficient CNN network model tailored for food image recognition. Our model leverages a block structure comprising global shuffle convolution and local convolution in parallel. The integration of global shuffle convolution adeptly addresses the dispersed distribution of ingredients in food images, leading to a notable enhancement in recognition accuracy. To complement this, we strategically reduce the number of layers in the rear portion of the network, capitalizing on the front-end's emphasis on capturing long-range information. This approach effectively mitigates the parameter count and FLOPs. Evaluation across four prominent food image databases demonstrates that our method outperforms existing CNN-based, ViT-based, and hybrid lightweight network models. The development of this lightweight network holds promise for enhancing server-side training efficiency and facilitating the deployment of food recognition applications on mobile platforms. This forms a robust foundation for individuals to make informed, environmentally conscious, and health-driven dietary choices in their daily lives.

Moving forward, our future endeavors will encompass adapting to diverse hardware architectures and operating system environments for end devices. In addition, we aim to deploy lightweight algorithms for food recognition, detection, and segmentation, ultimately offering personalized recommendations for environmentally sustainable and health-conscious dietary choices.

References

[1] S. H. Wittwer, *Food, Climate, and Carbon Dioxide: The Global Environment and World Food Production*. Boca Raton, FL, USA: CRC Press, 1995.

[2] S. J. Vermeulen, B. M. Campbell, and J. S. I. Ingram, "Climate change and food systems," *Annu. Rev. Environ. Resour.*, vol. 37, pp. 195–222, 2012.

[3] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, 2019.

[4] A. Ishino, Y. Yamakata, H. Karasawa, and K. Aizawa, "RecipeLog: Recipe authoring app for accurate food recording," in *Proc. ACM Multimedia Conf.*, 2021, pp. 2798–2800, doi: 10.1145/3474085.3478563.

[5] A. Rostami, N. Nagesh, A. Rahmani, and R. C. Jain, "World food atlas for food navigation," in *Proc. 7th Int. Workshop Multimedia Assist. Dietary Manage. Multimedia Assist. Dietary Manage.*, 2022, pp. 39–47, doi: 10.1145/3552484.3555748.

[6] A. Rostami, V. Pandey, N. Nag, V. Wang, and R. C. Jain, "Personal food model," in *Proc. 28th Int. Conf. Multimedia, Virtual Event*, 2020, pp. 4416–4424, doi: 10.1145/3394171.3414691.

[7] K. Nakamoto, S. Amano, H. Karasawa, Y. Yamakata, and K. Aizawa, "Prediction of mental state from food images," in *Proc. 1st Int. Workshop Multimedia Cooking, Eating, Related Appl.*, 2022, pp. 21–28, doi: 10.1145/3552485.3554937.

[8] Y. Yamakata, A. Ishino, A. Sunto, S. Amano, and K. Aizawa, "Recipeoriented food logging for nutritional management," in *Proc. 30th Int. Conf. Multimedia*, 2022, pp. 6898–6904.

[9] T. Yao et al., "Online latent semantic hashing for cross-media retrieval," *Pattern Recognit.*, vol. 89, pp. 1–11, 2019.

[10] J. Ródenas, B. Nagarajan, M. Bolaños, and P. Radeva, "Learning multisubset of classes for fine-grained food recognition," in *Proc. 7th Int. Workshop Multimedia Assist. Dietary Manage. Multimedia Assist. Dietary Manage.*, 2022, pp. 17–26, doi: 10.1145/3552484.3555754.

[11] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Trans. Image Process*, vol. 29, pp. 265–276, 2020.

[12] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *Proc. Winter Conf. Appl. Comput. Vis.*, Lake Tahoe, NV, USA, 2018, pp. 567–576, doi: 10.1109/WACV.2018.00068.

[13] J. Zhao et al., "Deep-learning-based automatic evaluation of rice seed germination rate," *J. Sci. Food Agriculture*, vol. 103, no. 4, pp. 1912–1924, 2023.

[14] Z. Huang et al., "Fast location and segmentation of high-throughput damaged soybean seeds with invertible neural networks," *J. Sci. Food Agriculture*, vol. 102, no. 11, pp. 4854–4865, 2022.

[15] W. Min et al., "Vision-based fruit recognition via multi-scale attention CNN," *Comput. Electron. Agriculture*, vol. 210, 2023, Art. no. 107911.

[16] W. Shafik et al., "Using a novel convolutional neural network for plant pests detection and disease classification," *J. Sci. Food Agriculture*, vol. 103, no. 12, pp. 5849–5861, 2023.

[17] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.

[18] G. Sheng, S. Sun, C. Liu, and Y. Yang, "Food recognition via an efficient neural network with transformer grouping," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 11465–11481, 2022.

[19] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[20] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[21] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," 2022, *arXiv:2206.02680*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[23] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 122–138.

[24] S. Mehta, M. Rastegari, L. G. Shapiro, and H. Hajishirzi, "ESPNetV2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9190–9200.

[25] M. Tan and V. Quoc Le, " EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, pp. 6105–6114.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[27] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[28] Y. Li et al., "Efficientformer: Vision transformers at mobilenet speed," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 12934–12949, 2022.

[29] T. Huang, L. Huang, S. You, F. Wang, C. Qian, and C. Xu, "LightViT: Towards light-weight convolution-free vision transformers," 2022, *arXiv:2207.05557*.

[30] H. Cai et al., "Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17302–17313.

[31] J. Zhang et al., "MiniViT: Compressing vision transformers with weight multiplexing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12135–12144.

[32] K. Wu et al., "TinyViT: Fast pretraining distillation for small vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 68–85.

[33] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5260–5269.

[34] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12165–12175.

[35] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[36] A. Srinivas, T. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529.

[37] J. Li et al., "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios," 2022, *arXiv:2207.05501*.

[38] J. Pan et al., "EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 294–311.

[39] S. Mehta and M. Rastegari, "MobileViT: Lightweight, general purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2022.

[40] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2249–2256.

[41] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101-mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[42] W. Min, L. Liu, Z. Luo, and S. Jiang, "Ingredient guided cascaded multi-attention network for food recognition," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1331–1339.

[43] W. Min et al., "Large scale visual food recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9932–9949, Aug. 2023.

[44] S. Horiguchi, S. Amano, M. Ogawa, and K. Aizawa, "Personalized classifier for food image recognition," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2836–2848, Oct. 2018.

[45] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 1085–1088.

[46] Y. Kawano and K. Yanai, "Real-time mobile food recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2013, pp. 1–7.

[47] S. Y. Kawano and K. Yanai, "FoodCam: A real-time food recognition system on a smartphone," *Multimedia Tools Appl.*, vol. 74, no. 14, pp. 5263–5287, 2015.

[48] P. Pouladzadeh and S. Shirmohammadi, "Mobile multi-food recognition using deep learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 3s, pp. 1–21, 2017.

[49] R. Z. Tan, X. Chew, and K. W. Khaw, "Neural architecture search for lightweight neural network in food recognition," *Mathematics*, vol. 9, no. 11, pp. 1245–2021, 2021.

[50] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated Residual Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 636–644.

[51] M. Klasson, C. Zhang, and H. Kjellström, "A hierarchical grocery store image dataset with visual and semantic labels," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2019, pp. 491–500.

[52] Y. Kawano and K. Yanai, "FoodCam-256: A. large-scale realtime mobile food recognition system employing high-dimensional features and compression of classifier weights," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 761–762.

[53] W. Min et al., "ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 393–401.

[54] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev*, vol. 60, no. 2, pp. 223–311, 2018.

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[56] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[57] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1577–1586.

[58] J. F. Yeh, K.-M. Lin, C.-Y. Lin, and J.-C. Kang, "Intelligent mango fruit grade classification using AlexNet-SPP with mask R-CNN-Based segmentation algorithm," *IEEE Trans. AgriFood Electron.*, vol. 1, no. 1, pp. 41–49, Jun. 2023.

Guorui Sheng received the M.E. degree in computer science from Kunsan National University, Gunsan, South Korea, in 2007, and the Ph.D. degree in computer application technology from Nankai University, Tianjin, China, in 2017.

From 2017 to 2018, he was a Research Assistant to Scholar Bruce Denby with the School of Computer Science and Technology, Tianjin University. He is currently a Lecturer with the Department of Information and Electrical Engineering, Ludong University, Yantai, China. He has authored or co-authored more than 20 peer-referenced papers in relevant journals and conferences, including *ACM Transactions on Multimedia Computing, Communications, and Applications* and *Nutrients*. His research interests include computer vision, deep learning, and food computing.

Weiqing Min (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

He is currently an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. He has authored or co-authored more than 50 peer-referenced papers in relevant journals and conferences, including *Patterns* (Cell Press), *ACM Computing Surveys*, *Trends in Food Science and Technology*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, *Food Chemistry*, ACM MM, AAAI, and IJCAI. His research interests include multimedia content analysis and food computing.

Mr. Win was a Senior Member of CCF. He was the recipient of the 2016 *ACM Transactions on Multimedia Computing, Communications, and Applications*, the Nicolas D. Georganas Best Paper Award, and the 2017 *IEEE Multimedia Magazine* Best Paper Award. He was the Guest Editor for the special issues on international journals, such as IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and *Foods*.

Tao Yao received the Ph.D. degree in multimedia retrieval from the Dalian University of Technology, Dalian, China, in 2017.

He is currently an Associate Professor with the Department of Information and Electrical Engineering, Ludong University and also a Researcher with Yantai Research Institute of New Generation Information Technology, Southwest Jiaotong University, Chengdu, China. He has authored or co-authored more than 30 peer-referenced papers in relevant journals and conferences, including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS CYBERNETICS, *ACM Transactions on Multimedia Computing, Communications, and Applications* and *Pattern Recognition*. His research interests include multimedia retrieval, computer vision, and machine learning.

Jingru Song received the B.E. degree in software engineering from the College of Computer Science, Liaocheng University, Liaocheng, China, in 2022. She is currently working toward the M.E. degree in computer science and technology with the College of Information and Electrical Engineering, Ludong University, Yantai, China.

Her research interests include multimedia processing, computer vision, and food computing.

Yancun Yang received the Ph.D. degree in management from Shandong University, Jinan, China, in 2008.

He is currently a Lecturer with the Department of Information and Electrical Engineering, Ludong University, Yantai, China. He has authored or co-authored more than 10 peer-referenced papers in relevant journals and conferences, including *ACM Transactions on Multimedia Computing, Communications, and Applications* and *Nutrients*. His research interests include computer vision, deep learning, and food computing.

Lili Wang received the M.E. and Ph.D. degrees in electromagnetic field and microwave technology from Electronic Engineering School, Beijing University of Posts and Telecommunication, Beijing, China, in 2006.

She is currently a Professor with the School of Information and Electrical Engineering, Ludong University, Yantai, China. Her research interests include broadband communication and multimedia communication.

Shuqiang Jiang (Senior Member, IEEE) received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, and a Professor with the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. He has authored or co-authored more than 150 articles. He was supported by the National Science Fund for Distinguished Young Scholars in 2021, the NSFC Excellent Young Scientists Fund in 2013, and the Young Top-Notch Talent of Ten Thousand Talent Program in 2014. His research interests include multimedia analysis and multimodal intelligence.

Mr. Jiang is a Senior Member of CCF and a Member of ACM. He was a TPC Member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM. He was the recipient of the Lu Jiaxi Young Talent Award from CAS in 2012 and the CCF Award of Science and Technology in 2012. He is the Vice Chair of the IEEE CASS Beijing Chapter and the ACM SIGMM China Chapter. He was the General Chair of ICIMCS in 2015 and the Program Chair of the 2019 ACM Multimedia Asia and PCM in 2017. He is an Associate Editor of *Multimedia Tools and Applications* and *ACM Transactions on Multimedia Computing, Communications, and Applications*.