Received: 30 June 2022 | Accepted: 11 August 2022

DOI: 10.1002/int.23050

RESEARCH ARTICLE

WILEY

Food recognition via an efficient neural network with transformer grouping

Guorui Sheng | **Shuqi Sun** | **Chengxu Liu** | **Yancun Yang**

Department of Artificial Intelligence, School of Information and Electrical Engineering, Ludong University, Yantai, China

Correspondence
Yancun Yang, School of Information and Electrical Engineering, Ludong University, 186, Middle Hongqi Road, Zhifu District, Yantai City 264025, Shandong Province, China.
Email: Sdeven95@126.com

Abstract

Recently, considerable research efforts have been devoted to food recognition for its great potential applications in human health. Much work so far has focused on directly extracted deep visual features via Convolutional Neural Networks, which require significant computational resources and training time. The high requirements on hardware resources severely limit the application of food recognition in mobile devices and the sustainable extension on the server side. Therefore, how to design an efficient and high-performance lightweight neural network for food recognition is the key to solve the problem. In this paper, we propose a Lightweight Transformer-Based Deep Neural Network for food image recognition, which can achieve effective recognition of food images with fewer parameters and lower computational cost. Through Transformer Grouping and Token Shuffling, we construct an efficient food image recognition network that effectively combines the advantages of Transformer to extract global features and MobileNet to extract local features. The proposed network architecture effectively copes with the particularly scattered distribution of salient features in food images, and improves the recognition rate. We conduct extensive experiments on three popular food data sets, demonstrating that our method achieves state-of-the-art

performance in applying lightweight neural networks to food image recognition.

KEYWORDS
deep learning, food recognition, lightweight, mobilenet, transformer

1 | INTRODUCTION

In the field of computer vision and multimedia, food computing[1] has received more and more attention in recent years. The importance of analyzing and understanding food images from different perspectives is obvious, such as nutrition estimation,[2] food choices,[3,4] food diaries,[5] healthy eating recommendations,[6–8] and the cafeteria.[9] Food recognition is an important fundamental step in gaining a deeper understanding of food.

As a recognition task, the key to food recognition is to extract discriminative visual features. Early research on food recognition was mainly about handcrafted features.[10–12] Recently, food image recognition is moving towards the use of deep learning as a general solution due to its powerful discriminative feature learning capability. For example, Meyers et al.[5] used the GoogLeNet network to train a multilabel classifier to predict the type of food present in a meal. Martinel et al.[13] proposed a wide slice residual network to capture the vertical structure of food images. Deep learning-based methods usually achieve better performance than handcrafted features because of their advantages in representation learning. While these deep learning models perform well on the food recognition task they are trained on, they may not be effective enough for direct deployment in the real world. Practitioners of deep learning-based food recognition may face some challenges when training or deploying models, which stem from three aspects:

- *Widely distributed subtle discriminative details in food images*: There are widely distributed subtle discriminative details in food images, which are more difficult to capture in many cases, resulting in food image recognition being a fine-grained recognition. For example, as shown in Figure 1, the food fried rice, which is very common in China, has a variety of food ingredients, seasonings, and cooking methods due to different regions. So the final image of fried rice also shows weak distinguishing features and a very scattered layout. Therefore, in many cases, the discriminative details are too subtle and scattered to be well represented by existing Convolutional Neural Networks (CNNs). Effective global feature extraction is particularly important for food images. The recently popular attention mechanism is expected to play a role in effectively recognizing food images.
- *Enabling on-device deployment*: Most food recognition applications, such as visual food diary,[5] health-aware recommendation,[2,14] and cafeterias,[9] need to run in real-time on Internet of Things (IoT) and smart devices, where model inference happens directly on the device. One of the reasons is that food consumption is closely related to mass life and happens every day. Therefore, most of these consumption activities take place on convenient portable devices so it becomes imperative to optimize the model for the target device. To the best of our knowledge, there is no research work in this area so far.
- *Sustainability extension on the server side*: Training and deploying large deep learning models are very resource-intensive. Although training may be a one-time consumption, deployment and long-

FIGURE 1 Some samples from food data set, with widely distributed subtle discriminative details [Color figure can be viewed at wileyonlinelibrary.com]

running inference can still become expensive in terms of consumption of server-side RAM, CPU, and so forth. With the global carbon reduction trends, the carbon footprint of a data center is very concerning. For example, famous organizations, like, Google, Facebook, Amazon, and so forth spend several billion dollars each per year in capital expenditure on their data centers. To achieve green and sustainable artificial intelligence (AI), we should propose building efficient neural architectures that can achieve the same level of accuracy while reducing carbon footprint by orders of magnitude. As for food culture in the world, there are so many types and shapes of food which directly leads to a large number of food images. Effective training of such a large and growing collection of food images places very high demands on hardware resources. Under the trend of carbon emission reduction, we should design a more efficient neural network for food image recognition.

Taking these factors into consideration, we propose a Lightweight Transformer-Based Deep Neural Network (LTBDNN) for food recognition, which is capable of achieving efficient classification of food images with fewer parameters. This framework mainly consists of two components, namely, MobileNet Part (MP) and Transformer Part (TP). MP takes the image as input and stacks MobileNet blocks, it extracts local features at pixel level by leveraging the efficient depthwise and pointwise convolution. TP takes six learnable tokens as input and stacks multihead attention (MHA) and feed-forward networks (FFNs). Global features of the food image are got from these tokens.

In view of the importance of global features for recognizing food images, we have carried out several targeted designs for the TP: (1) The way of Token Generation. We use the original image to be convolved and then unfolded as the input tokens of the TP. This strategy can enable

LTBDNN to learn global representations with spatial inductive bias. (2) The generated tokens are grouped and used as the input of Transformer Grouping. This strategy can effectively extract the global features of food images. (3) The features learned by each transformer group are shuffled before entering the next block so that the global features are further fused and the recognition rate is more effectively improved.

In each MobileNet–Transformer block, the global features extracted by the TP are fused with the local features extracted by the MP and finally used for food image recognition. During the fusion process, we use lightweight cross attention to model this bidirectional bridge by performing the cross attention at the bottleneck of MP where the number of channels is low, and removing projections (W^Q, W^K, W^V) from MP where the number of positions is large, but keeping them at TP. This strategy ensures that the model has a smaller number of parameters and lower computational consumption.

The final experimental results show that our method can extract more comprehensive and extensive global features for food images, and further fuse with the local features extracted by MobileNet, resulting in a very good recognition effect.

To evaluate our method, we conduct extensive experiments on three popular food data sets: ETH Food-101,[11] Vireo-Food 172,[15] and ISIA Food-500.[16] Our method achieves solid performance on these data sets.

The contributions of our paper can be summarized as follows:

- We introduce a new LTBDNN for food recognition, which can achieve an efficient classification of food images with fewer parameters and computational costs. To the best of our knowledge, we are the first research work to apply lightweight neural networks to food image recognition.
- We propose a wide-range attention mechanism based on Transformer Grouping and Token Shuffling, which can effectively deal with the very scattered key features of food images.
- We conduct an extensive evaluation of our proposed method to verify the effectiveness of our approach. As one strong baseline, code and models will also be released upon publication to support future research.

2 | RELATED WORK

Food recognition: Image recognition[17–21] is one of the most active directions in the field of AI recently, and the recognition of food images belongs to the more difficult fine-grained recognition. Traditional food image recognition consists of two steps: (1) food image feature extraction and (2) classification model training. Among them, image feature extraction and selection is the key to food image recognition. Features here mainly refer to handcrafted features, ranging from simple features, such as color, texture, shape, edge, and spatial relationship to Scale Invariant Feature Transform (SIFT),[22] histograms of oriented gradient,[23] and so forth. For example, Yang et al.[10] first use a semantic texton forest to calculate the component distribution of each pixel in an image, and then builds multidimensional histogram features as visual representations. Bettadapura et al.[12] combined different types of feature descriptors such as the original SIFT[22] and its variants into fused features for food recognition. Some works[24,25] use multikernel learning to fuse various types of image features such as SIFT, Gabor texture, and color histogram for food image recognition.

CNN can learn image features layer by layer: Its bottom layer is general features, such as image edges, textures, and so forth; high-level features are a combination of low-level features, which are specific features for specific tasks.[26–30] Due to the powerful expressive ability of CNN,

it was soon also applied to the field of food image recognition.[31,32] Some work extract features directly on pretrained networks, for example, Ming et al.[33] used the ResNet network to directly extract visual features for food image recognition. Some works fine-tune the existing deep networks on food image data sets, for example, the earliest work that applied deep learning network to dish image recognition[34] extracted image features by fine-tuning the AlexNet network; some works redesign deep neural networks for food image recognition tasks, Martinel et al.[13] proposed a WISeR network for food images with specific vertical structures (such as hamburger, pizza, cake, etc.), recognition is performed by fusing visual features from the Wide Residual Networks[35] and the proposed Slice Network in the paper.

Different from the existing methods that mainly focus on the improvement of the recognition rate and ignore the size and efficiency of the model, our work aims to achieve a more efficient and lightweight implementation on food recognition.

Lightweight deep neural network: Deep neural networks have achieved the highest precision in various tasks at the expense of many parameters, requiring significant computational resources and training time. Thus there is a huge demand for model compression and acceleration techniques before deploying to resource-constrained devices and real-time applications. In recent years, a growing number of methods have been presented for compressing and accelerating the network while making the slightest compromise with the model accuracy. Most approaches can be classified into the following categories: parameter pruning, network quantization, low-rank factorization, model distillation, and compact network design.

SqueezeNet[36] extensively uses 1×1 convolutions with squeeze and expand modules primarily focusing on reducing the number of parameters. More recent works shifts the focus from reducing parameters to reducing the number of operations (MAdds) and the actual measured latency. MobileNet V1[37] employs depthwise separable convolution to substantially improve computation efficiency. MobileNet V2[38] expands on this by introducing a resource-efficient block with inverted residuals and linear bottlenecks. ShuffleNet[39] utilizes group convolution and channel shuffle operations to further reduce the MAdds.

Recently, Vision Transformer (ViT)[40] demonstrates the advantage of global processing and achieves significant performance boost over CNNs. ViT divides an image into a sequence of nonoverlapping patches and then learns interpatch representations using multiheaded self-attention in transformers.[41] The general trend is to increase the number of parameters in ViT networks to improve performance.[42–44] However, these performance improvements come at the cost of model size (network parameters) and latency. Many real-world applications require visual recognition tasks (e.g., food recognition) to run on resource-constrained mobile devices in a timely fashion. To be effective, ViT models for such tasks should be lightweight and fast.

Chen et al.[45] propose a new network that parallelizes MobileNet and Transformer with a two-way bridge in between, named Mobile-Former. The bridge and Former consumes less than 20% of the total computational cost, but significantly improve the representation capability. Mehta et al.[46] introduce MobileViT, a lightweight and general-purpose ViT for mobile devices. MobileViT presents a different perspective for the global processing of information with transformers, that is, transformers as convolutions. The author claimed that MobileViT significantly outperforms CNN- and ViT-based networks across different tasks and data sets.

Our work is much inspired by the recent work,[45] in that we are both devoted to create a ViT-based lightweight neural network. However, we have significant differences in three-fold. (1) *Motivation*: We aim to create an efficient neural network for ViT-based food image recognition, while Chen et al.[45] attempt to find a general-purpose parallel design of MobileNet and Transformer. (2) *Methodology*: Chen et al.[45] introduce a Mobile-Former structure which contains very few tokens

that are randomly initialized, we take into consideration learning global representations with spatial inductive bias. To this end, we unfold input image into nonoverlapping flattened pathces result in not losing the patch order nor the spatial order of pixels within each patch. In addition, for the weak and scattered features of food images, we use Transformer Grouping method to make Transformer obtain broader and more global features. Further, we use the Token Shuffling strategy to fully communicate the global features learned by two Transformer group to further adapt to the characteristics of food images and achieve higher classification accuracy. We use three classic data sets of food images for classification comparison, confirming that our method achieves state-of-the-art results in classical lightweight neural networks including.[45]

3 | METHOD

In this section, we will introduce the proposed LTBDNN for food recognition. Figure 2 illustrates the architecture of LTBDNN.

The same as Mobile-Former,[45] LTBDNN has two branches, one of which is based on Transformer for extracting global features of food images, and the other is based on MobileNet for extracting local features. Global features and local features are fused in each Transformer-MobileNet block. Different with Mobile-Former,[45] LTBDNN performs Token Generation, Transformer Grouping, and Token Shuffling, to make the method more effective in food image recognition: In the process of global feature extraction, the convolution and unfold operations are sequentially applied to the original image to generate six tokens as input of TP. This Token Generation method enables LTBDNN to learn global representations with spatial inductive bias and helps it obtain global features better. In the TP, Transformer Grouping is used for multichannel simultaneous feature extraction, and the Token Shuffling is used to make the extracted multichannel features fully communicated. These two operations further improve the effect and efficiency of TP to extract global features. The following will introduce the details of the three steps Token Generation, Transformer Grouping, and Token Shuffling that play a decisive role in improving the performance of food image recognition, as well as the details of the information exchange between MP and TP.

3.1 | Token Generation

Different from the randomly initialized tokens in Mobile-Former, our method obtains the input tokens of TP by first applying the convolution operation to the original image, and then

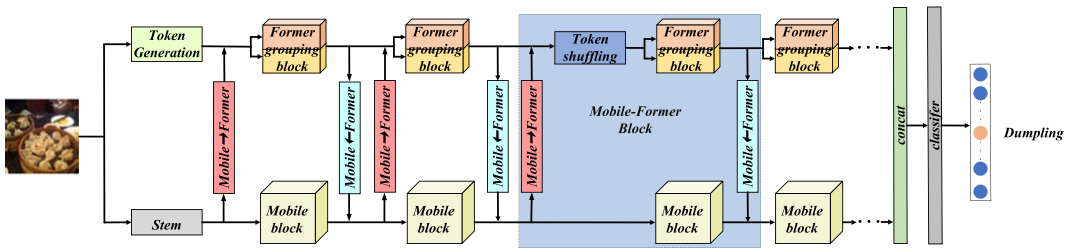

FIGURE 2 Overview of proposed Lightweight Transformer-based Deep Neural Network for food recognition [Color figure can be viewed at wileyonlinelibrary.com]

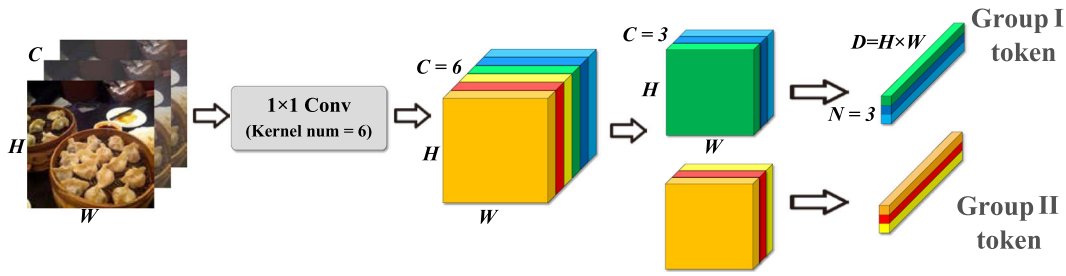

FIGURE 3 Proposed Token Generation process [Color figure can be viewed at wileyonlinelibrary.com]

applying the unfold operation to the obtained feature map. This Token Generation method enables LTBDNN to learn global representations with spatial inductive bias, which means neither the patch order nor the spatial order of pixels within each patch is lost in the generated tokens.

As shown in Figure 3, after performing a convolution operation on the input image ($X_0 \in \mathbb{R}^{H\times W\times 3}$), its corresponding feature map is generated, denoted as $X_{\text{token}} \in R^{H\times W\times c}$, where c, W, and H correspond to the number of channels, width, and height of the feature map, respectively. Then unfold the two dimensions of W and H of the generated feature map X_{token} to generate the tokens required for the transformer operation, denoted as $Z \in R^{M\times d}$, where d and M are the dimensions and number of tokens, respectively. To obtain comprehensive global information for the TP, we specify the correspondence between the feature image and the token, one channel of the feature map corresponds to one token, namely,

$$\begin{aligned} &unfold\left(X_{\text{token}}\right): \left[X_{\text{token}}^{H\times W\times c} \rightarrow X_{\text{token}}^{(H\times W)\times c} \rightarrow X_{\text{token}}^{d\times c}\right] \\ &Z = unfold\left(X_{\text{token}}\right), \quad Z \in R^{M\times d}. \end{aligned} \tag{1}$$

It should be noted that to match the correspondence mentioned above, it should be satisfied: $M = c$, here we set $M = c = 6$.

3.2 | Transformer Grouping

After generating the tokens, we divide them into two groups and pass them into two Transformers for training.

As shown in Figure 4, the MobileNet block and Transformer block in the network are connected through a two-way bridge whose direction is determined by the flow direction of the feature information. We define the bridge that flows from the MobileNet block to the Transformer block as MobileNet → Transformer, which flows from the Transformer block to the direction of the two-way bridge. The bridge of the MobileNet block is Transformer → MobileNet. We use a lightweight cross attention to model this bidirectional bridge by performing the cross attention at the bottleneck of MobileNet where the number of channels is low, and removing projections (W^Q, W^K, W^V) from MobileNet side where the number of positions is large, but keeping them at Transformer side. At the MobileNet → Transformer stage, we divide the feature map output from the MobileNet block to the Transformer block equally according to the channel dimension and perform cross-attention on each group of

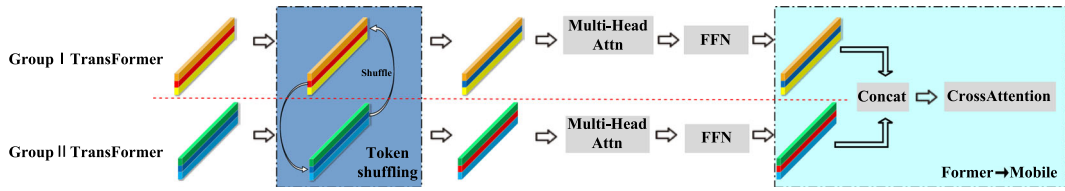

FIGURE 4 Proposed Transformer Grouping structure, where we also show Token Shuffling implementation. FFN, feed-forward network. [Color figure can be viewed at wileyonlinelibrary.com]

tokens in the two Transformer blocks, respectively. We define the output of the MobileNet block as X_i, where i represents the ith layer, x_i^c represents the ith channel of the feature map output by the cth layer, and Z_G^i represents the ith layer token of the Gth group Transformer:

$$Z_1^{\text{hidden}} = Z_1^i + \left[Attention\left(Z_1^h(W_1)_h^Q, x_i^c, x_i^c\right)\right]_{h=1:H} W^o, \quad c \in \left[0, \frac{C}{2}\right), \tag{2}$$

$$Z_2^{\text{hidden}} = Z_2^i + \left[Attention\left(Z_2^h(W_2)_h^Q, x_i^c, x_i^c\right)\right]_{h=1:H} W^o, \quad c \in \left[0, \frac{C}{2}\right). \tag{3}$$

Among them, $(W_1)_h^Q$ represents the query matrix of the hth head part in the first group of Transformers, W^o is used to combine the MHA results, C represents the total number of channels of the feature map, and $Attention(Q, K, V)$ is the standard attention equation, query Q, key K, and value V follow:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{4}$$

Here the K and V come from the feature map X_i obtained from the MobileNet block, and Q comes from the token in the Transformer block.

Similar to the previous part, in the Transformer → MobileNet module, the information of the Transformer block flows to the MobileNet block through a two-way bridge. Here we do not calculate the *Attention* twice but combine the tokens obtained by the two groups of Transformers in dimension. After that, the transmission of information is completed after an attention operation:

$$X_i = X_i + \left[Attention\left(X_h, Z_h W_h^K, Z_h W_h^V\right)\right]_{h=1:H}, \tag{5}$$

where W_h^K and W_h^V represent the projection matrix of *Key* and *Value*. The feature map obtained by the MobileNet block X_i is as the *Query*, the token in the Transformer block is used as *Key* and *Value*, respectively.

During the bidirectional bridge operation, the result of MobileNet → Transformer is used as the input of the Transformer block, and the input of the MobileNet is generated by Transformer → MobileNet block. To reduce the calculation amount of the model, three parameters W^Q, W^K, and W^V are only used in the TP, that is, the above three parameters only generate dot multiplication with the tokens but not with the feature map matrix.

The reason why the Transformer Grouping strategy is adopted is based on the characteristics of food images: common object recognition generally has a single global feature, while food images often have multiregional quasi-global features with little difference in recognition. The parallel structure of two transformers is conducive to extracting food image features from different perspectives at the same time, while the number of two transformers is a compromise between accuracy and model scale. Through Transformer Grouping, the model can learn the global features of food images more comprehensive, and make targeted improvements to the scattered and large range of food image features. The subsequent experimental results prove this.

3.3 | Token Shuffling

The correspondence between feature images and tokens is mentioned in Section 3.1, each token stores the global feature in each channel. Considering that the key features of food images are very scattered, let the information between the two groups of Transformers fully communicate by making the tokens shuffled in the Transformer, so that the two sets of Transformers take into account the different dimensional features of the input image during the optimization process. Use Z_G^n to represent each token, where G represents the group to which the token belongs, and n represents the token number:

$$\begin{aligned} &Z_1 = \left(Z_1^1 Z_1^2 Z_1^3\right), \quad Z_2 = \left(Z_2^1 Z_2^2 Z_2^3\right), \\ &replace\left(Z_1, Z_2\right) \rightarrow Z_1 = \left(Z_1^1 Z_2^2 Z_1^3\right), \quad Z_2 = \left(Z_2^1 Z_1^2 Z_2^3\right). \end{aligned} \tag{6}$$

It should be noted that the Token Shuffling operation occurs after MobileNet → Transformer, and before Z_i participates in the MHA calculation of the Transformer block, as shown in Figure 4.

The Token Shuffling here comes from the exchange of tokens from the two sets of transformers, which can achieve the purpose of fully communicating the global features obtained by the two transformers. The reason for only shuffle one token from each group is to effectively control the amount of computation on the TP side.

3.4 | Information flow between MP and TP

In the MobileNet block, the input is the output feature map $X_i \in R^{L \times C}$ by the previous layer of network, where C is the number of channels and L represents its spatial feature ($L = hw$, where h and w are height and width of the feature map). In the Transformer block, the grouped global tokens are used as the input for each Transformer: $Z_1^i \in R^{(M/2) \times d}$, $Z_2^i \in R^{(M/2) \times d}$, where M represents the total number of tokens in the Transformer block, and d represents the dimension of the tokens. The output of the MobileNet block is the input required by the next

layer of network, and its L and C are determined by the convolution and pooling operation of the current layer. The two sets of outputs generated by the two Transformers are spliced by dimension and used as the output of the entire Transformer block, after completing the Attention operation, the output needs to be further grouped and shuffled as the input of the next layer of Transformer blocks. In particular, for the first layer of Transformer block, its input is formed by unfolding the spatial dimension L of the feature map obtained after the original image is convolved by one layer (the number of convolution kernels is equal to M).

4 | EXPERIMENT

4.1 | Data set

Three data sets are selected to evaluate our method, namely, ETH Food-101,[11] VireoFood-172,[15] and ISIA Food-500.[16] The number of categories and images of the data sets increased in turn to verify the robustness of our model.

ETH Food-101 contains 101,000 images that belong to 101 food categories. There are 750 training images and 250 testing images for each category.

VireoFood-172 contains 110,241 food images from 172 categories. Similar to ETH food-101, in each food category, 60%, 10%, and 30% of images are randomly selected for training, validation, and testing, respectively.

ISIA Food-500 contains 500 categories from the list in Wikipedia, altogether 399,726 images. It is a large-scale ontology of food images and a more comprehensive food data set that surpasses existing popular benchmark data sets by category coverage and data volume.

4.2 | Experimental setup

Our model is implemented on the Pytorch platform. The images are resized to 224×224. The model is optimized using the stochastic gradient descent with a batch size of 100 and momentum of 0.9, weight decay of 10^{-5}. The learning rate is initially set to 10^{-3} and divided by 100 every 200 epochs. Top-1 accuracy and Top-5 accuracy are used as evaluation metrics.

4.3 | Experiment on food image data set

We evaluated LTBDNN against existing benchmark lightweight deep neural networks on three food image data sets mentioned above. We conduct experiments using LTBDNN generated from tokens of two different dimensions. Among them, 150 dimensional-token (denoted as TD-150) represents a smaller token, which makes the network parameters less, and 192 dimensional-token (denoted as TD-192) represents a higher dimension, which makes the network parameters slightly higher but can get better results.

Table 1 shows the performance comparison of LTBDNN and baseline lightweight network on three data sets. It can be seen from the experimental results on ETH Food-101 in Table 1 that (1) Our method exceeds all baseline methods. Compared with the recently published

TABLE 1 Performance comparison on three popular food image data sets

Model	#Params. (*M*)	ETH Food-101		VireoFood-172		ISIA-Food 500	
		Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
MobileNet V3(large)[47]	4.20	71.63	90.24	75.33	91.03	41.10	70.00
ShuffleNet V1[39]	1.16	63.56	84.92	67.81	87.27	37.38	65.28
ShuffleNet V2(×2.0)[48]	5.87	71.50	89.83	72.78	89.06	19.18	42.94
GhostNet[49]	4.03	69.57	88.27	72.08	89.01	45.51	73.86
Mobile-Former-294[45]	10.10	71.42	89.41	75.82	92.03	48.15	75.92
LTBDNN(TD-150)	**10.08**	**73.81**	**90.26**	**79.18**	**92.42**	**50.28**	**78.23**
LTBDNN(TD-192)	**12.23**	**76.83**	**92.50**	**78.82**	**93.58**	**49.83**	**77.31**

Note: Bold indicates the best experimental results.

Abbreviation: LTBDNN; Lightweight Transformer-Based Deep Neural Network.

lightweight network GhostNet, our method outperforms by 5 and 4 percentage points in Top-1 and Top-5 accuracy, respectively. (2) Our method outperforms the Transformer-based lightweight network Mobile-Former[45] by about 3.5 and 3.2 percentage points in Top-1 and Top-5 accuracy, respectively. Although marginal performance improvement, our method did not use additional data augmentation strategy, like, WISeR, which additionally applied various photometric distortions and AlexNet-style color augmentation.

Table 1 also shows the performance comparison of LTBDNN and baseline lightweight network on VireoFood-172 data set. The total number of images in the data set VireoFood-172 is basically the same as that of ETH Food-101, but the categories are increased by 71, which will bring greater challenges to the model - - how to maintain good performance with fewer trainable images.

We first compare LTBDNN with the baseline lightweight neural network on VireoFood-172. As shown in Table 1, we can see that LTBDNN achieves the state-of-the-art performance in both Top-1 accuracy and Top-5 accuracy. Compared with the best performing baseline lightweight network MobileNet V3, the accuracy of LTBDNN achieves significantly higher Top-1 accuracy (78.83% vs. 75.33%) and Top-5 accuracy (93.58% vs. 91.03%). Table 1 also shows that LTBDNN exceeds the Transformer-based lightweight network Mobile-Former, there is a performance improvement of about 3% and 1.5% in Top-1 and Top-5 accuracy, respectively.

Table 1 shows the performance of LTBDNN on ISIA Food-500. It can be seen that the experimental results of classic lightweight networks on this database are relatively low, which is attributed to the three major characteristics of this database: (1) *Larger data volume*: It has 399,726 images from 500 food categories. (2) *Larger category coverage*: It consists of 500 categories, which is about 3–5 times that of existing data sets, such as Food-101 and Vireo Food-172. (3) *Higher diversity*: Food categories from this data set covers various countries and regions including both eastern and western cuisines. As shown in Table 1, compared with other models, LTBDNN performs well. The reason is that LTBDNN can efficiently combine the global features and local features of food images, so it can make a more effective identification of the characteristics of Western dishes and the characteristics of Chinese food.

4.4 | Comparison among efficient CNNs and Mobile-Former

In terms of accuracy and number of parameters trade-offs, the comparisons are performed on the three food image data sets mentioned above. Table 1 shows the comparison between LTBDNN and classic efficient CNNs: (a) ShuffleNetV1,[39] V2[48] and (b) MobileNet V1,[37] V3[47] and (c) GhostNet[49] and (d) Mobile-Former.[45] The comparison covers a number of parameters that range from 1.16*M* to 12.23*M*. Compared with the latest efficient CNN GhostNet, although LTBDNN has a higher number of parameters (12.23*M* vs. 4.03*M*), LTBDNN achieves significantly higher Top-1 accuracy (76.83% vs. 69.57%). This demonstrates that our design improves the representation capability efficiently for food images.

Because LTBDNN adopts Transformer, it has more parameters than efficient CNN, but because of its stronger global feature expression ability, the recognition effect is much higher than that of efficient CNN. In the case of the rapid development of the hardware level, compared with the huge improvement brought by the performance, we believe that a moderate increase in the amount of parameters is acceptable.

Compared with Mobile-Former, the Top-1 accuracy of LTBDNN on ETH Food-101 is 2.4 percentage points higher with almost the same amount of parameters, and the accuracy is 5.4 percentage points higher with 2*M* more parameters. This significant performance improvement will greatly facilitate the deployment of LTBDNN on IoT devices.

4.5 | Ablation and discussion

In this section, we show LTBDNN is effective via several ablations performed on three data sets classification. Here, Mobile-Former-294M[45] is used and all models are trained for 300 epochs.

LTBDNN is more effective in food image recognition than Mobile-Former as it performs Token Generation, Transformer Grouping, and Token Shuffling while encodes global interaction via Transformer efficiently, resulting in more accurate prediction. As shown in Table 2, the three steps each show their important roles in improving the accuracy of food image recognition. Taking the experimental results on ETH food-101 as an example, adding Token Generation gains 1.5% Top-1 accuracy over the baseline that uses Mobile-Former alone. This validates our Token Generation design in LTBDNN, it can enable LTBDNN to learn global representations with spatial inductive bias. In addition, another ablation on the step of Transformer Grouping, shows that after adopting this step, the Top-1 recognition accuracy was further improved by 0.9 percentage points. Finally, the Token Shuffling step further improves the final Top-1 accuracy by 3.7 percentage points to 76.83%. The final two-step performance improvement validates our Transformer Grouping and Token Shuffling designs in LTBDNN, which can enable LTBDNN to learn the global features of food images more efficiently and accurately, thereby greatly improving recognition accuracy.

4.6 | Qualitative analysis and visualization

In this section, we will visualize why LTBDNN can achieve better performance in food image recognition.

Figure 5 shows the comparison of attention span for images containing regular objects and food dishes. It can be found that the key features used for class distinction are very obvious and concentrated in the images containing regular objects, so the attention mechanism needs to pay

TABLE 2 Ablation of Token Generation, Transformer Grouping, and Token Shuffling evaluated on three popular food image data sets

Model				ETH Food-101		Virio Food-172		ISIA Food-500	
Mobile-Former	Token Generation	Transformer Grouping	Token Shuffling	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
√	–	–	–	71.42	89.41	75.82	92.03	48.15	75.92
√	√	–	–	72.90	90.22	76.64	92.32	48.24	75.41
√	√	√	–	73.18	90.51	76.86	92.90	48.87	76.34
√	√	√	√	**76.83**	**94.62**	**78.82**	**93.58**	**49.83**	**77.31**

Note: Bold indicates the best experimental results.

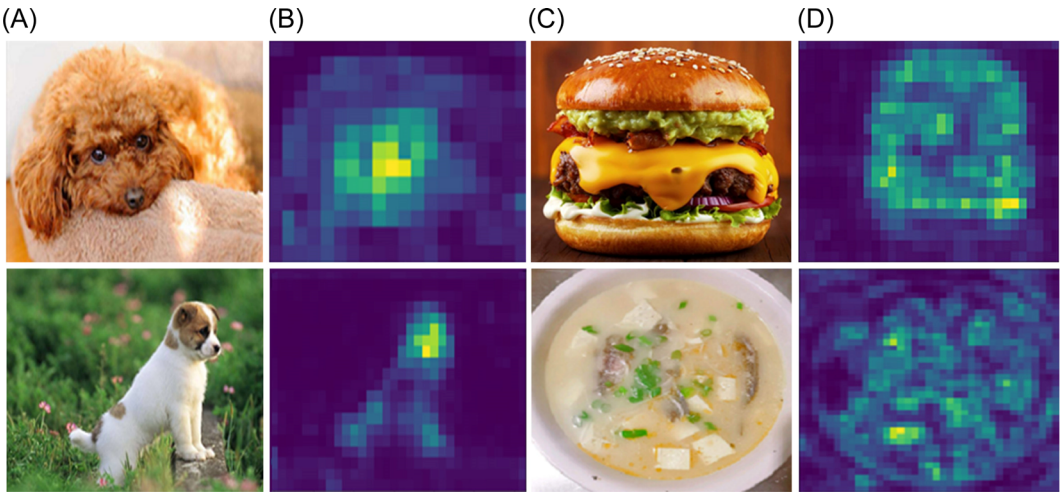

FIGURE 5 Attention distribution of general images and food images. (A, B) General images and its attention distribution and (C, D) food images and its attention distribution. [Color figure can be viewed at wileyonlinelibrary.com]

attention to a very small range. For food images, it can be found that the key features used to distinguish categories are very scattered, making it difficult to determine the category of food with a small range of attention. To deal with this situation, it is necessary to further expand the scope of attention and make it more precise.

On the basis of the fact that our method achieves a substantial improvement in the recognition rate of food images compared to the Mobile-Former method, we focus on analyzing this kind of food image: wrongly recognized in the Mobile-Former method, but correct in the LTBDNN method. Figure 6 shows the superiority of our proposed method. The four sets of example images are separated by black dashed lines, and the leftmost column of each set of images is the original food image, correctly identified in LTBDNN but misidentified by Mobile-Former. The middle column of each set of images is the attention heatmap generated by Mobile-Former method recognition, and the rightmost column of each set of images is the attention heatmap generated by LTBDNN.

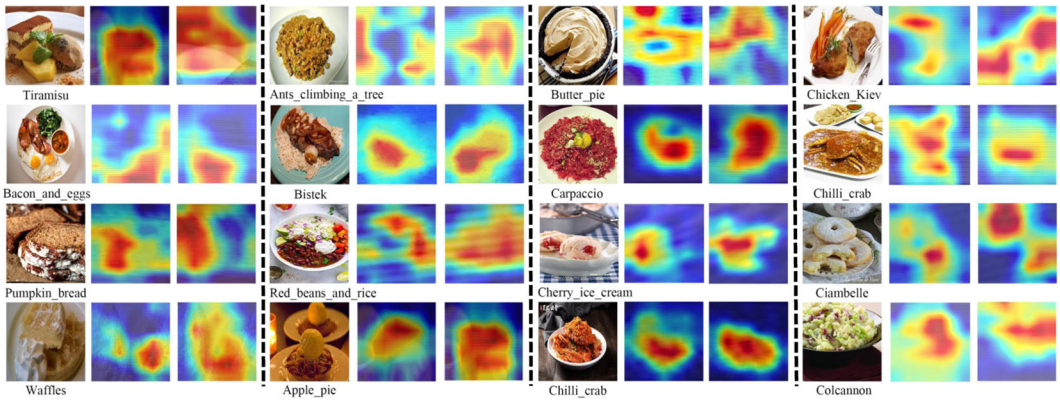

FIGURE 6 Attention span is further expanded and precise by LTBDNN. For each of the four sets of images, the left column is the original food image, the middle is the attention heatmap generated by Mobile-Former, and the right column is the attention heatmap generated by LTBDNN. LTBDNN; Lightweight Transformer-Based Deep Neural Network. [Color figure can be viewed at wileyonlinelibrary.com]

Through the heat map, it can be found that LTBDNN can effectively expand the attention range and more accurately locate the key features of food images, thereby effectively improving the recognition accuracy.

5 | CONCLUSIONS AND FUTURE WORKS

In this paper, we present an LTBDNN for food category prediction. It is capable of achieving efficient classification of food images with fewer parameters and computational cost by combining global features and local features efficiently. Especially for global features, our method uses Transformer Grouping and Token Shuffling to achieve an expanded attention span to better adapt to the characteristics of food images. Extensive evaluation on three benchmark data sets has verified its effectiveness.

Future work includes: (1) Further optimizing the deep neural network architecture based on the bidirectional bridge structure to achieve a more efficient fusion of local features and global features with a lower number of parameters and less computational consumption, and further improve the recognition accuracy of food images. (2) Continue to work on the efficient fusion of the global features extracted by the transformer and the local features extracted by the lightweight network to achieve higher recognition accuracy of food images. (3) Researching the application of lightweight neural networks to food image recognition, and further combine ingredient and recipe information to obtain personal health advice and analysis of eating habits. In the reality that people's daily life is highly dependent on mobile phones, research in this direction has great practical significance.

DATA AVAILABILITY STATEMENT

Data openly available in a public repository that issues data sets with DOIs.

ORCID

Guorui Sheng http://orcid.org/0000-0001-6790-0239
Yancun Yang http://orcid.org/0000-0003-0785-6007

REFERENCES

1. Min W, Jiang S, Liu L, Rui Y, Jain R. A survey on food computing. *ACM Comput Surv*. 2019;52(5):1-36.
2. Schäfer H, Elahi M, Elsweiler D, et al. User nutrition modelling and recommendation: balancing simplicity and complexity. In: Bieliková M, Herder E, Cena F, Desmarais MC, eds. *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP 2017, Bratislava, Slovakia, July 9–12, 2017*, ACM; 2017:93-96.
3. Elsweiler D, Trattner C, Harvey M. Exploiting food choice biases for healthier recipe recommendation. In: Degenhardt J, Kallumadi S, Rijke M, Si L, Trotman A, Xu Y, eds. *Proceedings of the SIGIR 2017 Workshop On eCommerce co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, eCOM@SIGIR 2017, Tokyo, Japan, August 11, 2017*; ACM; 2017:575-584.
4. Weiqing M, Shuqiang J, Ramesh J. Food recommendation: framework, existing solutions, and challenges. *IEEE Trans Multimedia*. 2019;22(10):2659-2671.
5. Meyers A, Johnston N, Rathod V, et al. Im2Calories: towards an automated mobile vision food diary. In: Society IEEE Computer, ed. *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015*. IEEE Computer Society; 2015:1233-1241.
6. Trattner C, Rokicki M, Herder E. On the relations between cooking interests, hobbies and nutritional values of online recipes: implications for health-aware recipe recommender systems. In: Bieliková M, Herder E, Cena F, Desmarais MC, eds. *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP 2017, Bratislava, Slovakia, July 9–12, 2017*; ACM; 2017:59-64.
7. Chen J-J, Ngo C-W, Feng F-L, Chua T-S. Deep understanding of cooking procedure for cross-modal recipe retrieval. In: Boll S, Lee KM, Luo J, et al., eds. *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22–26, 2018*; 2018:1020-1028.
8. Chen J-J, Ngo C-W, Chua T-S. Cross-modal recipe retrieval with rich food attributes. In: Liu Q, Lienhart R, Wang H, et al., eds. *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23–27, 2017*; ACM; 2017:1771-1779.
9. Aguilar E, Remeseiro B, Bolaños M, Radeva P. Grab, pay, and eat: semantic food detection for smart restaurants. *IEEE Trans Multimedia*. 2018;20(12):3266-3275.
10. Yang S, Chen M, Pomerleau D, Sukthankar R. Food recognition using statistics of pairwise local features. In: Society IEEE Computer, ed. *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010*. IEEE; 2010:2249-2256.
11. Bossard L, Guillaumin M, Gool LV. Food-101-mining discriminative components with random forests. In: Fleet DJ, Pajdla T, Schiele B, Tuytelaars T, eds. *Proceedings of the 13th European Conference on Computer Vision, ECCV 2014, Zurich, Switzerland, September 6–12, 2014, Part VI*. Springer; 2014:446-461.
12. Bettadapura V, Thomaz E, Parnami A, Abowd GD, Essa I. Leveraging context to support automated food recognition in restaurants. In: Society IEEE Computer, ed. *2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, January 5–9, 2015*. IEEE; 2015:580-587.
13. Martinel N, Foresti GL, Micheloni C. Wide-slice residual networks for food recognition. In: Society IEEE Computer, ed. *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12–15, 2018*. IEEE; 2018:567-576.
14. Yuan F, Chen S, Liang K, Xu L. *Research on the Coordination Mechanism of Traditional Chinese Medicine Medical Record Data Standardization and Characteristic Protection Under Big Data Environment*. Shandong People's Publishing House; 2021.
15. Chen J, Ngo C-W. Deep-based ingredient recognition for cooking recipe retrieval. In: Hanjalic A, Snoek C, Worring M, et al., eds. *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15–19, 2016*; ACM; 2016:32-41.
16. Min W, Liu L, Wang Z, et al. ISIA food-500: a dataset for large-scale food recognition via stacked global–local attention network. In: Chen CW, Cucchiara R, Hua X-S, et al., eds. *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event/Seattle, WA, USA, October 12–16, 2020*; ACM; 2020: 393-401.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Society IEEE Computer, ed. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*. IEEE; 2016:770-778.

18. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R, eds. *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, June 9–15, 2019, Long Beach, CA, USA*. Proceedings of Machine Learning Research. Vol 97. PMLR; 2019: 6105-6114.
19. Ai S, Koe A, Sandor V, Huang T. Adversarial perturbation in remote sensing image recognition. *Appl Soft Comput*. 2021;105(5):107252.
20. Huang T, Zhang Q, Liu J, Hou R, Wang X, Li Y. Adversarial attacks on deep-learning-based SAR image target recognition. *J Network Comput Appl*. 2020;162:102632.
21. Cai J, Qian K, Luo J, Zhu K. SARM: service function chain active reconfiguration mechanism based on load and demand prediction. *Int J Intell Syst*. 2022;37(1):6388-6414.
22. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vision*. 2004;60(2):91-110.
23. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Society IEEE Computer, ed. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), June 20–26, 2005, San Diego, CA, USA*. IEEE; 2005:886-893.
24. Joutou T, Yanai K. A food image recognition system with multiple kernel learning. In: IEEE, ed. *Proceedings of the International Conference on Image Processing*. IEEE; 2009:285-288.
25. Hoashi H, Joutou T, Yanai K. Image recognition of 85 food categories by feature fusion. In: Society IEEE Computer, ed. *12th IEEE International Symposium on Multimedia, ISM 2010, Taichung, Taiwan, December 13–15, 2010*. IEEE; 2010:296-301.
26. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet DJ, Pajdla T, Schiele B, Tuytelaars T, eds. *Proceedings of the 13th European Conference on Computer Vision—ECCV 2014, Zurich, Switzerland, September 6–12, 2014, Part I*. Springer; 2014:818-833.
27. Zhu T, Zhou W, Ye D, Cheng Z, Li J. Resource allocation in IoT edge computing via concurrent federated reinforcement learning. *IEEE Internet Things J*. 2022;9(2):1414-1426.
28. Zhu T, Li J, Hu X, Xiong P, Zhou W. The dynamic privacy-preserving mechanisms for online dynamic social networks. *IEEE Trans Knowl Data Eng*. 2022;34(6):2962-2974.
29. Li J, Ye H, Li T, et al. Efficient and secure outsourcing of differentially private data publishing with multiple evaluators. *IEEE Trans Dependable Secur Comput*. 2022;19(1):67-76.
30. Li J, Huang Y, Wei Y, et al. Searchable symmetric encryption with forward search privacy. *IEEE Trans Dependable Secur Comput*. 2021;18(1):460-474.
31. Jiang S, Min W, Liu L, Luo Z. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Trans Image Process*. 2019;29:265-276.
32. Min W, Liu L, Luo Z, Jiang S. Ingredient-guided cascaded multi-attention network for food recognition. In: Amsaleg L, Huet B, Larson MA, et al., eds. *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21–25, 2019*; ACM; 2019:1331-1339.
33. Ming Z-Y, Chen J, Cao Y, Forde C, Ngo C-W, Chua TS. Food photo recognition for dietary tracking: system and experiment. In: Schoeffmann K, Chalidabhongse TH, Ngo C-W, et al, eds. *Proceedings of the 24th International Conference on MultiMedia Modeling, MMM 2018, Bangkok, Thailand, February 5–7, 2018, Part II*. Springer; 2018:129-141.
34. Kagaya H, Aizawa K, Ogawa M. Food detection and recognition using convolutional neural network. In: Hua KA, Rui Y, Steinmetz R, Hanjalic A, Natsev A, Zhu W, eds. *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 3–7, 2014*; ACM; 2014:1085-1088.
35. Zagoruyko S, Komodakis N. Wide residual networks. arXiv preprint arXiv:1605.07146. 2016.
36. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv preprint arXiv:1602.07360*. 2016.
37. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. 2017.
38. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. In: Society Computer Vision Foundation/IEEE Computer, ed. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*; Computer Vision Foundation/IEEE Computer Society; 2018:4510-4520.
39. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Society Computer Vision Foundation/IEEE Computer, ed. *2018 IEEE Conference on Computer*

Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018; Computer Vision Foundation/IEEE Computer Society; 2018:6848-6856.
40. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 2020.
41. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30: 5998-6008.
42. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Meila M, Zhang T, eds. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, July 18–24, 2021, Virtual Event*. PMLR; 2021:10347-10357.
43. Graham B, El-Nouby A, Touvron H, et al. LeViT: a vision transformer in convNet's clothing for faster inference. In: IEEE, ed. *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021*. IEEE; 2021:12259-12269.
44. Wu H, Xiao B, Codella N, et al. CvT: introducing convolutions to vision transformers. In: IEEE, ed. *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021*. IEEE; 2021:22-31.
45. Chen Y, Dai X, Chen D, et al. Mobile-Former: bridging MobileNet and Transformer. In: Society IEEE Computer, ed. *The Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, Louisiana, USA, June 21–24, 2022*. IEEE; 2022:5270-5279.
46. Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*. 2021.
47. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. In: IEEE, ed. *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019*. IEEE; 2019: 1314-1324.
48. Ma N, Zhang X, Zheng H-T, Sun J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Proceedings of the 15th European Conference on Computer Vision—ECCV 2018, Munich, Germany, September 8–14, 2018, Part XIV*; Springer; 2018: 116-131.
49. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. GhostNet: more features from cheap operations. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020*; Computer Vision Foundation/IEEE; 2020:1580-1589.

How to cite this article: Sheng G, Sun S, Liu C, Yang Y. Food recognition via an efficient neural network with transformer grouping. *Int J Intell Syst*. 2022;37: 11465-11481. doi:10.1002/int.23050