**Frontiers in Computing and Intelligent Systems**
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

# Sichuan Cuisine Recognition Method based on Residual Neural Network

**Weifu Li, Yichong Cai and Jinyu Huang**

School of Sichuan University of Science and Engineering, Zigong 643000, China

**Abstract:** To address issues such as the high number of parameters, significant variations among images of similar dishes, weak geometric invariance, and low recognition rates in Sichuan cuisine recognition methods, a lightweight Sichuan cuisine recognition model, RGBNet, based on residual neural network, is proposed. The model employs dilated convolutions to increase the receptive field of convolutional kernels while maintaining a consistent parameter count, thus obtaining more shallow-level features. An RGB module is constructed using asymmetric convolutions to enhance the model's geometric invariance, feature non-linear expression, and feature extraction capabilities. Finally, the DFC long-range attention mechanism is introduced to effectively capture long-range information, thereby improving adaptive learning capabilities. To validate the model's performance, the classic ChineseFoodNet benchmark dataset is utilized. A MiniChineseFood dataset is created by extracting 30 classes totaling 20,000 images for experimentation. The recognition accuracy is measured using the top1 method of image recognition performance, achieving a final image recognition accuracy of 96.62%. Compared to models such as EfficientNet, ShuffNet, FasterNet, and MobileNetV2, RGBNet demonstrates respective accuracy improvements of 16.57%, 18.52%, 17.12%, and 16.35%. This presents a novel approach for industrial food recognition.

**Keywords:** Residual Neural Network; Attention Mechanism; Image Classification; Sichuan Cuisine Recognition.

## 1. Introduction

All Sichuan cuisine, also known as Sichuan-style cuisine, is a regional culinary style represented by Sichuan Province. Sichuan cuisine has profound effects on human health, nutrition, and various aspects of life. With the increase in per capita consumption levels, more researchers are turning their attention to food science, aiming to achieve health regulation by analyzing the nutritional components and ingredient combinations of dishes [1].

Methods for recognizing Sichuan cuisine have progressed from early traditional wireless RF signal methods, traditional machine learning methods, to deep learning-based recognition methods. Traditional RF methods [2] involve implanting wireless RF chips in utensils containing dishes to identify and analyze the dishes. Although RF-based methods have high accuracy, they require customizing utensils in advance, and the process is cumbersome, with limited functionality and poor maintainability. Traditional machine learning methods [3] rely on manually selecting features, conducting statistical analysis, and inputting the results into classifiers. The accuracy is not ideal. Recognition methods based on deep learning [4] are characterized by lossless, real-time, and pollution-free features. They can identify dish categories by capturing images through cameras. Compared to traditional machine learning methods, which require manual feature extraction, convolutional neural networks can automatically learn and extract features. By stacking multiple convolutional and pooling layers, abstract features are extracted layer by layer, achieving more accurate and efficient classification.

Research on food recognition is interdisciplinary, spanning fields such as computer vision [5], new media [6], industrial informatics [7], agriculture, medicine, and nutrition science [8]. The widespread use of portable devices (such as smartphones and cameras) and the development of artificial intelligence have led to extensive applications in Sichuan cuisine image recognition. Therefore, the development of real-time and accurate methods and technologies for Sichuan cuisine recognition has significant practical value. Haiyan Wang [9] and others improved local skeleton information learning by integrating asymmetric convolutions to enhance dish feature extraction. Deng Zhiliang [10] proposed a dish recognition network model that integrates multiscale features to extract semantic information from deep-level images, and calculates inter-class similarity using triplet loss. Wu Zhengdong [11] introduced a multiscale sampling module to address the limitations of fully connected layers on input sizes. Additionally, an attention-based bilinear network was proposed to construct an attention network from both channel and spatial directions to enhance feature extraction capabilities. Liao Enhong [12] addressed the issue of accuracy errors caused by the large inter-class similarity in Sichuan cuisine dish images using the maximum inter-class loss function.

Although the aforementioned methods can effectively identify dish categories, they often come with a huge number of parameters and seldom consider the issue of lightweight design. Therefore, a lightweight Sichuan cuisine recognition method is proposed by improving the residual neural network model. This method enhances the convolutional neural network backbone based on the characteristics of the Sichuan cuisine image dataset and incorporates attention mechanisms to capture pixel-level long-range relationship information. To validate the model's performance, this study conducted comparative experiments, comparing the proposed method with lightweight network models such as EfficientNet [13], ShuffNet [14], FasterNet [15], and MobileNetV2 [16].

## 2. CNN-Based Sichuan Cuisine Image Recognition

### 2.1. The Fundamental Principles of CNN

Convolutional Neural Network (CNN) [17] is a deep

learning model widely utilized in computer vision and image processing tasks. The core of CNN is the Convolutional Layer, which employs convolution operations to extract features from input data. The convolutional layer effectively captures spatial local features in the image, such as edges and textures. Simultaneously, the convolutional layer possesses characteristics of parameter sharing and sparse connections, significantly reducing the number of network parameters and enhancing computational efficiency.

Although increasing the number of network layers improves the model's generalization ability to some extent, the high time and space complexity constrain the application of deep convolutional neural networks in resource-constrained environments such as mobile phones and embedded devices [18]. To address the issue of low computational efficiency in large convolutional network models, a network structure is constructed using the residual neural network-based approach.

## 2.2. Optimizing the Design of CNN

### 2.2.1. Introducing Dilated Convolution

In dish recognition networks, the first layer's convolution operation is typically employed to extract low-level features from the input image, such as edges and color information. This layer performs convolution operations on the pixel values of the input image with convolutional kernels, resulting in a new set of feature maps that better represent the texture information of the input image.

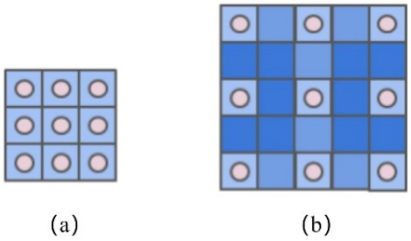


**Fig 1.** Standard Convolution and Dilated Convolution

The advantage of using dilated convolution [19] for feature extraction in the first layer lies in its ability to increase the receptive field of the convolutional kernel [20] while maintaining the output resolution. This enhancement contributes to an improved perceptual capability of the network. In comparison to standard convolution, dilated convolution also effectively reduces the number of parameters, mitigates the risk of overfitting, and accelerates the speed of convolutional computations, thereby enhancing the operational efficiency of the model. Figure 1 illustrates examples of standard convolution (Figure 1a) and dilated convolution (Figure 1b).

### 2.2.2. Fusion of Asymmetric Convolution in RGB Bottleneck

Traditional lightweight networks primarily rely on depth-wise separable convolution for feature extraction. Although this method enhances computational efficiency, splitting the convolution operation into two parts during the separable operation leads to the loss of partial multi-scale feature information.

To address these limitations, a method to improve convolutional neural networks is proposed by introducing asymmetric convolution blocks [21]. The aim is to enhance the modeling of geometric deformations by strengthening the information in the convolutional kernel skeleton, thereby improving the network's ability to model geometric deformations and enhance generalization performance.

The RGB Bottleneck block of asymmetric convolution consists of three parallel layers, each using convolutional kernels of sizes n×n, 1×n, and n×1 to slide and extract features. After convolution, Batch Normalization is applied to the outputs of the three branches, and then the outputs of each branch are summed to obtain a rich feature space. Non-square convolution layers, such as 1×d and d×1, are utilized. The additivity property of convolution is leveraged, as shown in the following formula (Equation 1).

$$A = \mathrm{C} * \mathrm{K}_1 + \mathrm{C} * \mathrm{K}_2 + \cdots + \mathrm{C} * \mathrm{K}_p = C * \left(\mathrm{K}_1 \oplus \mathrm{K}_2 \oplus \cdot \cdot \oplus \mathrm{K}_p\right) = C * K \quad (1)$$

Where A represents the equivalent output, C is the input, K is the 2D convolutional kernel, and P is the number of convolutional kernels, if there exist P size-compatible 2D kernels (Kp) that, when applied with the same stride on the same input C, generate outputs with the same resolution, and if the sum of these outputs is denoted as A, then the corresponding kernels at each position can be summed to form an equivalent kernel K. This equivalent kernel K produces the same output A when applied to the same input.

The RGB bottleneck structure is illustrated in Figure 2, where Figure 2(a) depicts the RGB bottleneck structure with a stride of 1, and Figure 2(b) represents the RGB bottleneck structure with a stride of 2.

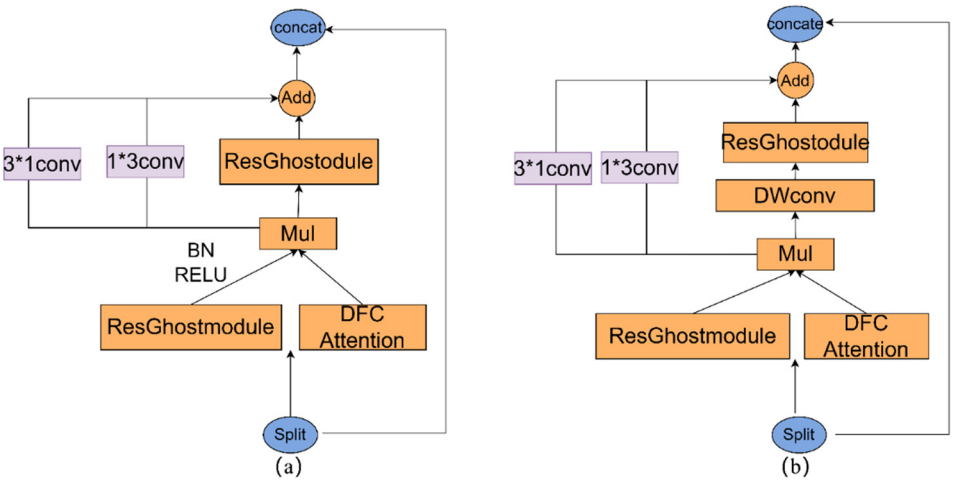


**Fig 2.** RGB bottleneck

### 2.2.3. Attention Module to Enhance Geometric Deformation Performance

By introducing a decoupled fully connected (DFC) attention mechanism branch, implemented with asymmetric convolutions employing unequal horizontal and vertical convolutions [22], into the RGB bottleneck structure, the model captures richer image features using distinct convolution kernels in different directions. This significantly enhances the model's capability to capture long-range spatial information and representation power in images.

For a given input $Z \in \mathbb{R}^{H\times W\times C}$, which can be regarded as a tensor of size H*W, the mathematical expression for implementing the attention map using an ordinary fully connected layer is shown in Equation (2):

$$a_{hw} = \sum_{h'w'} F_{hw,h'w'} \odot Z_{h'w'} \quad (2)$$

In the equation, $\odot$ denotes element-wise multiplication, represents the learnable weights in the fully connected layer, and is the obtained attention map, and Zis the original feature.

By decoupling Equation (1) along both the horizontal and vertical directions, long-range correlations in both directions can be captured, resulting in attention weights. The feature aggregation processes in the horizontal and vertical directions are illustrated in Equations (3) and (4).

$$a'_{hw} = \sum_{h'=1}^{H} F^{H}_{h,h'w} \odot z_{h'w}, h = 1,2\ldots,H, w = 1,2\ldots W \quad (3)$$

$$a_{hw} = \sum_{w'=1}^{W} F^{W}_{w,hw'} \odot a'_{hw'} h = 1,2,\ldots,H, w = 1,2,\ldots,W \quad (4)$$

In the equation, In the formula, $F^H$ represents horizontal weights, $F^W$ represents vertical weights. The DFC (Decoupled Fully Connected) attention mechanism is illustrated in Figure 3.

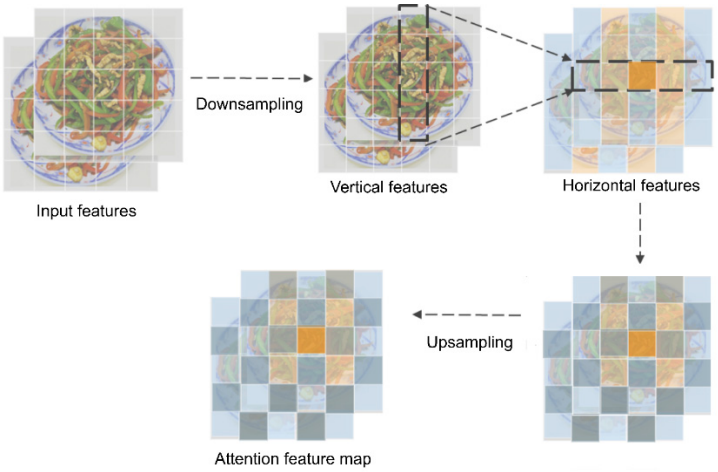


**Fig 3.** DFC attention mechanism

## 3. RGBNet Model Architecture

Considering the small inter-class differences and large intra-class differences in the Chinese cuisine dataset, a lightweight recognition approach for Sichuan dishes is proposed, aiming to address the task of recognizing Sichuan cuisine images on edge devices such as mobile phones. The model consists of three main parts. Firstly, in the first layer of the network, dilated convolution is employed for feature extraction, significantly reducing the model's parameter count while maintaining accuracy, thereby enhancing computational efficiency. The second part is primarily composed of the ResGhost Bottleneck structure proposed in this paper, which employs cost-effective operations [23] to break down larger convolutional layers into subnetworks with shared weights. Residual convolution is utilized to improve the model's generalization ability and efficiency. The third part introduces the Decoupled Fully Connected attention mechanism (DFC), designed to facilitate long-range communication. It incorporates dynamic normalization parameters to adjust feature values in different regions, ultimately achieving information exchange and feature calibration between different regions. The structure of the established RGBNet network is shown in Table 1.

**Table 1.** RGBNet Network Architecture

| Input features | Operation | Repetition count | Output depth | Stride |
|---|---|---|---|---|
| 2242 x 3 | 5x5dilated convolution | 1 | 16 | 2 |
| 1122 x 16 | RGB Bottlenet | 2 | 24 | 2 |
| 562 x 24 | RGB Bottlenet | 2 | 40 | 2 |
| 282 x 40 | RGB Bottlenet | 4 | 80 | 2 |
| 142 x 112 | RGB Bottlenet | 6 | 160 | 2 |
| 72 x 160 | RGB Bottlenet | 4 | 160 | 1 |
| 72 x 160 | 1x1conv | 1 | 960 | 1 |
| 72 x 960 | 7x7 Average pooling | 1 | - | - |
| 12 x 960 | 1x1conv | 1 | 1280 | 1 |
| 12 x 1280 | FC Convolution | 1 | 30 | - |

## 4. MiniChineseFood Dataset

ChineseFoodNet [24] is a large-scale dataset of food images, comprising 208 classes with a total of 180,000 images featuring various culinary styles from different regions in China. Each dish is represented by images capturing significant variations in angles, lighting conditions, and plating. However, due to the dataset's inclusion of a substantial number of visually distinct dishes, which tend to achieve high scores during network training, it lacks representativeness.

To address this, 30 classes were extracted from the ChineseFoodNet dataset, resulting in a practical set of 20,000 images. Through data augmentation techniques, including random augmentation [25] and random erasing [26], the original image count was expanded to 100,000 images. This extended dataset is named MiniChineseFood, and it was divided into a training set and a test set in a 4:1 ratio for model training. Some images from the MiniChineseFood dataset are shown in Figure 4.


**Fig 4.** MiniChineseFood Dataset

The selected images from the ChineseFoodNet dataset adhere to the following criteria: ensuring low intra-class similarity among extracted dish images, such as different shapes of ingredients (a and b), varied plating (c and d), and distinct types of ingredients despite having the same name (e, f); maintaining high inter-class similarity among extracted dish images, for instance, different types of ingredients (g and h), different cooking methods (i and j) for the same ingredient, and dishes with different ingredients and cooking methods that appear similar but belong to different categories (k and l).

## 5. Experimental Design and Analysis

### 5.1. Model Training and Result Analysis

To evaluate the performance of the RGBNet network in recognizing Sichuan cuisine images, this study selected representative lightweight CNN networks for performance comparison, including MobileNetV2, ShuffleNet, EfficientNet, and FasterNet.

In the RGBNet, the optimizer is SGD, with a learning rate (lr) of 0.045, momentum set to 0.9, and weight decay (weight_decay) of 4e-05. Additionally, gradient clipping was not applied. The learning rate was configured using a "step" decay strategy, where the learning rate is multiplied by a parameter gamma (set to 0.98) after the first epoch and gradually reduced in subsequent epochs to fine-tune the model parameters. The changes in model accuracy and loss function values are illustrated in Figure 5.

From the graph, it can be observed that selecting top-1 accuracy as the evaluation metric, at the beginning of training, RGBNet, like other models, exhibits relatively low accuracy. However, starting from the 10th epoch, RGBNet's accuracy growth rate gradually surpasses other networks and reaches its peak at the 196th epoch. In comparison to other models, which achieve convergence around the 85th epoch, RGBNet converges more slowly. This is attributed to the ResGhost module requiring more data for learning to achieve better model fitting capability. After thorough learning, all performance metrics surpass those of the comparative models, strongly indicating that RGBNet possesses robust feature extraction and fitting capabilities.

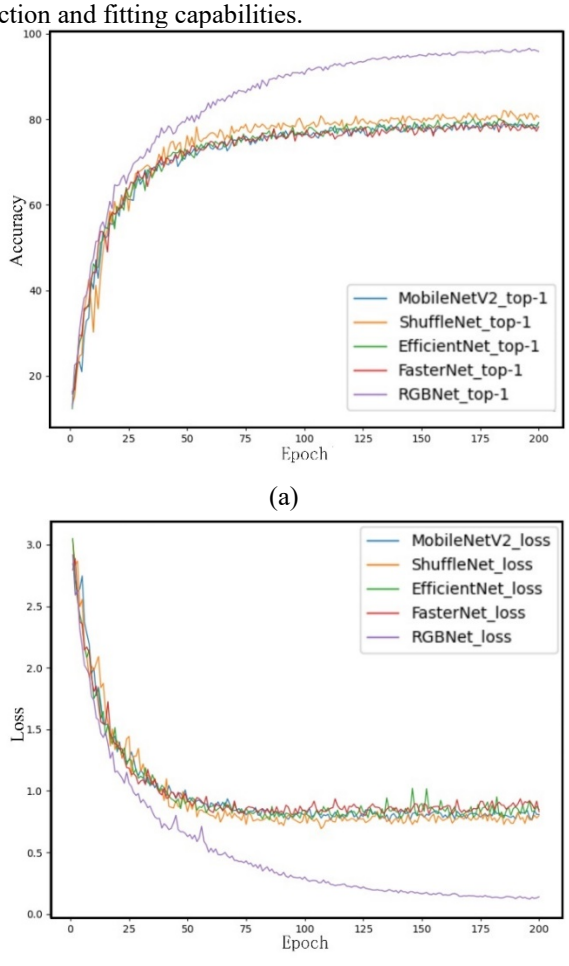


**Fig 5.** (a) Model Accuracy Curve, (b) Model Loss Curve

Table 2 presents the performance of the model on the test set after training. Compared to other models, the RGBNet demonstrates excellent performance, achieving an accuracy of 96.72%, recall of 0.9646%, precision of 0.9739%, and an F1 score of 0.9625%.

**Table 2.** Data Results

| Model Name | Accuracy Macro | Recall Macro | Precision Macro | F1 Score |
|---|---|---|---|---|
| MobileNetV2 | 80.37% | 0.8056% | 0.8173 | 0.8048 |
| ShuffleNet | 78.20% | 0.7726% | 0.7812 | 0.7719 |
| EfficientNet | 80.15% | 0.7969% | 0.8116 | 0.8023 |
| FasterNet | 79.60% | 0.7893% | 0.7965 | 0.7847 |
| RGBNet | 96.72% | 0.9646% | 0.9739 | 0.9625 |

### 5.2. Ablation Experiment

To verify the impact of dilated convolution and DFC attention mechanism on model recognition accuracy, RGBNet (conv2d), RGBNet (Dilated Conv), and RGBNet (Dilated Conv+DFC) were selected for ablation experiments. In RGBNet (conv2d), the first layer uses regular 3×3 convolution; in RGBNet (Dilated Conv), the first layer uses dilated convolution; and in RGBNet (Dilated Conv+DFC), DFC attention mechanism is incorporated alongside dilated convolution. The experimental results are shown in Table 4. This fully demonstrates that integrating DFC effectively captures long-range information, enhancing accuracy, and combining dilated convolution with the lightweight RGB module can achieve better recognition results.

**Table 3.** Ablation Experiment

| Model Name | Accuracy Macro | Recall Macro | Precision Macro | F1 Score |
|---|---|---|---|---|
| RGBNet | 90.44% | 0.9026% | 0.9089% | 0.9018% |
| RGBNet (Dilated Conv) | 91.34% | 0.9147% | 0.9226% | 0.9116% |
| RGBNet (Dilated Conv+ DFC) | 96.72% | 0.9646% | 0.9739% | 0.9621% |

## 6. Conclusion

For the Sichuan cuisine recognition task, a lightweight RGBNet network model based on a residual neural network is proposed. In the backbone network, standard convolution is channel-split to merge multiple asymmetric convolutions, acquiring richer features from multiple directions. Subsequently, DFC long-range attention is introduced to capture long-distance dependencies in sequences. Compared to existing convolutional neural networks, the proposed model performs better on the MiniChineseFood dataset.

However, there are still areas for further improvement in the deep learning approach used by the model. Firstly, due to the diverse nature of Sichuan cuisine dishes and variations in ingredient types and cooking styles, the data samples extracted in the experiments are relatively singular. In future research, more advanced object detection algorithms will be integrated to handle multi-object cuisine datasets. Secondly, although RGBNet achieves good performance in lightweight models, its feature representation capability is weaker compared to some complex models. This can be addressed by incorporating more network layers and using more complex features to enhance the algorithm for better practical applications in real life.

## Acknowledgments

Sichuan Province's Returnee Science and Technology Activities Project for Overseas Students; Talent Introduction Project of Sichuan University of Science and Technology (2021RC13)

## References

[1] Min Weiqing, Liu Linhu, Liu Yuxin, et al. A Survey of Food Image Recognition Methods[J]. Journal of Computer Science. Vol.45 (2008) No. 03, p.542-566.

[2] Wang Hang. Application of RFID Technology in University Cafeteria Practice[J]. China's Strategic Emerging Industries Vol. 172 (2008) No. 40, p. 256.

[3] Chen Mei, Kapil Dhingra, Wu Wen, Yang Lei, et al. PFID: Pittsburgh fast-food image dataset[C]// Proceedings of the International Conference on Image Processing. Cairo, Egypt, 2009: 289-292.

[4] Martinel N, Foresti G L, Micheloni C. Wide-slice residual networks for food recognition[C]// Proceedings of the International Conference on Applications of Computer Vision. Lake Tahoe, USA,2018:567-576.

[5] Austin Meyers, Nick Johnston, Vivek Rathod, et al. Im2Calories: towards an automated mobile vision food diary [C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile,2015:1233–124.

[6] Deng L, Chen J., Sun Q, He, X. et al. Mixed-dish recognition with contextual relation networks[C]//Proceedings of the 27th ACM International Conference on Multimedia,2019:112-120.

[7] Xiao G, Wu Q, Chen H, et al. A deep transfer learning solution for automating food material procurement using electronic scales[J]. IEEE Transactions on Industrial Informatics, 2019, 16 (4):2290-2300.

[8] Lo F P W, Sun Y, Qiu J, et al. Image-based food classification and volume estimation for dietary assessment: A review[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(7):1926-1939. DOI:10.1109/JBHI.2020.2987943.

[9] Wang Haiyan, Zhang Miao, Liu Hulin, et al. Chinese Cuisine Image Recognition Method Based on Improved ResNet Network[J]. Journal of Shanxi University of Science and Technology. Vol. 40 (2022) No. 01, p. 154-160.

[10] Deng Zhiliang, Li Lei. Chinese Dish Recognition Model Based on Improved Residual Networks[J]. Progress in Laser and Optoelectronics. Vol. 58 (2021) No. 06, p. 264-272.

[11] Wu, Zhengdong. Research on Image Classification Algorithm of Chinese Cuisine (Master, University of Electronic Science and Technology of China, Chengdu, China, 2020). p. 25-33.

[12] Liao Enhong, Li Huifang, Wang Hua, et al. Food Image Recognition Based on Convolutional Neural Networks. Journal of South China Normal University (Natural Science Edition). Vol. 51 (2019) No. 4, p. 113-119.

[13] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR ,2019:6105-6114.

[14] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices [C]// Proceedings of the IEEE conference on computer vision and pattern recognition.2018:6848-6856.

[15] Chen J, Kao S, He H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[J]. arXiv preprint arXiv: 2303. 03667, 2023.

[16] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[17] Gai Rongli, Cai Jianrong, Wang Shiyu, et al. A Comprehensive Review on the Application of Convolutional Neural Networks in Image Recognition[J]. Small & Miniaturized Computer Systems. Vol. 42 (2021) No. 9, p. 1980-1984.

[18] Lin Jingdong, Wu Xinyi, Chai Yi, et al. Overview of Optimizing Convolutional Neural Network Structures. Acta Automatica Sinica. 2020, 46(1): 24-37.

[19] Wu Haohao, Wang Fangshi. Application of Multi-Scale Dilated Convolution in Image Classification. Computer Science. 2020, 47(6A): 166-171.

[20] Li Y, Zhang X, Chen D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes[C]// Proceedings of the IEEE conference on computer vision and pattern recognition.2018:1091-1100.

[21] Ding X, Guo Y, Ding G, et al. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks [C]// Proceedings of the IEEE/CVF international conference on computer vision.2019:1911-1920.

[22] Tang Y, Han K, Guo J, et al. GhostNetV2: Enhance Cheap Operation with Long-Range Attention[J].arXiv preprint arXiv: 2211. 12905, 2022.

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

[24] Chen X, Zhu Y, Zhou H, et al. Chinesefoodnet: A large-scale image dataset for Chinese food recognition[J].arXiv preprint arXiv: 1705.02743, 2017.

[25] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops.2020:702-703.

[26] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI conference on artificial intelligence.2020,34(07):13001-13008.