Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230300

Lightweight food classification model based on MobileViT and ULSAM

Tian Zhang

Shanghai Film Academy, Shanghai University, Shanghai, China, 200072

2876187491@qq.com

Abstract. This paper presents a novel approach to enhance the image classification performance by incorporating Unified Local-Scale Attention Module mechanism into the lightweight MobileViT architecture. The MobileViT+ Ultra-Lightweight Subspace Attention Module model achieved remarkable accuracy on the ISIA food-500 dataset, while maintaining computational efficiency and parameter quantity similar to the original MobileViT model. Moreover, the MobileViT+ Ultra-Lightweight Subspace Attention Module model outperforms other lightweight models such as MobileNetV2 and LCNet. The ablation experiments confirmed the effectiveness of Ultra-Lightweight Subspace Attention Module in enhancing classification accuracy and its ability to uniformly optimize multiple model structures. Additionally, this paper explored a more lightweight model that significantly reduced FLOPs and parameter quantity while maintaining the same model performance. Overall, this research provides a practical and resource-efficient approach for improving image classification performance in various deep learning.

Keywords: food classification, lightweight, subspace attention, ISIA food-500, MobileViT.

1. Introduction

Recent advancements in machine learning and deep learning have enabled researchers to make significant progress in several fields. Deep learning techniques have been successful in various visual perception tasks, including object and action recognition, image segmentation, super-resolution, and visual question answering, while being increasingly used in everyday life.

Object recognition and semantic recognition, including popular models such as Generative Pre-trained Transformer and diffsinger, are examples of the use of machine learning and deep learning techniques in everyday life [1-4]. The recognition and classification of food is one of the sub-tasks of object recognition. Often when people encounter a new food, they want to identify it, so meeting this need requires a lightweight food recognition network model that can be used on mobile and embedded devices. Related work in computer vision includes food classification, recipe generation , and food image retrieval [5-12].

While deep learning technology has achieved better accuracy on several datasets, certain image classification problems, especially for similar types of food, can still pose a challenge to current methods, primarily due to variations resulting from different production methods. Therefore, we wanted to explore a more effective approach to food image recognition. Nonetheless, since the main objective of this article is to allow users to quickly identify broad food categories.

DOI: 10.54254/2755-2721/27/20230300

Addressing the above-mentioned challenges involves finding a suitable model for task classification and optimizing it for mobile and embedded devices. Some studies have achieved impressive results in food classification using deep neural networks optimized for this task [13-14]. Additionally, many lightweight models (MobileViT [15], EfficientNet [16], LCNet [17]) have been developed for traditional tasks with satisfactory results. However, currently, combining both approaches is not ideal, as the models with high accuracy are large and cumbersome, whereas the models with low complexity lack accuracy.

This study propose a novel method that combines the MobileViT model with a new spatial attention module called the Ultra-Lightweight Subspace Attention Module (ULSAM) to enhance its ability to capture spatial information for food image classification. The author proposed model achieves high accuracy while maintaining low computational complexity and parameter quantity, making it a practical and efficient alternative for deep learning applications.

2. Related work

2.1. *Food ecognition*

Food recognition for healthcare applications is receiving more attention, following the success of image recognition. Yang segmented each image into eight different types of components using Semantic Texton Forest and classified them with SVM, making it one of the earliest works on food recognition [18]. Recently, Martinel et al revealed that food has unique characteristics in the vertical direction, by introducing a sliced convolutional block to capture the food layer and merge it with the depth residual block output [13]. Qiu et al. created a model based on adversarial erasure that concentrates on "maintaining the basic accuracy of classifying input images" and "adversarial Mining discriminative food regions" with the help of assisted adversarial networks [14]. Moreover, they introduced the Sushi-50, a new, fine-grained food dataset.

2.2. Light weighted model

Since 2017, numerous practical and substantial lightweight network architectures have emerged. Google proposed the MobileNet model as a lightweight network architecture for mobile devices [10-21]. It utilizes the depthwise separable convolution technique and the residual structure by performing Expansion and Projection operations, among others.

It also constructs an inverted residual network module, which is referred to as the Inverted residual block. In an effort to enhance the convolutional feature expression capability of mobile networks and resolve the limitations of the channel attention mechanism (such as SE), CANet introduced a novel attention mechanism referred to as Coordinate Attention [16].

LCNet, also a lightweight network architecture, enhances accuracy without a corresponding increase in inference time [17]. Combining these strategies results in a more optimized balance between accuracy and speed. Lastly, the MobileViT network focuses on skillfully combining the inductive bias advantages of CNNs and the global receptive field capability of ViT, resulting in a lightweight, general-purpose, and low-latency end-to-side network architecture [15].

2.3. Deep learning for food recognition

Tahir et al. proposed an open continuous learning framework that uses transfer learning to extract deep features, Relief F for feature selection, and an adaptive degraded incremental kernel extreme learning machine (ARCIELM) for classification [22]. The transfer learning approach capitalizes on the high generalization ability of deep learned features, while Relief F reduces computational complexity by sorting and selecting the most significant features.

MSMVFA combines high-level semantic features, mid-level attribute features, and deep visual features to create a unified representation that captures food image semantics with the highest probability[23]. Horiguchi et al. suggested a personalized incremental learning framework for each user,

DOI: 10.54254/2755-2721/27/20230300

which combines the nearest class mean classifier and 1-nearest neighbor classifier using deep features to address the personalized food classification problem [24].

2.4. Dataset

Datasets play a critical role in the development and evaluation of food image recognition models. ISIA food-500 is a fine-grained food dataset with 500 categories and almost 400,000 images sourced from Wikipedia [25]. Despite providing comprehensive coverage of food categories and data volume, many classes exhibit high similarity, making it difficult to classify images accurately. To simplify the classification task, the authors combined 80 classes by either merging or deleting subclasses belonging to the same superclass. Each class comprises approximately 450 to 1000 images to ensure a well-balanced distribution. Despite aggregation, the dataset retains certain fine-grained features. Some samples are shown in the figure 1.

Figure 1. Some images from the ISIA-food500 dataset.

3. Networks and methods

3.1. MobileViT based

ViT was a model proposed in 2020 by Google's team for applying the Transformer to image classification [27]. Although previous studies have attempted to use Transformer for visual tasks, the ViT model's effectiveness lies in its simplicity, versatility, and scalability.

The food image classification task is a challenging and relevant computer vision problem that involves recognizing food items based on their type, style, cooking method, and ingredients, requiring a model with strong detail perception and abstraction skills, particularly in complex and diverse conditions.

Due to the mobile nature of such tasks, an ideal model must be both lightweight and fast. Therefore, this paper propose the MobileViT model, providing efficient high-precision results at minimal cost [15]. In previous studies, the patch is projected, and the Transformer subsequently learns global information between patches, leading to a loss of the image's inductive bias. Consequently, these models need more parameters and wider and deeper models. However, the MobileViT, with its convolution and Transformer structures, combines the Transformer's global modeling and CNN's inductive bias, requiring fewer parameters than ViT, making it mobile-friendly.

Furthermore, the MobileViT blocks effectively encode both local and global information, with the added benefit of having differing perspectives for global representation learning. Standard convolution entails un-folding, local processing, and folding. Figure 2 shows the substrate structure.

DOI: 10.54254/2755-2721/27/20230300

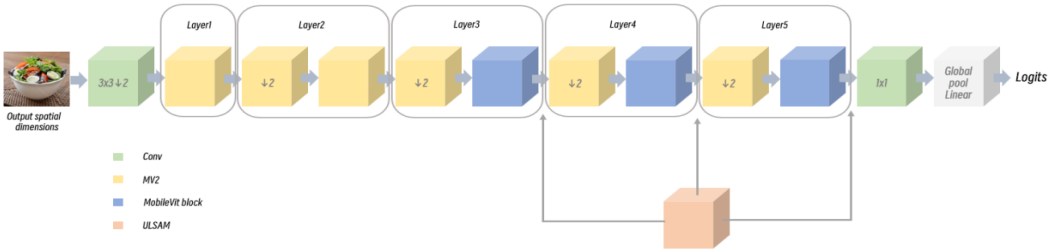

Figure2. The basic network structure of MoblieViT and where the ULSAM module is added.

The MV2 module utilizes two convolution kernels of size 1x1 and a convolution kernel of size 3x3 to perform deep separable convolution on the input feature when the convolution step size is set to 2. Conversely, a residual connection is introduced when the step size is set to 1 to prevent potential problems like gradient explosion or disappearance that can be caused by an excessive number of output features. This facilitates network parameter propagation across layers. Moreover, the MV2 module follows the upsample-extract-downsample approach, which involves an inverse operation to upsample the input data, the subsequent feature extraction using deep separable convolution, and lastly, downsampling, ultimately maintaining the input data dimension, significantly reducing computational burden and model parameters.

3.2. ULSAM block

Physical food images exhibit strong spatial features, such as soup foods that are mainly contained in bowls or pots, where the container is situated below the food. Similarly, pastries and breads items exhibit a strong vertical hierarchy, with certain food presentations creating spatial interrelationships. The spatial attention mechanism provides many advantages for the performance and accuracy of food recognition and classification in several ways. Firstly, it focuses the neural networks on critical image areas, such as different ingredients and side dishes, improving the visibility and classification accuracy. Secondly, it enables multi-scale processing, which is critical for food images containing ingredients and side dishes of varying sizes and proportions. By doing so, it better perceives image details and local features. Thirdly, it enhances interpretability and visualizes the processing actions of the neural network through the pixel attention weight calculation. Therefore, we chose an ultra-lightweight subspace attention mechanism called ULSAM [28]. The large computational overhead and parameter number associated with existing attention mechanisms are not desirable in compact CNNs. ULSAM effectively learns cross-channel interdependencies for each feature map subspace, making it the first attention module to achieve this goal. In particular, it divides extracted features into g groups, spatially recalibrating each subfeature of a group, and finally joining the g group features together. The approximate structure of the model is shown in the figure 3.

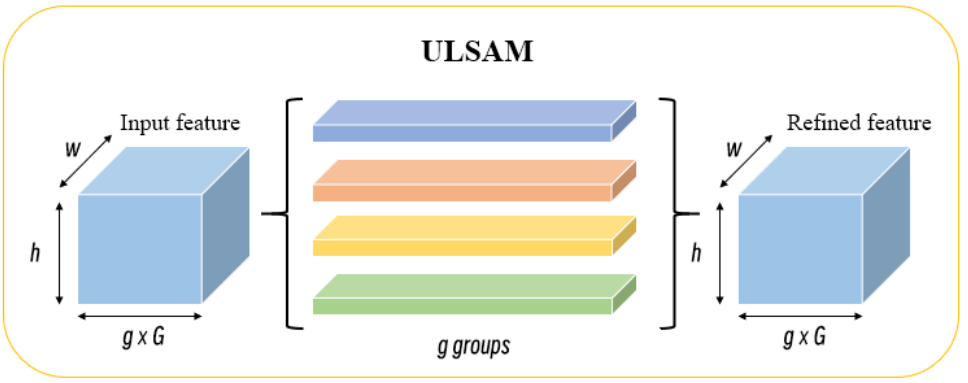

Figure 3. ULSAM Block.

DOI: 10.54254/2755-2721/27/20230300

3.3. MobileViT+ ULSAM

It is crucial to integrate ULSAM into the framework structure of MobileViT. Deeper layers are associated with global features that possess stronger location information and contain semantic information with coarser spatial details while having smaller dimensions than that of the feature maps in the initial layer. Moreover, evidence suggests that adding ULSAM to a shallower layer creates a loss of effective information and interferes with spatial relationships [28], hindering the training effect. Therefore, integrating self-attention in deeper layers is beneficial for better learning of global information interactions. MobileViT comes with its MobileViT Block, and including a spatial attention module after the block can enhance the model's perception of spatial relationships.

This paper presents three insertion methods(Figure 2) and the specific parameters are shown in the table1, table2, table3:

V1: adding the spatial attention module after layer 5,

V2: adding it after layers 4 and 5,

V3: adding it after layers 3, 4, and 5.

Table1. MobileVit + ULSAM V1.

Layer(V3)	out_channels	dim	g
layer1	16	128×128	None
layer2	24	64×64	None
layer3	48	32×32	None
layer4	64	16×16	None
layer5	80	8×8	None
ULSAM	80	8×8	16

Table2. MobileVit + ULSAM V1.

Layer(V4)	out_channels	dim	g
layer1	16	128×128	None
layer2	24	64×64	None
layer3	48	32×32	None
layer4	64	16×16	None
ULSAM	64	16×16	8
layer5	80	8×8	None
ULSAM	80	8×8	16

Table3. MobileVit + ULSAM V1.

Layer(V5)	out_channels	dim	g
layer1	16	128×128	None
layer2	24	64×64	None
layer3	48	32×32	None
ULSAM	48	32×32	4
layer4	64	16×16	None
ULSAM	64	16×16	8
layer5	80	8×8	None
ULSAM	80	8×8	16

DOI: 10.54254/2755-2721/27/20230300

4. Experiment

4.1. Dataset preprocessing

To start the image preparation for training, the dataset was split into two sets: one for training and the other for validation, in a 4:1 ratio respectively. For the training set, we perform center cropping of the images to resize them into 224×224 pixels, and augment the dataset by randomly flipping them horizontally to enhance its diversity. On the other hand, for the validation set, images are resized to 224×224 pixels without cropping for consistency.

Then, we convert all samples in both sets by transforming them into tensors, which is a multi-dimensional array that represents the numerical data of images. Subsequently, the tensor values are normalized with parameters ([0.616, 0.520, 0.418], [0.232, 0.248, 0.268]) for effective learning convergence. The normalization serves to standardize the mean and variance of the pixel values, making them easier for the model to learn in a consistent and stable manner.

4.2. Hardware and environment

The proposed method in this article was tested on a computer system equipped with an Intel Core i7 13700k CPU, 32GB of memory, and a NVIDIA GeForce RTX 3090Ti GPU. Specifically, the software environment used was a 64-bit Windows 11 operating system, and the code was developed using the PyTorch framework, a popular deep learning framework for machine learning research.

4.3. Experimental setup

For the training process, we used MobileViT as the baseline architecture, and selected XXS for scaling the model for its lightweight nature. On the ISIA food-500 dataset, we used the stochastic gradient descent (SGD) optimizer with a batch size of 256, a momentum of 0.9, and a weight decay rate of 5 × 10^−4 with an initial learning rate of 0.045. The SGD optimizer is an efficient algorithm for optimizing large scale deep learning models, and the author found these hyperparameters to produce good results in our experiments.

To tackle the issue of diverging or slow convergence caused by an excessively large or small learning rate respectively, this paper adopts a cosine annealing algorithm that dynamically modifies the learning rate.

The cosine annealing algorithm partitions the training process into several cycles, each with a pre-defined initial and minimum learning rate. For each cycle, a cosine function value is computed as a ratio of the current learning rate to the initial learning rate, based on the current training step and the overall iteration. The formula for calculation follows:

$$lr = lr_{min} + 0.5 \cdot (lr_{max} - lr_{min}) \cdot \left(1 + \cos\frac{\pi \cdot iter}{total_{iter}}\right) \quad (1)$$

In the equation, lr denotes the learning rate in the current iteration, $\mathrm{lr_{min}}$ represents the minimum learning rate, $\mathrm{lr_{max}}$ represents the initial learning rate, iter denotes the number of completed training steps, and $\mathrm{total_{iter}}$ represents the total number of training steps.

At the start of every cycle, the current learning rate is identical to the initial learning rate. As training steps progress, the current learning rate gradually decreases and aligns with the minimum value in the shape of a cosine wave. Eventually, at the end of each cycle, the current learning rate matches the minimum rate. This approach continues throughout the entire process, allowing for successful convergence. In the experiment, each cycle lasted 130 epochs, accumulating to a total of 390 epochs.

In addition, we compared our algorithm to several other established models, including MobileNetV2, ResNet, and LCNet. All of the models followed the same training methodology adopted in this paper.

4.4. Results

According to the Table4, the model frameworks of V1, V2, and V3 versions have achieved respective improvements of 0.8%, 1,0%, and 1.3% compared to the original model. Furthermore, it can be seen

DOI: 10.54254/2755-2721/27/20230300

that V3 outperforms V2, while V2 performs better than V1. Therefore, it can be concluded that incorporating ULSAM after the MobileViT module in a deep network structure can enhance the model's ability to capture spatial information. This demonstrates that the method in this research can effectively leverage the feature extraction capabilities of MobileVit, and improve the performance of image classification by enhancing the spatial attention mechanism through ULSAM. This method not only outperforms the original MobileViT in accuracy but also maintains a similar level of computing complexity and parameter quantity, demonstrating its high practical value.

Table4. Comparison of the original MobileVit with three different methods of adding ULSAM.

Models	ACC	FLOPs	Params
MobileViT	0.810	318.83M	0.98M
MobileViT+ULSAM*V1*	0.818	318.86M	0.98M
MobileViT+ULSAM*V2*	0.820	318.96M	0.98M
MobileViT+ULSAM*V3*	0.823	319.24M	0.98M

Compared with ResNet, the MobileVit+ULSAM(Table5) model exhibits significant advantages in terms of FLOPs (floating-point operations per second) and parameter quantity, while also demonstrating superior training effectiveness. These results indicate that the MobileVit+ULSAM model can substantially decrease the demand for computing resources without compromising model performance, providing a resource-efficient and high-accuracy alternative for deep learning applications. In addition, the MobileVit+ULSAM model outperforms other lightweight network models in terms of accuracy, achieving a 5.5% increase in accuracy compared to MobileNet. Although LCNet incorporates a spatial attention module and performs well, there is still 1.0% difference compared to the method proposed in this paper. These findings provide evidence of the potential of MobileVit+ULSAM model in diverse deep learning applications.

Table5. MobileVit + ULASM V3 compared to other networks.

Models	ACC	FLOPs	Params
MobileNet V2	0.768	333.05M	2.33M
ResNet34	0.819	3682.01M	21.33M
LCNET	0.813	167.19M	1.77M
MobileViT+ULSAM*V3*	0.823	319.24M	0.98M

4.5. Ablation experiments

The effectiveness of ULSAM was further tested and verified by replacing the MobileViT Block with a regular MV2 Block (named MobileViT) and evaluating the performance of the model (MobileViT and MobileViT+ULSAM) with and without the ULSAM module, as shown in Table 6. This experiment demonstrates that the ULSAM module can optimize food images for classification with or without the MobileViT module. Moreover, the MobileViT Block also contributes to this classification task.

The MobileViT Block in the MobileViT model was substituted with a regular MV2 Block and the renamed model MobileViT was used for further comparison of the performance impact of ULSAM modules under two conditions: with ULSAM retained after layer 3, layer 4, and layer 5, and without ULSAM.

Experiments have shown that the ULSAM module improves food image classification performance, independent of whether or not the MobileViT Block is present. This indicates that the ULSAM module can consistently optimize different model structures, enhancing their performance and robustness.

Upon comparison, we discovered that after substituting the Mobile Vision Transformer Block with the mv2 structure, the classification accuracy of the model decreased to 76.0%. However, upon adding three layers of ULSAM structure the accuracy significantly improved to 81.7%, surpassing the original accuracy of the MobileViT. Moreover, this network structure's FLOPs are one-third of the original and

DOI: 10.54254/2755-2721/27/20230300

the number of parameters is only one-fifth of the Mobile Vision Transformer Block. These features allow for further lightweighting and improve the accuracy in line with the original model.

Table 6. Pilot Experiments.

Models	ACC	FLOPs	Params
MobileViT*	0.760	105.60M	0.21M
MobileViT	0.811	318.83M	0.98M
MobileViT*+ULSAM*V3*	0.817	106.00M	0.21M
MobileViT+ULSAM*V3*	0.823	319.24M	0.98M

5. Conclusion

The current experimental findings suggest that MobileViT plus ULSAM possesses certain advantages in real-world image classification tasks, yet it still faces certain limitations. To enhance the model's performance and generalization capability, this study intends to undertake the following tasks in the future.

Firstly, we aim to investigate new data augmentation techniques, such as color perturbation and rotation, to diversify and challenge the dataset, which will enhance the model's resilience to noise and alterations. Secondly, we plan to evaluate various optimizers and learning rate adjustment strategies, including Adam, RMSProp, and Warmup, to determine the optimum parameters and training process for MobileViT + ULSAM. Thirdly, it is planned to further scrutinize and visualize MobileViT plus ULSAM through methods like Grad-CAM or Saliency Map to understand the regions and features that the model prioritizes. Employing methods such as t-SNE or UMAP to present the clustering and feature space learned by the model is also under consideration. Fourthly, we aim to use additional food datasets, such as Fruit360, Food2K, and ChineseFood or Caltech256, to assess the model's ability to classify varied categories and scenarios. This would facilitate the classification of fruits, dishes, and everyday objects. Finally, the author also plan to utilize MobileViT plus ULSAM in other domains and tasks, like medical image analysis, remote sensing image recognition, and video understanding, to assess its versatility and effectiveness.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[4] Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2022, June). Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 11020-11028).

[5] Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X., & Jain, R. (2015). Geolocalized modeling for dish recognition. IEEE transactions on multimedia, 17(8), 1187-1199.

[6] Deng, L., Chen, J., Sun, Q., He, X., Tang, S., Ming, Z., Zhang, Y. & Chua, T. S. (2019, October). Mixed-dish recognition with contextual relation networks. In Proceedings of the 27th ACM International conference on multimedia (pp. 112-120).

[7] Wang, Y., Chen, J. J., Ngo, C. W., Chua, T. S., Zuo, W., & Ming, Z. (2019, June). Mixed dish recognition through multi-label learning. In Proceedings of the 11th Workshop on Multimedia

DOI: 10.54254/2755-2721/27/20230300

for Cooking and Eating Activities (pp. 1-8).

[8] Salvador, A., Drozdzal, M., Giró-i-Nieto, X., & Romero, A. (2019). Inverse cooking: Recipe generation from food images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10453-10462).

[9] H. Lee, H., Shu, K., Achananuparp, P., Prasetyo, P. K., Liu, Y., Lim, E. P., & Varshney, L. R. (2020, April). RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system. In Companion Proceedings of the Web Conference 2020 (pp. 181-184).

[10] Wang, H., Lin, G., Hoi, S. C., & Miao, C. (2020). Structure-aware generation network for recipe generation from images. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16 (pp. 359-374). Springer International Publishing.

[11] Ciocca, G., Napoletano, P., & Schettini, R. (2017). Learning CNN-based features for retrieval of food images. In New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers 19 (pp. 426-434). Springer International Publishing.

[12] Shimoda, W., & Yanai, K. (2017, April). Learning food image similarity for food image retrieval. In 2017 IEEE Third International Conference on Multimedia Big Data (BigMM) (pp. 165-168). IEEE.

[13] Martinel, N., Foresti, G. L., & Micheloni, C. (2018, March). Wide-slice residual networks for food recognition. In 2018 IEEE Winter Conference on applications of computer vision (WACV) (pp. 567-576). IEEE.

[14] Qiu, J., Lo, F. P. W., Sun, Y., Wang, S., & Lo, B. (2022). Mining discriminative food regions for accurate food recognition. arXiv preprint arXiv:2207.03692.

[15] Mehta, S., & Rastegari, M. (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178.

[16] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13713-13722).

[17] Cui, C., Gao, T., Wei, S., Du, Y., Guo, R., Dong, S., Lu, B., Zhou, Y., Lv, X., Liu, Q., Hu, X., Yu, D. & Ma, Y. (2021). PP-LCNet: A lightweight CPU convolutional neural network. arXiv preprint arXiv:2109.15099.

[18] Yang, S., Chen, M., Pomerleau, D., & Sukthankar, R. (2010, June). Food recognition using statistics of pairwise local features. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 2249-2256). IEEE.

[19] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[20] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. IEEE. Published online 2018. doi:10.1109/CVPR.2018.00474

[21] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V. & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324).

[22] Hou, S., Feng, Y., & Wang, Z. (2017). Vegfru: A domain-specific dataset for fine-grained visual categorization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 541-549).

[23] Jiang, S., Min, W., Liu, L., & Luo, Z. (2019). Multi-scale multi-view deep feature aggregation for food recognition. IEEE Transactions on Image Processing, 29, 265-276.

[24] Horiguchi, S., Amano, S., Ogawa, M., & Aizawa, K. (2018). Personalized classifier for food image recognition. IEEE Transactions on Multimedia, 20(10), 2836-2848.

DOI: 10.54254/2755-2721/27/20230300

[25] Min, W., Liu, L., Wang, Z., Luo, Z., Wei, X., Wei, X., & Jiang, S. (2020, October). Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 393-401).

[26] Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–mining discriminative components with random forests. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13 (pp. 446-461). Springer International Publishing.

[27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[28] Saini, R., Jha, N. K., Das, B., Mittal, S., & Mohan, C. K. (2020). Ulsam: Ultra-lightweight subspace attention module for compact convolutional neural networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1627-1636).