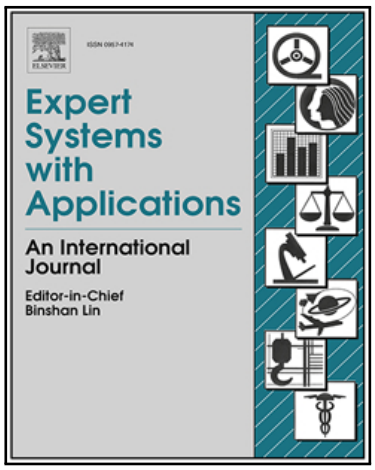


Channel Grouping Vision Transformer for Lightweight Fruit and Vegetable Recognition

Chengxu Liu, Weiqing Min, Jingru Song, Yancun Yang, Guorui Sheng, Tao Yao, Lili Wang, Shuqiang Jiang

PII: S0957-4174(25)02255-9
DOI: https://doi.org/10.1016/j.eswa.2025.128636
Reference: ESWA 128636

To appear in: *Expert Systems With Applications*

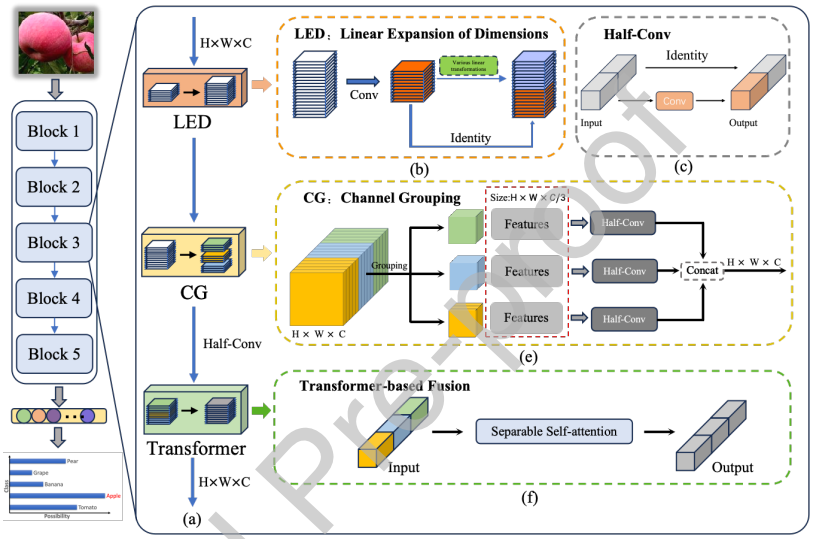
Received date: 14 August 2024
Revised date: 19 May 2025
Accepted date: 13 June 2025

Please cite this article as: Chengxu Liu, Weiqing Min, Jingru Song, Yancun Yang, Guorui Sheng, Tao Yao, Lili Wang, Shuqiang Jiang, Channel Grouping Vision Transformer for Lightweight Fruit and Vegetable Recognition, *Expert Systems With Applications* (2025), doi: https://doi.org/10.1016/j.eswa.2025.128636

Graphical Abstract

Channel Grouping Vision Transformer for Lightweight Fruit and Vegetable Recognition

Chengxu Liu, Weiqing Min, Jingru Song, Yancun Yang, Guorui Sheng, Tao Yao, Lili Wang, Shuqiang Jiang



We present a novel lightweight classification network, CGViT(Channel Grouping Vision Transformer), specifically designed for fruit and vegetable recognition. CGViT utilizes a channel grouping structure combined with the Transformer-based fusion to extract distinguishing features from fruit and vegetable images.

Highlights

Channel Grouping Vision Transformer for Lightweight Fruit and Vegetable Recognition

Chengxu Liu, Weiqing Min, Jingru Song, Yancun Yang, Guorui Sheng, Tao Yao, Lili Wang, Shuqiang Jiang

- We adopt a Channel Grouping Vision Transformer (CGViT) for lightweight fruit and vegetable recognition.
- We benchmark various lightweight deep learning networks on these four fruit datasets.
- Evaluations on four fruit and vegetable datasets demonstrate that our approach achieves state-of-the-art performance while consuming fewer resources.

Channel Grouping Vision Transformer for Lightweight Fruit and Vegetable Recognition

Chengxu Liu[a], Weiqing Min[b,c], Jingru Song[a], Yancun Yang[a], Guorui Sheng[a,*], Tao Yao[a], Lili Wang[a], Shuqiang Jiang[b,c]

[a] *School of Information and Electrical Engineering, Ludong University, 264025, Yantai, China*
[b] *The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China*
[c] *University of Chinese Academy of Sciences, 100049, Beijing, China*

Abstract

Recognizing fruit and vegetable is crucial for improving processing efficiency, automating harvesting, and facilitating dietary nutrition management. The diverse applications of fruit and vegetable recognition require deployment on end devices with limited resources, such as memory and computing power. The key challenge lies in designing lightweight recognition algorithms. However, current lightweight methods still rely on simple CNN-based networks, which fail to deeply explore and specifically analyze the unique features of fruit and vegetable images, resulting in unsatisfactory recognition performance. To address this challenge, we propose a novel lightweight recognition network termed Channel Grouping Vision Transformer (CGViT). CGViT utilizes a channel grouping mechanism and half-convolution to enhance feature extraction capability while reducing complexity. This design enables the model to capture three discriminative types of features from images. Subsequently, the Transformer is employed for feature fusion and global information extraction, ultimately creating an efficient neural network model for fruit and vegetable recognition. The proposed CGViT approach achieved

*Corresponding author
Email addresses: chengxuliu@m.ldu.edu.cn (Chengxu Liu), minweiqing@ict.ac.cn (Weiqing Min), songjingru@m.ldu.edu.cn (Jingru Song), Harryyang@ldu.edu.cn (Yancun Yang), shengguorui@ldu.edu.cn (Guorui Sheng), yaotao@ldu.edu.cn (Tao Yao), wanglili@ldu.edu.cn (Lili Wang), sqjiang@ict.ac.cn (Shuqiang Jiang)

recognition accuracies of 71.26%, 99.99%, 98.92%, and 61.33% on four fruit and vegetable datasets, respectively, outperforming state-of-the-art methods (MobileViTV2, MixNet, MobileNetV2). The maximum memory usage during training is only 6.48GB, which is merely 13.8% of that required by state-of-the-art methods(MobileViTv2). The fruit and vegetable recognition model proposed in this study offers a more profound and effective solution, providing valuable insights for future research and practical applications in this domain. The code is available at https://github.com/Axboexx/CGViT.

Keywords: Fruit Recognition, Vegetable Recognition, Lightweight, Deep Learning, Computer Vision.

1. Introduction

The recognition of fruit and vegetable plays a crucial role in automated harvesting, quality inspection and analysis, intelligent food processing, and intelligent nutrition management of diet. Traditional methods for the recognition of fruit and vegetable mainly involve manual operations, which are highly subjective and costly(Aleixos et al., 2002; Yang et al., 2023). In recent years, with rapid technological advancements, automated recognition has gradually matured. Techniques such as infrared imaging, multispectral and hyperspectral technologies have yielded promising results(Feng et al., 2019; Gaikwad & Tidke, 2022; He et al., 2024), but their reliance on expensive equipment or complex spectroscopic methods makes them difficult to widely deploy in industrial settings. The rapid development of artificial intelligence, particularly computer vision, has brought qualitative improvements to fruit and vegetable recognition, offering advantages such as high efficiency, accuracy, and low cost(Faria et al., 2021; Escamilla et al., 2024). These technologies can adapt to most operational scenarios, largely replacing manual labor and significantly driving industrial progress. For example, automated fruit and vegetable recognition has improved production line efficiency, providing substantial benefits to the food processing industry(Rehman et al., 2019; Dhanush et al., 2023). In the domain of food computing(Min et al., 2019), the recognition of fruit and vegetable is increasingly important for various tasks such as selection, classification, nutritional analysis, and dietary recommendations(Nyalala et al., 2019; Xu et al., 2019; Siddique & Srizon, 2023). Accurate and efficient recognition in automated harvesting(Xu et al., 2022; Bai et al., 2023), quality inspection and analysis(Zhu et al., 2022; Li

et al., 2023), and subsequent food processing is critical as an early-stage step. The potential of fruit and vegetable recognition is also reflected in self-checkout services in supermarketsHameed et al. (2020).

Currently, convolutional neural networks (CNNs) have demonstrated outstanding performance in vision-based fruit and vegetable recognition. Gill et al. (2022),Taner et al. (2024), andPan et al. (2024) have employed CNNs or their various derivatives for fruit and vegetable recognition, proving that deep models outperform traditional methods and offer significant advantages. The role of deep neural networks in fruit and vegetable recognition has also been explored by Nguyen et al. (2021) and Gupta & Tripathi (2024). Related works extend to tasks such as fruit and vegetable volume estimation(Ziaratban et al., 2017; Saikumar et al., 2023), disease(Gupta et al., 2024, 2025) , grade classification(Yogesh et al., 2020; Lee et al., 2020; Mputu et al., 2024), and control issues in processing(Li et al., 2021; Wang et al., 2021). However, most vision-based fruit and vegetable recognition methods still rely on existing general-purpose deep models without fully considering the unique characteristics of fruit and vegetable images. Consequently, the effectiveness of these methods is somewhat limited.

Meanwhile, improving the efficiency of training and inference for fruit and vegetable recognition models and deploying these models on resource-constrained edge devices has become increasingly important. On the one hand, with the rapid development of the Internet of Things (IoT) and the widespread use of mobile phones, fruit and vegetable recognition can only be truly productive when deployed on such edge devices, enabling it to play a role in front-line production and daily life. For instance, users can conveniently perform real-time fruit and vegetable recognition and information retrieval on mobile devices. In fields such as shopping, dining, and agriculture, smartphones or small cameras can be used to identify fruit and vegetable, providing real-time data on product quality and nutritional information, thereby enabling more informed decision making. On the other hand, efficient training on the server side requires model compression to achieve the same or even better results with fewer resources, thereby promoting energy conservation, emission reduction, and green artificial intelligence.

Current fruit and vegetable recognition algorithms primarily rely on high-performance computing devices, such as servers and desktop computers, to process complex image data and perform model inference. Although the computational power of edge devices has been steadily increasing, they still face significant limitations compared to traditional large-scale computing equip-

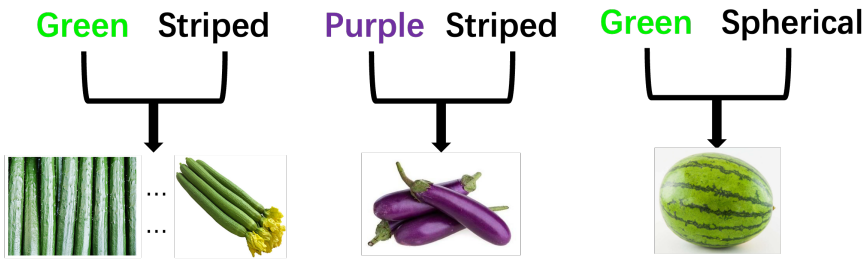


Figure 1: Differences in the first type of characteristics can make a great difference in the results.

ment in terms of storage capacity, memory resources, power consumption, and processing speed. Therefore, it is crucial to design specialized lightweight fruit and vegetable recognition algorithms that adhere to resource constraints while ensuring high accuracy in computationally limited work environments. Based on these considerations, we have thoroughly explored the unique characteristics of fruit and vegetable images and designed a lightweight deep neural network specifically for fruit and vegetable recognition.

Fruit and vegetable images exhibit three hierarchical levels of features: (1) The first level consists of relatively apparent surface attributes, such as color and shape. By analyzing these attributes, we can preliminarily infer the identity of the fruit or vegetable. Although these features are relatively straightforward and can be extracted without complex operations, they play a crucial role in fruit and vegetable recognition. As shown in Fig. 1, if the information includes a long, green object, it may correspond to vegetables like luffa or zucchini. If the color changes to purple, it could indicate foods like eggplant. If the shape becomes spherical, it might represent fruits like watermelon. This first category of features is simple and evident, guiding us toward the correct direction in reasoning. The same applies to deep models, where the first level of features plays a critical role in the model's judgment. However, relying solely on these features is insufficient for accurate decision-making. In different working environments, relying on shallow features alone may lead to erroneous judgments under low-light conditions. (2) Compared to the more obvious first category, the second level of features is less conspicuous and harder to identify. These features include complex textures, patterns, smoothness, and small granular protrusions in fruit and vegetable images. For example, although both luffa and zucchini are long, green vegeta-

bles, the surface of luffa is rougher, while the surface of zucchini is smoother. Similarly, the surface of a cucumber is relatively rough and often has small granular protrusions. Therefore, hidden features such as texture and pattern are critical to the accuracy of fruit and vegetable recognition. (3) The third level of features involves depth information that is difficult for the human eye to capture, as well as deep connections between the different manifestations of similar fruits and vegetables. Fruits and vegetables may exhibit numerous variations, such as changes in growth stages and different forms of presentation. For instance, a whole apple differs in characteristics from an apple cut into pieces or slices, but there remains an inherent connection between them. This connection may represent deeper features that are challenging for humans to comprehend. To address this limitation, we have developed a specialized network model. To our knowledge, most current lightweight fruit and vegetable recognition methods do not address these limitations. Some studies suggest improving model performance by incorporating additional high-level semantic information, but this approach involves adding additional information and computational resources(Rachmawati et al., 2022).

Based on the aforementioned characteristics of fruit and vegetable images, we used the partial convolution from FasterNet proposed byChen et al. (2023) and self-attention mechanisms as a baseline to construct our model. We introduced an improvement to the partial convolution, termed half-convolution, which enhances its feature extraction capability with a slight increase in the number of parameters. This enhancement allows for better extraction of the three types of feature while maintaining the requirements for a lightweight model. To extract these three types of features more effectively, we divided the feature map channels into different channel groups, with each group containing one third of the total channels. Then, half-convolution was applied separately to each group to improve feature extraction performance. Since there is no information exchange between channel groups during the convolution stage, global information between different groups may be lost. To address this issue, we adopted the lightweight Transformer mechanism, MobileViT, proposed byMehta & Rastegari (2021), to merge the output feature maps of different channel groups. Current research also explores the use of attention mechanisms in the domain of fruit and vegetable recognition(Min et al., 2023), which improves the model's ability to handle global information in fruit and vegetable images.

Based on the aforementioned analysis, we propose a lightweight network model for fruit and vegetable recognition, termed CGViT (Channel Group-

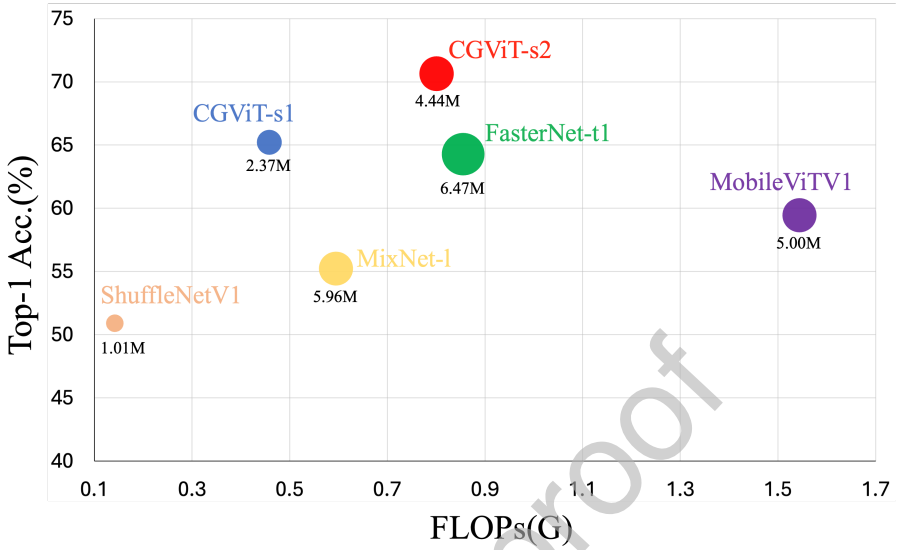


Figure 2: A simple comparison between CGViT and other lightweight networks on the Fru92 dataset(Hou et al., 2017). With roughly equivalent FLOPs and parameters, CGViT achieved the best performance. More detailed and comprehensive experimental results can be found in the experimental section of the paper.

ing Vision Transformer). This model integrates a novel lightweight convolutional module based on half-convolution as its backbone. Additionally, we incorporate a lightweight Transformer module to improve the neural network sensitivity to global features. CGViT is capable of capturing various types, levels, and granularities of features, leading to more robust and comprehensive recognition results. We conducted extensive evaluations on four benchmark fruit datasets, which differ in categories, the number of images, and shooting conditions, ranging from single fruits or vegetables per image to complex scenes captured in grocery stores or markets. The datasets cover between 50 and 131 categories, with approximately 3,500 to over 70,000 images. Our method achieved outstanding recognition performance across all datasets. Fig. 2 displays a subset of the experimental results. To further assess the effectiveness of the model, we visualized CGViT and found that it successfully extracts salient and discriminative features from fruit and vegetable images. Moreover, we conducted extensive comparative experiments with several lightweight models, including the ShuffleNet series, MobileNet series, and MobileViT series, across the four fruit and vegetable datasets.

The experimental results demonstrate the effectiveness of our approach.We summarize our contributions as follows:

1. Based on the characteristics of fruit and vegetable images, we employed the channel grouping for three types of features extraction, combined with the improved half-convolution, which improved recognition accuracy while reducing the model size.
2. By enhancing global feature extraction capabilities through Transformer-based fusion, we designed and implemented the lightweight fruit and vegetable recognition model CGViT, effectively reducing parameter and computational complexity while ensuring high accuracy.
3. Extensive experiments were conducted on four benchmark fruit and vegetable datasets to evaluate CGViT and various state-of-the-art(SOTA) lightweight deep neural networks, validating the effectiveness of CGViT.

2. Materials and Methods

2.1. Fruit and vegetable datasets

Through extensive research, many large-scale fruit and vegetable image datasets have been released. These datasets contain a wide variety of images of fruit and vegetable captured in different scenes. They are designed for specific tasks, often featuring a broader range of categories and diversity than images typically gathered in laboratory or industrial settings. The datasets include a rich array of fruit and vegetable characteristics and dynamic backgrounds, allowing for more accurate evaluation of a model's performance in real-world production environments. Consequently, this paper selects the following four fruit and vegetable datasets to comprehensively evaluate our model.The datasets is available at https://huggingface.co/datasets/Axboexx/CGViT_Datasets/tree/main

Fru92 (Hou et al., 2017) is a subset of the VegFru dataset and contains 92 fruit categories with a total of 69,614 images. Most of the images in the VegFru dataset were obtained from online searches and were then carefully filtered to ensure high quality. In Fru92(Hou et al., 2017), each fruit category has at least 200 images. We used the first 100 images from each category for training, the next 50 for validation, and the remaining images for testing. The images were sourced from various platforms, including Google and Flickr.

Fruits-360 (Muresan & Oltean, 2018) contains 73,410 images across 107 fruit types, making it the largest dataset in this category. It was developed

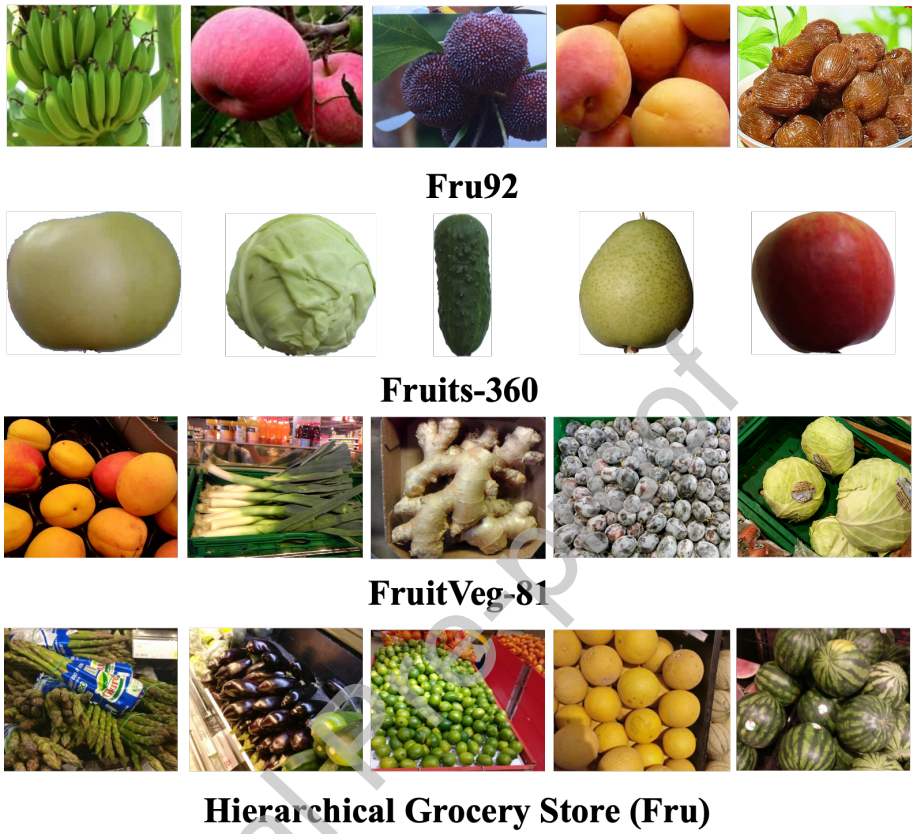


Figure 3: Images from different datasets.

in a controlled laboratory setting, using a low-speed motor that rotates the fruit against a white background to capture images from various angles. The motor rotates at 3 revolutions per minute (rpm), and a 20-second video is recorded to capture a 360-degree view of the fruit. The training set contains 54,963 images, while the test set contains 18,447 images.

FruitVeg-81 (Waltner et al., 2017) comprises 15,737 images from 81 categories of fresh fruit and vegetable, taken by five mobile phones in a SPAR grocery store. The training set consists of 9,378 images, while the test set consists of 6,359 images.

Hierarchical Grocery Store (Fru) (Klasson et al., 2019) is part of the "Hierarchical Grocery Store" dataset and comprises 3,480 images covering 50 different categories. Klasson et al. (2019) collected these images from

18 different grocery stores, including their fruit and vegetable sections. To ensure that the images were captured in natural conditions, they were taken with a 16-megapixel Android smartphone camera from different distances and angles. Notably, background noise was preserved to more closely resemble a typical grocery store environment. For each category, 60% of the images were randomly selected for training, 10% for validation, and the remaining 30% for testing.

2.2. CGViT

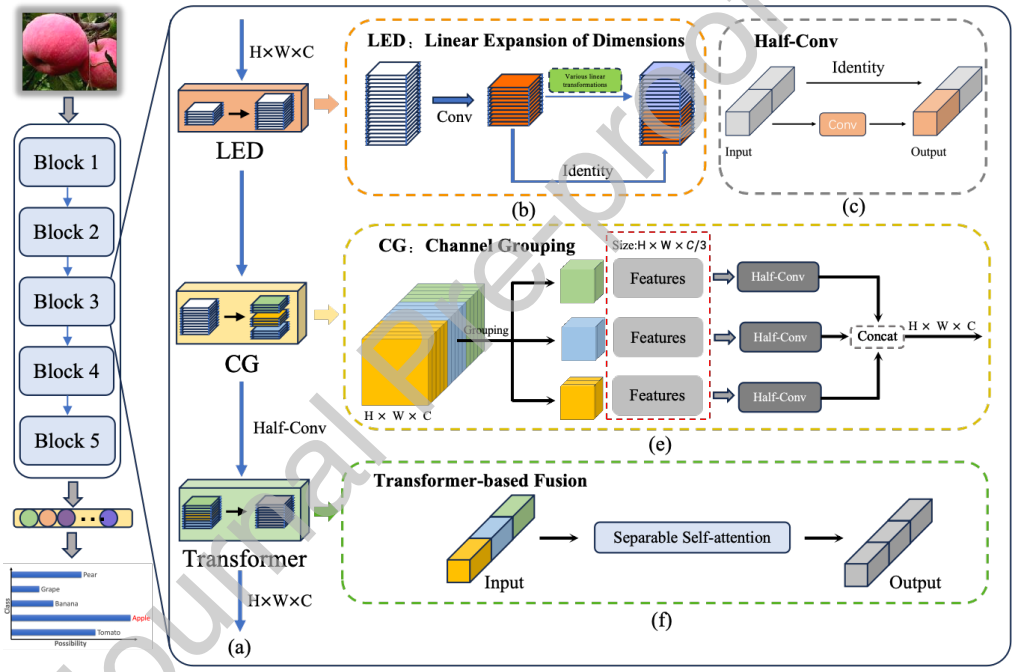


Figure 4: Overview of CGViT. (a) Specific elements of each block, (b) efficiently and cost-effectively generate additional feature maps. (c) a replacement for traditional convolution. (e) enables richer and more comprehensive feature extraction. (f) facilitates global information exchange and integrates feature maps from (e). Note that (f) is applied in Block 3-5.

Fig. 4 shows the basic framework of CGViT.The Introduction highlights that fruit and vegetable images have three distinct types of feature. The first type consists of simple features, such as color and shape, which are captured in the shallow layers of the network. The second type of features,

extracted primarily by intermediate layers, includes complex textures, surface smoothness, and fine granular protrusions. These features are crucial in distinguishing visually similar but intrinsically different categories of fruits and vegetables. For example, subtle differences in surface roughness can differentiate between similarly shaped green vegetables such as luffa and zucchini. The third type of characteristic, captured by the deep layers, includes not only depth information but also intrinsic relationships among different forms of the same fruit or vegetable, such as whole, sliced or peeled forms.

The channel grouping processes different channels of the feature map through multiple independent convolutions, making it more sensitive to discriminative features in fruit and vegetable images. This ensures that the extracted features are diverse and comprehensive. For example, in the early stages of the network, fewer convolutional operations are performed, primarily to extract simpler features such as color and shape. Channel grouping enables differential feature extraction across distinct convolution operations, ensuring diversity and completeness in feature extraction.

However, since the channel grouping isolates feature map channels and processes them independently throughout the feature extraction process, no information exchange occurs between channel groups. This limits the network's ability to gather global information, which is often critical to accurate recognition. To address this limitation, a Transformer-based fusion is incorporated to ensure that CGViT captures sufficient global information. These two mechanisms: channel grouping and Transformer-based fusion, constitute the most critical components of CGViT for feature extraction. The remaining components primarily contribute to reducing CGViT parameters and FLOPs.

For the input image, CGViT first performs a cost-effective linear operation (Han et al., 2020) to increase the dimension of the feature map. It maps different information of the feature map to different dimensions in the early stage of the network, then groups the feature map by channels, using half-convolution to extract features for each channel group. This approach enhances CGViT's generalization ability without increasing the network parameters and FLOPs. After several sets of half-convolution calculations, a Transformer-based fusion at the end fuses and exchanges feature information between different channels. The architecture of CGViT ensures the targeted extraction of different discriminant features of fruit and vegetable images while avoiding the problem of missing global features through the Transformer-based fusion. The design of CGViT actually encompasses

some modules with broad applicability, yet its core optimizations are tailored to the characteristics of fruit and vegetable images. The reasons have been described above. Its performance advantages in this task stem from a deep consideration of the characteristics of fruit and vegetable images and resource-constrained environments. Next, we describe the main components of CGViT in detail. The following is the pseudo code of CGViT. Let $Input$ and $Output$ denote the input data and the corresponding output results, respectively. The variables h and w represent the height and width of the feature map, while c denotes the number of channels and a indicates the number of channels per group. The value 3 is used illustratively in the pseudo code to indicate the number of groups. The variable $classes$ refers to the total number of classification categories. As the feature map is processed, its spatial dimensions and channel count may vary, resulting in corresponding changes in c and a. The symbol f' denotes the convolution kernel, and k represents its size. Variables m and s represent the number of channels in the new feature maps X_1 and X_2.

Algorithm 1 CGViT

Require: $Input \in \mathbb{R}^{h\times w\times 3}, a = c/3$
Ensure: $Output \in \mathbb{R}^{1\times classes}$
  $X \leftarrow \text{Resize } (Input) \in \mathbb{R}^{224\times 224\times 3}$
  for $i \leftarrow 1, 2, 3, 4, 5$ **do**
    $X_1 \leftarrow X * f', f' \in \mathbb{R}^{c\times k\times k\times m}, X_1 \in \mathbb{R}^{h'\times w'\times m}$
    $X_2 \leftarrow \Phi_{\alpha,\beta}(X_1), \forall \alpha = 1, \cdots, m, \beta = 1, \cdots, s, X_2 \in \mathbb{R}^{h'\times w'\times s}$
    $X \leftarrow Concat(X_1, X_2)$
    $A_j \leftarrow split(X, [a, a, a]), j = 1, 2, 3$
    $X \leftarrow Concat\{Conv(A_j[h, w, 0 : \frac{a}{2}]), A_j[h, w, \frac{a}{2} + 1 : a]\}, j = 1, 2, 3$
    if $i \geq 3$ **then**
      $X \leftarrow Transformer\_based\_fusion(X)$
    end if
  end for
  $Output \leftarrow classifier(X)$

2.2.1. Channel Grouping

The channel grouping mechanism is widely used in image recognition and classification. Variations in fruit and vegetable images are simpler and

rarely undergo significant changes unless they are prepared as a dish. The channel grouping mechanism employs multiple independent convolutions to process different channels of the feature map, thereby being more sensitive to the discriminative features of fruit and vegetable images. This ensures that the features extracted are more diverse and complete. For instance, in the early stages of the network, the image undergoes fewer convolutional computations, extracting only simpler features (such as color, shape, etc.). The channel grouping mechanism allows different convolutions to perform differentiated feature extraction, ensuring the diversity and completeness of image feature extraction. Therefore, this strategy is particularly suitable for recognizing fruit and vegetable images, as illustrated in Fig. 4(e).

Moreover, channel grouping can effectively reduce the number of parameters and computations required by the model, which is crucial for constructing lightweight models. The channel grouping mechanism and its variants have been extensively studied in computer vision tasks. For example, in ShuffleNet (Zhang et al., 2018), the dense point convolution (He et al., 2016) in the Bottleneck module of ResNet is replaced by channel grouping convolution. This improves performance of the network while reducing the number of parameters and computations, making it lighter. The concept of channel grouping is also applied in MixNet (Tan & Le, 2019), where convolutions of different sizes are used for different channels of the feature map. This approach captures feature extraction modes of varying resolutions, achieving excellent performance.

Specifically, the input of the channel grouping mechanism is a three-dimensional tensor:

$$X^l \in \mathbb{R}^{h \times w \times c} \tag{1}$$

where w is the width of the feature map, h is the height, c is the number of channels, and l represents a layer in the basic module. The output result is a new feature map of the same size as X^l:

$$A^l \in \mathbb{R}^{h \times w \times c} \tag{2}$$

Channel Grouping of CGViT as follows:

For the input data:$X^{input} \in \mathbb{R}^{h \times w \times c}$, it is divided into three parts evenly according to c, that is $X_i^{input} \in \mathbb{R}^{h \times w \times a}, i = 1, 2, 3; a = c/3$, Maybe i is not divisible by three, we add the undivided part to X_3^{input}. Therefore, $X^{input} = Concat\{X_1^{input}, X_2^{input}, X_3^{input}\}$. Through channel grouping, CGViT

can extract diverse features from fruit and vegetable images with fewer parameters and reduced computational complexity. The shallow layers are used to extract simple features such as color and shape, the middle layers extract more subtle features such as texture, smoothness, and patterns, and the deep layers extract abstract and complex information, including the connections between different appearances of the same fruit or vegetable, such as slicing and peeling. Although traditional CNNs and high-performance methods have demonstrated effectiveness in general fruit and vegetable recognition (Sun et al., 2021), existing research primarily aims to achieve high accuracy in environments where resources are not a constraint. However, lightweight neural networks designed for resource-constrained scenarios often require specialized design, while also maintaining the ability to extract features from fruit and vegetable images, such as subtle surface textures, internal structural connections, and variations between different forms. This limitation has prompted the development of CGViT, in which we integrate mechanisms such as channel grouping to enhance feature extraction while maintaining a lightweight architecture suitable for deployment on edge devices. Channel grouping helps CGViT extract diverse features from fruit and vegetable images. Additionally, the feature map is divided into three parts because CGViT is a lightweight network, and c is not large in each layer. If the number of groups increases, the number of channels in each group will be very small, and different groups may contain very similar features. This would increase the computational cost without improving network performance. In the experimental section, experiments with different numbers of groups will be designed to verify that $c = 3$ is better.

2.2.2. Half-convolution

Since CGViT is a lightweight network designed for mobile devices and micro-end devices, we do not use traditional convolutions that are more computationally intensive, but use half-convolutions improved based on partial convolutions (Chen et al., 2023). Research in Chen et al. (2023) points out that partial convolution(Chen et al., 2023) extracts spatial features more effectively while reducing redundant calculations and memory accesses. half-convolution that we improved on this also has this advantage. We first review the traditional Depthwise Separable Convolution(Howard, 2017). For an input of size $I \in \mathbb{R}^{h\times w\times c}$, c convolution kernels of size $W \in \mathbb{R}^{k\times k}$ are used to calculate the output $O \in \mathbb{R}^{h\times w\times c}$. Each filter performs spatial sliding on the input channel and contributes to one output channel. The Depthwise Sepa-

rable Convolution(Howard, 2017) reduces its FLOPs to $h \times w \times k^2 \times c$, which can effectively reduce FLPOs. However, the subsequent Pointwise Convolution(Howard, 2017) may cause a relatively large loss of accuracy. Therefore, in practice, the channel number c of Depthwise Convolution(Howard, 2017) increases to c' ($c' > c$), which will trigger more frequent memory accesses, may bring non-negligible delay, and reduce the overall computing speed, especially for I/O affected limited terminal equipment. According to the research of Chen et al. (2023), the number of memory accesses is upgraded to

$$h \times w \times 2c' + k^2 \times c' \approx h \times w \times 2c' \tag{3}$$

which will be higher than the number of accesses of normal convolution, that is

$$h \times w \times 2c + k^2 \times c^2 \approx h \times w \times 2c. \tag{4}$$

Therefore, to reduce the computational cost and reach the limitations of mobile devices or micro-end devices, we use half-convolution, which is an improvement on partial convolution to extract features from fruit and vegetable images. Fig. 4(c) shows the basic principle of half-convolution.

Specifically, half-convolution only applies normal convolution to half of the input channels for feature extraction. (The partial convolution used as the baseline only uses a quarter of the number of channels. Since CGViT is a lightweight network with fewer network layers, we use more channels to improve its performance.), the remaining channels remain constant. Simultaneously, the input channels of the half-convolution are equal to the output channels, resulting in the FLOPs of the half-convolution being

$$h \times w \times k^2 \times c_b^2 \tag{5}$$

where $c_b = \frac{c}{2}$, the FLOPs of a folded convolution are one-fourth that of a regular convolution. At the same time, half-convolution also has smaller memory access, that is

$$h \times w \times 2c_b + k^2 \times c_b^2 \approx h \times w \times 2c_b \tag{6}$$

compared with traditional convolution, the memory access is reduced by half.

For the second half of the input channel, we keep it unchanged. Although these channels have not been subjected to convolution operations, they still contain relevant features of fruit and vegetable images and will also play an important role in subsequent calculations, such as in the Transformer

fusion stage, and will also be involved in calculations. So, the formula for half-convolution is as follows, defining the input data and output results as $X \in \mathbb{R}^{h\times w\times c}, Y \in \mathbb{R}^{h\times w\times c}$

$$Y = Concat\{Conv\left(X_{\left[h,w,0:\frac{c}{2}\right]}\right), X_{\left[h,w,\frac{c}{2}+1:c\right]}\}. \tag{7}$$

2.2.3. Transformer-based Fusion

In Section 2.2.1, we propose channel grouping to efficiently extract three distinct types of feature from fruit and vegetable images at a lower computational cost. But it will also bring some disadvantages. Since the input channels are separated from the beginning and processed individually through half-convolution throughout the entire feature extraction process, each channel group remains isolated with no information exchange with other channel groups. This is not conducive to the network model collecting global information on fruit and vegetable images, and global information usually plays an important role in the recognition process. Therefore, we introduce the separable self-attention mechanism (Mehta & Rastegari, 2022), which is an improved method based on the MHA (multi-head attention mechanism) in MobileViTV1 (Mehta & Rastegari, 2021). Compared with the $O\left(k^2\right)$ time complexity of MHA, separable self-attention can achieve a linear time complexity of $O\left(k\right)$. The separable self-attention mechanism will have greater advantages when faced with mobile devices or micro-end devices with limited computing power. In CGViT, the lightweight fusion mechanism based on the separable self-attention mechanism(Mehta & Rastegari, 2022) is at the end of each basic CGViT module and is responsible for the task of information exchange between channels.

It is worth noting that we did not use the Transformer for the block in the early part of the network. The reason is that in the early stage of the network, the images are not fully utilized, and the number of channels of the feature map is also small, and it does not contain a lot of depth information, if the Transformer-based fusion is applied, the effect is not obvious and it will also consume more computing resources and storage resources. In Section 4.1, ablation experimental results will be presented to support this view.

2.2.4. Linear Expansion of Dimensions

The deep network must continuously increase the dimensionality of the feature map during calculation to extract more depth information. Using normal convolutions to increase the dimensionality of feature maps results

in more parameters and an increased computational load, which is inconsistent with our goal of building a lighter fruit and vegetable recognition model. Therefore, we use the method proposed in GhosNet (Han et al., 2020) to linearly increase the dimensionality of the feature map, as shown in Fig. 4(b). We add the linear dimensionality-raising operator to CGViT to generate more feature maps with fewer parameters and FLOPs. Based on the findings of Han et al. (2020), conventional neural networks often generate numerous redundant feature maps. The authors analyzed ResNet50 (He et al., 2016) and observed that many of these feature maps were strikingly similar. By employing a straightforward linear transformation method, these redundant feature maps can be derived with significantly fewer parameters and computational effort, resulting in a more efficient process. Specifically, according to the research by Han et al. (2020), given data $X \in R^{c\times h\times w}$, where c is the number of channels, h and w are the height and width of the input data. If the convolution operation to generate n feature maps can be expressed as:

$$Y = X * f + b \tag{8}$$

where $*$ represents the convolution operation, b represents the deviation term, $Y \in \mathbb{R}^{h'\times w'\times n}$ is the output data of this layer, with n channels, $f \in \mathbb{R}^{c\times k\times k\times n}$ is the convolution kernel of this layer, and k is the convolution kernel size. Therefore, the number of FLOPs required for this layer is $n \times h' \times c \times k^2$. The author emphasizes that it is unnecessary to utilize many FLOPs and parameters to create redundant feature maps. Many feature maps can be derived through simple linear transformations using a limited number of original feature maps. The process starts with generating a set of m original feature maps through standard convolution:

$$Y' = X * f' \tag{9}$$

$f' \in \mathbb{R}^{c\times k\times k\times m}$,$m < n$. To further obtain the required n feature maps, the author proposes to use a simple linear operation on each original feature map in Y' to generate new feature maps:

$$y_{i,j} = \Phi_{i,j}\left(y'_i\right), \forall i = 1, \ldots m,\ \ j = 1, \ldots, s \tag{10}$$

where y'_i is the $i-th$ original feature map of Y', and $\Phi_{i,j}$ is the $j-th$ linear operation used to generate the $j-th$ new feature map $y^{i,j}$. Therefore, $m \times s$ new feature maps can be obtained, and finally Y' and $y^{i,j}$ are spliced to obtain the output of the Linear Expansion of Dimensions.

3. Model training and evaluation

3.1. Model training

The operating system version is Ubuntu 20.04 LTS. We use the Pytorch 1.12.0(Paszke et al., 2017), Python 3.8 to construct our model, which is then trained on a NVIDIA A800 GPU (80GB), Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz, 64GB RAM, 1TB SSD. During the experimental data preprocessing, the input image size was resized to 256 × 256 and then randomly cropped to 224 × 224. Additionally, random horizontal flipping was applied to augment the images during model training. During testing, the input images were resized to 256 × 256 and then center-cropped to 224 × 224. All networks are trained directly on four datasets, optimized using stochastic gradient descent with a batch size of 128, a momentum of 0.9, and a weight decay of 10^{-4}. The initial learning rate is set to 10^{-2}, and the learning rate is adjusted using CosineAnnealingLR.

3.2. Model evaluation

The software and hardware settings in the evaluation are the same as Section 3.1. Model training. In the evaluation stage, the center of the 256 × 256 pixel image is cropped to 224 × 224 pixels and normalized. We use both Top-1 accuracy (Top-1 Acc.) and Top-5 accuracy (Top-5 Acc.) as evaluation metrics. Top-1 accuracy represents the percentage of predictions in which the top guess of the model matches the category of ground truth. Similarly, Top-5 accuracy represents the percentage of predictions where the correct category is among the top five guesses made by the model.

4. Results and discussion

4.1. Ablation study

We first evaluate the effectiveness of each component within CGViT. Our baseline network is FasterNet. We gradually add different components, namely the Channel Grouping and the Fusion mechanism, to verify that each method plays a positive role in fruit and vegetable image recognition.

The experimental results are presented in Table 1. The results indicate that the performance of the model improves progressively with the addition of different methods. The channel grouping mechanism enables the model to

Table 1: CGViT ablation experiment results (%).

Method	Fru92		Fruits-360		FruitVeg-81		Hierarchical Grocery Store (Fru)	
	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
Base	55.37	83.02	99.89	99.90	95.37	99.70	56.14	87.96
Base+Channel Grouping	61.50	84.61	99.90	99.99	96.46	99.79	57.69	88.50
Base+Channel Grouping+ Transformer	**64.83**	**88.07**	**99.99**	**100.00**	**98.54**	**99.90**	**58.33**	**89.58**

extract features from different categories in a more targeted manner, thereby enhancing the robustness of method. Additionally, the fusion mechanism addresses the issues introduced by the channel grouping mechanism. This mechanism facilitates the exchange of information between different channel groups, further improving the ability of the method to extract global information.

We conducted additional experiments to evaluate the impact of the number of channel groups on the performance of CGViT. Specifically, the number of channel groups was set to 2 and 4, and the experiments on the Fru92 dataset(Hou et al., 2017). The results are as follows:

Table 2: The experimental results of CGViT on Fru92 with 2 groups.

Method	Groups	Param(M)	FLOPs(G)	Top-1(%)	Top-5(%)
CGViT-s1	2	2.54	0.54	65.61	87.07
CGViT-s2		4.43	0.79	68.33	89.83

Table 3: The experimental results of CGViT on Fru92 with 4 groups.

Method	Groups	Param(M)	FLOPs(G)	Top-1(%)	Top-5(%)
CGViT-s1	4	2.29	0.46	65.91	88.12
CGViT-s2		4.36	0.74	67.37	89.68

According to Table 2 and Table 3, the impact on CGViT-s2 is greater. When the number of groups decreases or increases, the results are reduced. When the number of groups is 3, the Top-1 accuracy of CGViT-s2 on the Fru92 dataset(Hou et al., 2017) is 71.26%. Therefore, from the experimental results, the number of groups 3 is more suitable for CGViT.

Table 4: Experimental results of adding Transformer-based fusion.

	Param(M)	FLOPs(G)	Top-1(%)	Top-5(%)
CGViT-s1-T	2.46	0.92	68.11	89.25
CGViT-s2-T	7.71	8.58	72.54	91.25

The results in Table 4 show that adding more Transformer-based fusion in the early stage of CGViT does not significantly improve the performance, but consumes more resources.

4.2. Comparison with state-of-the-art

We list the state-of-the-art lightweight networks on four fruit and vegetable datasets, including MobileNetV2 (Sandler et al., 2018), GhostNet (Han et al., 2020), MobileViTv2(Mehta & Rastegari, 2022) and FasterNet (Chen et al., 2023). The experimental results are listed in Table 5. The experimental results indicate that the performance of CGViT surpasses other methods (s1, s2 represent CGViT with different FLOPs and parameters). Our method achieves the best performance in both Top-1 and Top-5 accuracy.

Our method demonstrates superior performance compared to all baseline and other lightweight models on simpler fruit datasets such as Fruits-360 and FruitVeg-81(Waltner et al., 2017). These results confirm the effectiveness of CGViT in fruit and vegetable image recognition. Furthermore, CGViT exhibits robust performance on more complex datasets, indicating its potential to excel in real-world scenarios involving intricate fruit and vegetable images. MobileViTV2(Mehta & Rastegari, 2022) also stands out on the Fru92(Hou et al., 2017) dataset, likely due to its higher parameter count and

Table 5: Comparison of performance on four datasets (%).

Method	Input Size	Fru92		Fruits-360		FruitVeg-81		Hierarchical Grocery Store (Fru)	
		Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
ShuffleNetV1 (Zhang et al., 2018)	224 × 224	51.76	81.26	99.95	99.99	98.71	99.92	53.22	84.86
ShuffleNetV2-2.0 (Ma et al., 2018)	224 × 224	46.78	78.39	99.81	99.98	98.83	99.94	61.03	89.08
MobileNetV2 (Sandler et al., 2018)	224 × 224	44.26	76.76	99.97	99.98	95.72	99.85	46.88	88.56
GhostNet (Han et al., 2020)	224 × 224	50.13	77.15	99.92	99.98	97.40	99.85	44.48	83.57
MixNet-l (Tan & Le, 2019)	224 × 224	55.46	81.02	99.61	99.96	98.50	99.92	51.41	87.56
MixNet-s (Tan & Le, 2019)	224 × 224	51.83	77.39	99.49	99.95	98.08	99.83	46.77	85.04
MobileViTV1 (Mehta & Rastegari, 2021)	256 × 256	58.38	84.01	99.94	99.99	98.08	99.89	57.25	86.10
MobileViTv2-2.0 (Mehta & Rastegari, 2022)	256 × 256	68.48	88.22	99.98	**100.00**	**98.92**	99.92	57.81	88.76
FasterNet-t0 (Chen et al., 2023)	224 × 224	63.17	88.11	99.98	**100.00**	98.82	99.89	56.80	89.50
FasterNet-t1 (Chen et al., 2023)	224 × 224	64.80	88.30	99.98	**100.00**	98.88	99.83	58.92	88.57
CGViT-s1	224 × 224	64.83	88.07	**99.99**	**100.00**	98.54	99.90	58.33	**89.58**
CGViT-s2	224 × 224	**71.26**	**90.43**	**99.99**	**100.00**	**98.92**	**99.97**	**61.33**	87.56

FLOPs. Similarly, ShuffleNetV2's(Ma et al., 2018) strong performance on the FruitVeg-81 dataset(Waltner et al., 2017) can be attributed to its high FLOPs. Despite its low parameter count and FLOPs, our method consistently delivers strong results across the various datasets. Further experiments will provide more precise comparisons regarding parameter counts.

As illustrated in Fig. 5, CGViT demonstrates superior convergence performance during training, achieving significantly lower final loss values of 0.75 and 0.81 compared to other lightweight architectures. Furthermore, when analyzed in conjunction with the data presented in Table 6, it becomes evident that CGViT exhibits a substantial reduction in model parameters, thereby reinforcing its advantage in terms of model compactness. The synergistic combination of enhanced convergence characteristics and reduced parameter count underscores the effectiveness of CGViT as an efficient architecture for fruit and vegetable recognition tasks. These empirical results collectively indicate that CGViT successfully establishes an optimal balance between recognition accuracy and lightweight design, positioning it as an ideal deployment solution for resource-constrained environments. The architectural superiority of CGViT is particularly manifested in its ability to maintain high performance while achieving remarkable efficiency, making it a compelling choice for practical implementation scenarios where computational resources are limited.

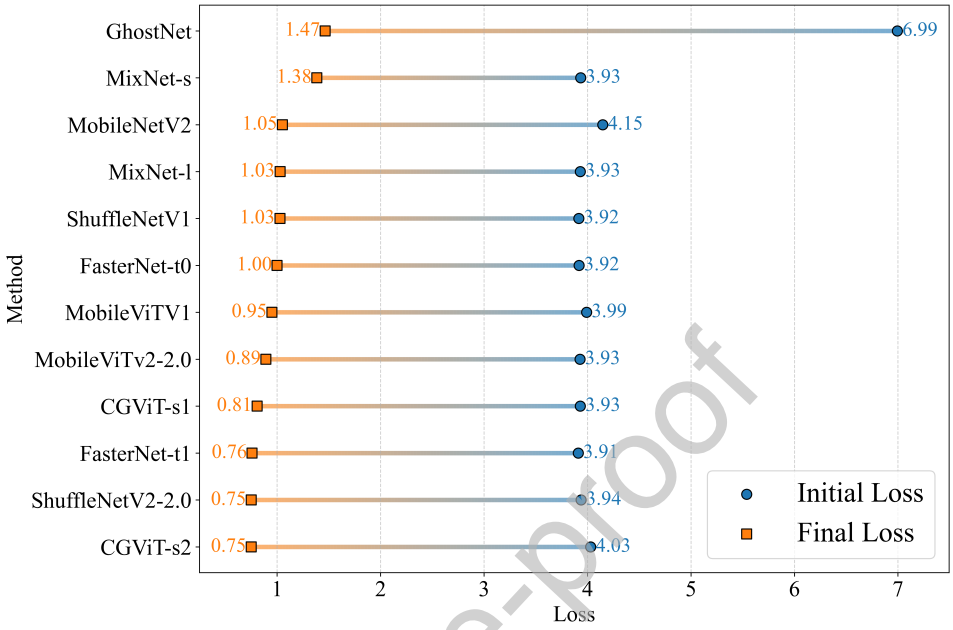


Figure 5: Comparison of initial and final training losses for lightweight networks

We calculated the parameter count and FLOPs for CGViT and compared them with those of other networks listed in Table 6. To ensure consistency, we standardized the settings across all experiments on the Fru92 dataset(Hou et al., 2017), using a batch size of 256, 8 workers, and a single A800 (80GB) GPU. The results in Table 6 show that our method has fewer parameters and FLOPs compared to several other lightweight networks. Moreover, our method requires less time to complete each epoch. Although GhostNet(Han et al., 2020) also has low time consumption per epoch, its performance across various datasets is subpar. This suggests that GhostNet(Han et al., 2020) sacrifices a significant amount of performance for speed. Our method employs a linear dimension expansion module similar to GhostNet(Han et al., 2020) but includes additional algorithms designed to enhance accuracy in fruit and vegetable images recognition. Consequently, our method outperforms lightweight networks with similar parameter counts or FLOPs. The higher accuracy achieved within the same training cycles can significantly reduce training costs.

Table 6: Comparison of parameters, Flops, Epoch, and accuracy in Fru92 with different lightweight networks.

Method	Param(M)	FLOPs(G)	Time for each epoch(s)	Memory(GB)	Top-1 Acc. (%)
ShuffleNetV1 (Zhang et al., 2018)	1.01	0.14	28	4.60	51.76
ShuffleNetV2-2.0 (Ma et al., 2018)	5.54	9.03	60	70.90	46.78
MobileNetV2 (Sandler et al., 2018)	2.34	0.31	11	11.50	44.26
GhostNet (Han et al., 2020)	5.18	0.15	7	6.45	50.13
MixNet-l (Tan & Le, 2019)	5.96	0.59	13	22.09	55.46
MixNet-s (Tan & Le, 2019)	2.73	0.25	7	12.62	51.83
MobileViTV1 (Mehta & Rastegari, 2021)	5.00	1.55	12	38.90	58.38
MobileViTv2-2.0 (Mehta & Rastegari, 2022)	17.52	5.63	18	46.96	**68.48**
FasterNet-t0 (Chen et al., 2023)	2.74	0.34	4	3.78	63.17
FasterNet-t1 (Chen et al., 2023)	6.47	0.85	3	7.10	64.80
CGViT-s1	2.37	0.48	5	5.74	**64.83**
CGViT-s2	4.44	0.79	7	6.48	**71.26**

Additionally, we tested the memory usage of various models during training to determine their impact on GPU memory. The "Memory" column in Table 6 provides the exact figures. We observed that our method uses a maximum of only 6.48GB of memory, while most other models require significantly more GPU memory. For instance, a larger lightweight model like MobileViTV2(Mehta & Rastegari, 2022) requires almost 47GB. Among the models with comparable GPU memory usage, our model has a significant accuracy advantage. It is worth noting that ShuffleNetV2(Ma et al., 2018) also demands a considerable amount of memory, which we attribute to its high FLOPs. This further confirms that our method stands out for its efficiency in memory usage.

Finally, we investigated the generalization capability of CGViT on different datasets. Given that the Fru92 dataset(Hou et al., 2017) contains a wide variety of fruit and vegetable images with different appearances, including intact, sliced and peeled images, and has a higher proportion of test images, CGViT can benefit significantly from transferring to other datasets. We present the

Table 7: Performance of visual representations transferred from Fru92 to the other three datasets (%).

Dataset	Fruits-360		FruitVeg-81		Hierarchical Grocery Store (Fru)	
	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.	Top-1 Acc.	Top-5 Acc.
CGViT-s1	99.99	100.00	98.54	99.90	58.33	89.58
CGViT-s2	99.99	100.00	98.92	99.97	61.33	87.56
CGViT-s1-Fine tuned	99.99	100.00	99.26	99.85	61.32	84.27
CGViT-s2-Fine tuned	99.99	100.00	99.49	99.87	67.54	93.60

experimental results of CGViT on three datasets in Table 7. In these experiments, CGViT denotes direct training on the respective dataset, while CGViT + Fine-tuned indicates that CGViT was pre-trained on Fru92(Hou et al., 2017) and then fine-tuned on the other three datasets for evaluation. As shown in Table 7, the fine-tuned CGViT achieved better performance when transferred to the three datasets, outperforming the results obtained from direct training on the target datasets. Notably, the fine-tuned model demonstrated more significant improvements on the FruitVeg-81(Waltner et al., 2017) and Hierarchical Grocery Store(Fru)(Klasson et al., 2019) datasets, with approximately a 1% increase on FruitVeg-81(Waltner et al., 2017) and a 3% and 6% increase on Hierarchical Grocery Store(Fru)(Klasson et al., 2019) for CGViT-s1 and CGViT-s2, respectively. However, the fine-tuned CGViT maintained a performance of 99.99% on the Fruits-360 dataset. This is because the Fruits-360 dataset(Muresan & Oltean, 2018) contains simpler images with plain white backgrounds and intact fruits without any slicing or segmentation. Thus, CGViT shows no significant improvement on this dataset.

4.3. Qualitative evaluation

We used Grad-CAM(Selvaraju et al., 2017) to perform a visual analysis of CGViT. First, we chose visually similar fruit images to check whether our model could extract the most critical features. Then, we visualized a set of fruit images with high variability to assess the model's ability to identify visual patterns.

First, we conducted experiments on multiple groups of visually similar images from the Fru92 dataset(Hou et al., 2017), as illustrated in Fig. 6. Our model can identify distinctive features in fruit images effectively. For

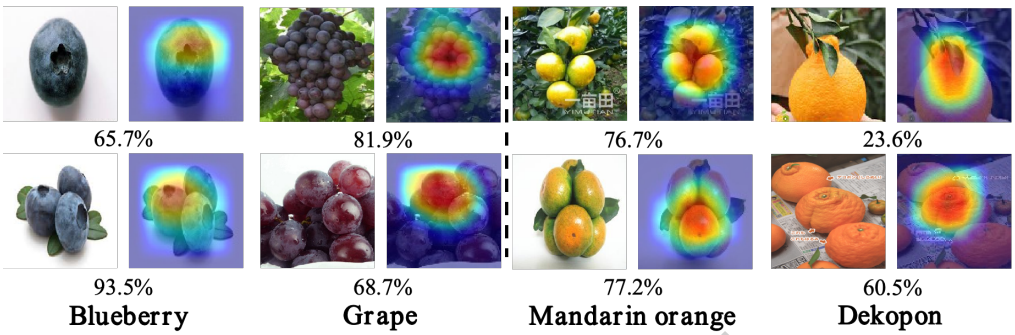


Figure 6: Attention visualization for similar images from Fru92.The probability in the figure represents the probability of the image being classified as the specified category.

example, although blueberries and grapes appear similar, the tops of blueberries are not as smooth and have unique characteristics that distinguish them from grapes. Similarly, Mandarin oranges share a similar color and shape with Dekopon; however, Dekopon has a small protrusion and a rougher surface, setting it apart from mandarin oranges.

In Fig. 7, we present the visualization of various categories of fruits. CGViT demonstrates its ability to extract visual patterns of figs, regardless of their setting, whether in a bowl or on a tree. For kiwifruit, CGViT can identify different visual patterns, whether it is the outer pattern when whole or the inner pattern when sliced open. CGViT also performs exceptionally well on fruits that undergo significant shape changes after being sliced. For instance, starfruit, when sliced, has a star-like shape that contrasts with its whole form; however, CGViT can accurately recognize it. Our method shows stronger activation in target object regions, indicating that CGViT has an enhanced capability to focus on more distinctive areas of fruit and vegetable images compared to other approaches.

Lastly, we visualized different stages of CGViT, revealing attention regions via backward propagation and showing only the positive gradient for a specific category. Fig. 8 illustrates the following: (1) For each input image, when the background is simple and clearly distinguishable from the target, CGViT quickly differentiates between the background and the fruit, focusing on extracting fruit-specific features in the later stages for classification. (2) In cases with complex backgrounds where the background color closely resembles the target color, CGViT shifts its focus to the target's shape and contour, allowing the model to concentrate on the target. Our method demonstrates

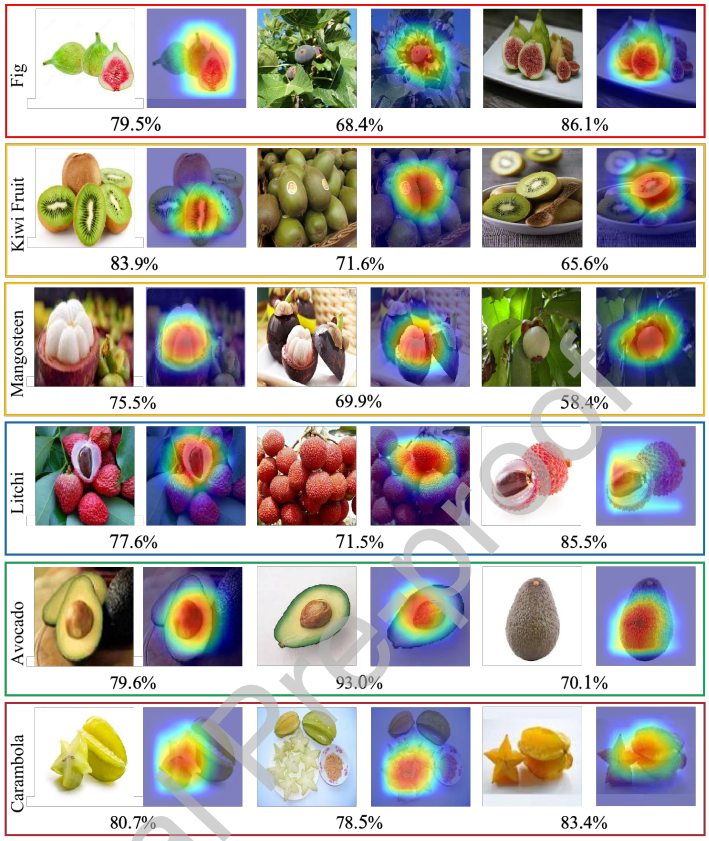


Figure 7: Results of the CGViT visualization experiment for different visual patterns (Images from Fru92). The probability in the figure represents the probability of the image being classified as the specified category.

a strong ability to understand various fruits, as the model can learn rich feature information and aggregate it through the Transformer-based fusion.

(3) The Transformer-based fusion module in CGViT gathers more global information, helping the model determine where attention should be focused. In contrast, the Channel Grouping module focuses on extracting more distinctive features, providing the model with a robust foundation for fruit and vegetable classification.

In each panel of Fig. 8, the left side presents images of four specific categories of fruits or vegetables, while the right side displays the most commonly misclassified categories for each. From the figure, it can be observed that even within the same type of fruit or vegetable, external factors such as

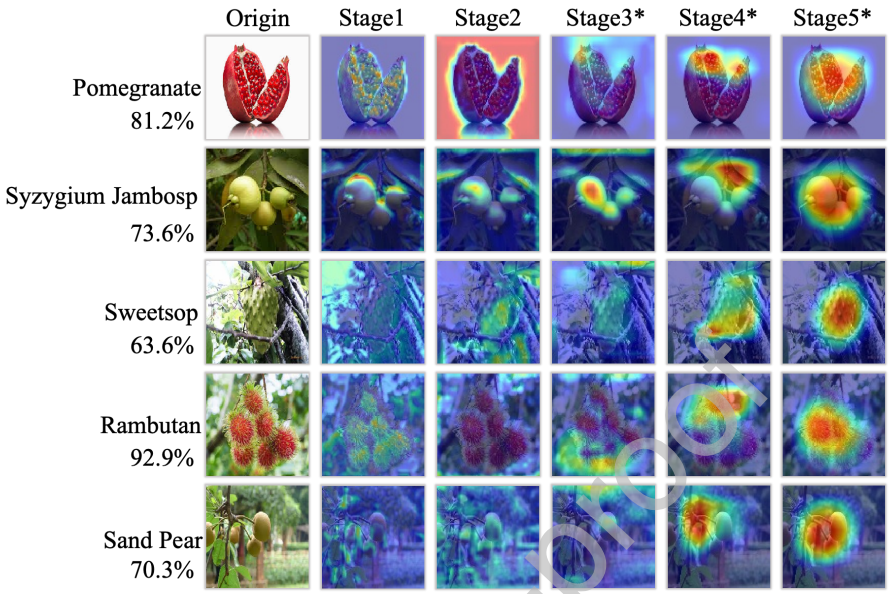


Figure 8: Visualization experiments at different stages of CGViT. The probability in the figure represents the probability of the image being classified as the specified category. (The * in the upper right corner of "Stage" indicates that Transformer-based Fusion is applied in this stage)

camera angle, lighting conditions, and ripeness can cause their appearance to resemble other categories. (1) Under certain lighting conditions, apples and gandaria exhibit similar gloss and color. Causes apple to be identified as gandaria or cherry. (2) As shown in the lower-left section of the figure, the fruits are all orange-yellow in color. When images are captured at close angles or when size-related features are diminished, the model may struggle to distinguish between them. (3) Due to variations in shooting height, many fine details of the characteristics of fruits and vegetables may not be apparent, making it difficult for the neural network to differentiate them, as shown on the right side of the figure.

Although subtle distinguishable features, such as shape details or texture variations, are still present in these images, the weight of these critical features may be diminished when the overall similarity of the feature is high, leading to classification errors. This phenomenon highlights the limitations of the model when handling complex visual features, particularly in fine-grained classification tasks.

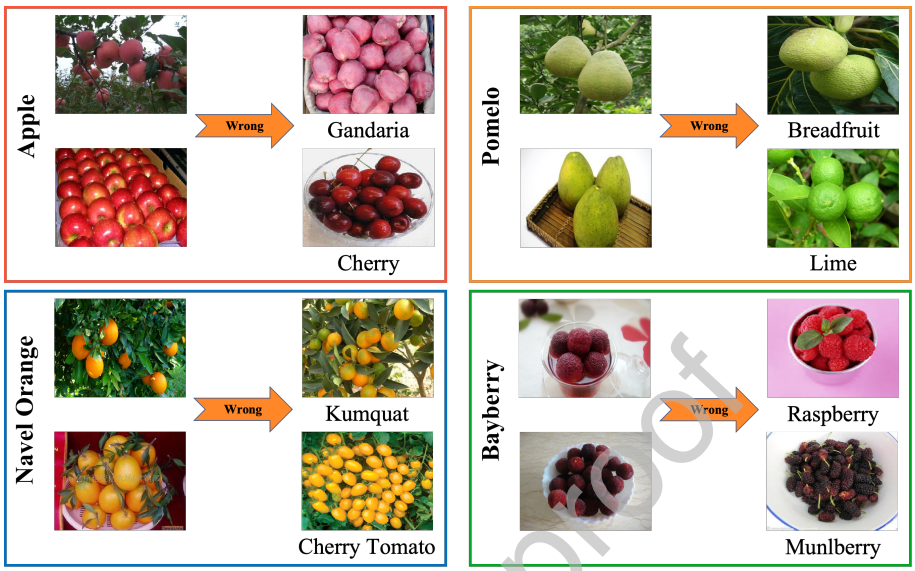


Figure 9: Some images that CGViT recognizes incorrectly.

4.4. Discussions

In our study, CGViT emerges as a novel lightweight neural network, making groundbreaking progress in the field of fruit and vegetable recognition. Through an extensive evaluation of four datasets, CGViT demonstrated outstanding performance. This innovation not only advances research in lightweight deep learning networks, but also achieves higher efficiency and accuracy on resource-constrained end devices compared to existing studies. This aligns with the findings of Yang et al. (2023) and Escamilla et al. (2024), who emphasised the critical role of real-time recognition technologies in improving automation and quality inspection efficiency within the food industry. Our work not only contributes to the theoretical advancement of this field, but also has significant practical value for real-world applications.

The success of CGViT lies in its innovative integration of channel grouping mechanisms with Transformer architectures in the domain of fruit and vegetable recognition. This enables CGViT to effectively extract diverse features from fruit and vegetable images while maintaining its lightweight nature. Not only does this approach enhance the recognition capability of CGViT, but it also reduces the number of parameters and computational complexity. Our findings suggest that CGViT has significant potential for practical applications such as automated harvesting, quality inspection, and

food processing, particularly in resource-constrained environments.

Future research could explore multi-scale feature extraction and fusion, insights from large models, and the integration of multimodal information to further enhance the performance of fruit and vegetable images recognition. Multi-scale feature extraction and fusion can capture subtle yet critical texture and shape variations in fruit and vegetable images, aiding in the differentiation of the same fruit type at different stages of maturity. The large models also offer inspiration for the recognition of lightweight fruit and vegetable images, such as transferring knowledge from large models through distillation techniques, incorporating efficient attention mechanisms, or employing multimedia learning strategies that combine image features with semantic information to improve accuracy and robustness. Moreover, integrating multimodal information (e.g. semantic data) can address the limitations of visual features, providing promising solutions to implement fruit and vegetable recognition models on resource-constrained devices.

5. Conclusions

In this work, We present a novel lightweight network model for fruit and vegetable recognition, termed CGViT. By leveraging the distinctive characteristics of fruit and vegetable images, CGViT incorporates a channel grouping mechanism to reduce parameter and computational complexity of method while extracting multi-level features. Additionally, the model employs half-convolution to further streamline its architecture while preserving robust feature extraction capabilities. The Transformer-based fusion module enhances the capacity of CGViT to aggregate global information. Extensive experimental results demonstrate that CGViT surpasses all other state-of-the-art lightweight benchmark models across four widely used fruit and vegetable datasets. As an efficient and effective lightweight solution, CGViT exhibits high accuracy and efficiency in visual-based fruit and vegetable recognition, positioning it as a viable option for deployment on resource-constrained edge devices and facilitating more efficient training in server-side environments.

ORCID Information

Chengxu Liu Orcid: 0009-0003-2874-436X e-mail: chengxuliu@m.ldu.edu.cn Weiqing Min: orcid: 0000-0001-6668-9208 e-mail: minweiqing@ict.ac.cn Jingru Song: orcid: 0009-0006-6637-2212 e-mail: songjingru@m.ldu.edu.cn Yancun Yang: orcid: 0000-0003-0785-6007 e-mail: Harryyang@ldu.edu.cn Guorui

Sheng(Corresponding author) Orcid: 0000-0001-6790-0239 e-mail: shengguorui@ldu.eud.cn Tao Yao: orcid: 0000-0003-2660-1050 e-mail: yaotao@ldu.edu.cn Lili Wang: orcid: 0000-0002-1025-3955 e-mail: wanglili@ldu.edu.cn Shuqiang Jiang orcid: 0000-0002-1596-4326 e-mail: sqjiang@ict.ac.cn

CRediT author statement

Chengxu Liu: Methodology, Software, Validation, Writing - Original Draft Weiqing Min: Conceptualization, Writing - Review & Editing Jingru Song: Writing - Review & Editing Yancun Yang: Software, Data Curation Guorui Sheng: Writing - Review & Editing Tao Yao: Resources Lili Wang: Project administration Shuqiang Jiang: Supervision

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aleixos, N., Blasco, J., Navarron, F., & Moltó, E. (2002). Multispectral inspection of citrus in real-time using machine vision and digital signal processors. *Computers and electronics in agriculture*, *33*, 121–137. https://www.sciencedirect.com/science/article/pii/S0168169902000029.

Bai, Y., Mao, S., Zhou, J., & Zhang, B. (2023). Clustered tomato detection and picking point location using machine learning-aided image analysis for automatic robotic harvesting. *Precision Agriculture*, *24*, 727–743. https://link.springer.com/article/10.1007/s11119-022-09972-6.

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., & Chan, S.-H. G. (2023). Run, don't walk: chasing higher flops for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12021–12031). http://openaccess.thecvf.com/content/CVPR2023/html/Chen_Run_Dont_Walk_Chasing_Higher_FLOPS_for_Faster_Neural_Networks_CVPR_2023_paper.html.

Dhanush, G., Khatri, N., Kumar, S., & Shukla, P. K. (2023). A comprehensive review of machine vision systems and artificial intelligence algorithms for the detection and harvesting of agricultural produce. *Scientific African*, (p. e01798). https://www.sciencedirect.com/science/article/pii/S2468227623002545.

Escamilla, L. D. V., Gómez-Espinosa, A., Cabello, J. A. E., & Cantoral-Ceballos, J. A. (2024). Maturity recognition and fruit counting for sweet peppers in greenhouses using deep learning neural networks. *Agriculture*, *14*, 331. https://search.proquest.com/openview/7518f19f99ceddf99d7c2bc6fd662645/1?pq-origsite=gscholar&cbl=2032441.

Faria, F. A., dos Santos, J. A., Rocha, A., & Torres, R. d. S. (2021). Automatic classifier fusion for produce recognition. In *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images* (pp. 252–259). IEEE. https://ieeexplore.ieee.org/abstract/document/6382764/.

Feng, J., Zeng, L., & He, L. (2019). Apple fruit recognition algorithm based on multi-spectral dynamic image analysis. *Sensors*, *19*, 949. https://www.mdpi.com/1424-8220/19/4/949.

Gaikwad, S., & Tidke, S. (2022). Multi-spectral imaging for fruits and vegetables. *International Journal of Advanced Computer Science and Applications*, *13*, 743–760. https://www.academia.edu/download/83941681/Paper_87-Multi_Spectral_Imaging_for_Fruits_and_Vegetables.pdf.

Gill, H. S., Khalaf, O. I., Alotaibi, Y., Alghamdi, S., & Alassery, F. (2022). Multi-model cnn-rnn-lstm based fruit recognition and classification. *Intelligent Automation & Soft Computing*, *33*. https://cdn.techscience.cn/ueditor/files/iasc/TSP_IASC-33-1/TSP_IASC_22589/TSP_IASC_22589.pdf.

Gupta, S., & Tripathi, A. K. (2024). Fruit and vegetable disease detection and classification: Recent trends, challenges, and future opportunities. *Engineering Applications of Artificial Intelligence*, *133*, 108260. https://www.sciencedirect.com/science/article/pii/S0952197624004184.

Gupta, S., Tripathi, A. K., & Lewis, N. (2025). Pre-trained noise based unsupervised gan for fruit disease classification in imbalanced datasets. *Pattern Analysis and Applications*, *28*, 39.

Gupta, S., Tripathi, A. K., & Pandey, A. C. (2024). Potcapsnet: an explainable pyramid dilated capsule network for visualization of blight diseases. *Neural Computing and Applications*, *36*, 23251–23274.

Hameed, K., Chai, D., & Rassau, A. (2020). A sample weight and adaboost cnn-based coarse to fine classification of fruit and vegetables at a supermarket self-checkout. *Applied Sciences*, *10*, 8667. https://www.mdpi.com/2076-3417/10/23/8667.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1580–1589). http://openaccess.thecvf.com/content_CVPR_2020/html/Han_GhostNet_More_Features_From_Cheap_Operations_CVPR_2020_paper.html.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

He, M., Li, C., Cai, Z., Qi, H., Zhou, L., & Zhang, C. (2024). Leafy vegetable freshness identification using hyperspectral imaging with deep learning approaches. *Infrared Physics & Technology*, *138*, 105216. https://www.sciencedirect.com/science/article/pii/S1350449524001002.

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). IEEE.

Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, . https://arxiv.org/abs/1704.04861.

Klasson, M., Zhang, C., & Kjellström, H. (2019). A hierarchical grocery store image dataset with visual and semantic labels. In *2019 IEEE winter conference on applications of computer vision (WACV)* (pp. 491–500). IEEE. https://ieeexplore.ieee.org/abstract/document/8658240/.

Lee, J., Nazki, H., Baek, J., Hong, Y., & Lee, M. (2020). Artificial intelligence approach for tomato detection and mass estimation in precision agriculture. *Sustainability*, *12*, 9138. https://www.mdpi.com/2071-1050/12/21/9138.

Li, K., Wang, J., Jalil, H., & Wang, H. (2023). A fast and lightweight detection algorithm for passion fruit pests based on improved yolov5. *Computers and Electronics in Agriculture*, *204*, 107534. https://www.sciencedirect.com/science/article/pii/S0168169922008420.

Li, X., Liu, Y., Gao, Z., Xie, Y., & Wang, H. (2021). Computer vision online measurement of shiitake mushroom (lentinus edodes) surface wrinkling and shrinkage during hot air drying with humidity control. *Journal of Food Engineering*, *292*, 110253. https://www.sciencedirect.com/science/article/pii/S0260877420303447.

Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 116–131). http://openaccess.thecvf.com/content_ECCV_2018/html/Ningning_Light-weight_CNN_Architecture_ECCV_2018_paper.html.

Mehta, S., & Rastegari, M. (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, . https://arxiv.org/abs/2110.02178.

Mehta, S., & Rastegari, M. (2022). Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*, . https://arxiv.org/abs/2206.02680.

Min, W., Jiang, S., Liu, L., Rui, Y., & Jain, R. (2019). A survey on food computing. *ACM Computing Surveys (CSUR)*, *52*, 1–36. https://dl.acm.org/doi/abs/10.1145/3329168.

Min, W., Wang, Z., Yang, J., Liu, C., & Jiang, S. (2023). Vision-based fruit recognition via multi-scale attention cnn. *Computers and Electronics in*

Agriculture, *210*, 107911. https://www.sciencedirect.com/science/article/pii/S0168169923002995.

Mputu, H. S., Abdel-Mawgood, A., Shimada, A., & Sayed, M. S. (2024). Tomato quality classification based on transfer learning feature extraction and machine learning algorithm classifiers. *IEEE Access*, . https://ieeexplore.ieee.org/abstract/document/10388315/.

Muresan, H., & Oltean, M. (2018). Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, *10*, 26–42. https://intapi.sciendo.com/pdf/10.2478/ausi-2018-0002.

Nguyen, H. H. C., Luong, A. T., Trinh, T. H., Ho, P. H., Meesad, P., & Nguyen, T. T. (2021). Intelligent fruit recognition system using deep learning. In *International Conference on Computing and Information Technology* (pp. 13–22). Springer. https://link.springer.com/chapter/10.1007/978-3-030-79757-7_2.

Nyalala, I., Okinda, C., Nyalala, L., Makange, N., Chao, Q., Chao, L., Yousaf, K., & Chen, K. (2019). Tomato volume and mass estimation using computer vision and machine learning algorithms: Cherry tomato model. *Journal of Food Engineering*, *263*, 288–298. https://www.sciencedirect.com/science/article/pii/S0260877419302973.

Pan, H., Xie, R., & He, Q. (2024). Fruit detection and recognition with deep learning. In *Fourth International Conference on Computer Vision and Data Mining (ICCVDM 2023)* (pp. 562–565). SPIE volume 13063. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13063/130632C/Fruit-detection-and-recognition-with-deep-learning/10.1117/12.3021359.short.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch, . https://openreview.net/forum?id=BJJsrmfCZ.

Rachmawati, E., Supriana, I., Khodra, M. L., & Firdaus, F. (2022). Integrating semantic features in fruit recognition based on perceptual color and semantic template. *Information Processing in Agriculture*,

9, 316–334. https://www.sciencedirect.com/science/article/pii/S2214317321000214.

Rehman, T. U., Mahmud, M. S., Chang, Y. K., Jin, J., & Shin, J. (2019). Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and electronics in agriculture*, *156*, 585–605. https://www.sciencedirect.com/science/article/pii/S0168169918304289.

Saikumar, A., Nickhil, C., & Badwaik, L. S. (2023). Physicochemical characterization of elephant apple (dillenia indica l.) fruit and its mass and volume modeling using computer vision. *Scientia Horticulturae*, *314*, 111947. https://www.sciencedirect.com/science/article/pii/S030442382300122X.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520). http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626). http://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Siddique, M. A. I., & Srizon, A. Y. (2023). An effective dimensionality reduction workflow for the enhancement of automated date fruit recognition utilizing several machine learning classifiers. In *International Conference on Big Data, IoT and Machine Learning* (pp. 363–378). Springer. https://link.springer.com/chapter/10.1007/978-981-99-8937-9_25.

Sun, Q., Chai, X., Zeng, Z., Zhou, G., & Sun, T. (2021). Multi-level feature fusion for fruit bearing branch keypoint detection. *Computers and Electronics in Agriculture*, *191*, 106479. https://www.sciencedirect.com/science/article/pii/S0168169921004968.

Tan, M., & Le, Q. V. (2019). Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, . https://arxiv.org/abs/1907.09595.

Taner, A., Mengstu, M. T., Selvi, K. Ç., Duran, H., Gür, İ., & Ungureanu, N. (2024). Apple varieties classification using deep features and machine learning. *Agriculture*, *14*, 252. https://www.mdpi.com/2077-0472/14/2/252.

Waltner, G., Schwarz, M., Ladstätter, S., Weber, A., Luley, P., Lindschinger, M., Schmid, I., Scheitz, W., Bischof, H., & Paletta, L. (2017). Personalized dietary self-management using mobile vision-based assistance. 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. 385–393). Springer. https://link.springer.com/chapter/10.1007/978-3-319-70742-6_36.

Wang, Y., Li, L., Liu, Y., Cui, Q., Ning, J., & Zhang, Z. (2021). Enhanced quality monitoring during black tea processing by the fusion of nirs and computer vision. *Journal of Food Engineering*, *304*, 110599. https://www.sciencedirect.com/science/article/pii/S0260877421001242.

Xu, M., Wang, J., & Gu, S. (2019). Rapid identification of tea quality by e-nose and computer vision combining with a synergetic data fusion strategy. *Journal of Food Engineering*, *241*, 10–17. https://www.sciencedirect.com/science/article/pii/S0260877418303091.

Xu, P., Fang, N., Liu, N., Lin, F., Yang, S., & Ning, J. (2022). Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation. *Computers and Electronics in Agriculture*, *197*, 106991. https://www.sciencedirect.com/science/article/pii/S0168169922003088.

Yang, Y., Han, Y., Li, S., Yang, Y., Zhang, M., & Li, H. (2023). Vision based fruit recognition and positioning technology for harvesting robots. *Computers and Electronics in Agriculture*, *213*, 108258. https://www.sciencedirect.com/science/article/pii/S0168169923006464.

Yogesh, Dubey, A. K., Ratan, R., & Rocha, A. (2020). Computer vision based analysis and detection of defects in fruits causes due to nutrients deficiency. *Cluster Computing*, *23*, 1817–1826. https://link.springer.com/article/10.1007/s10586-019-03029-6.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848–6856). http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_ShuffleNet_An_Extremely_CVPR_2018_paper.html.

Zhu, Y., Gu, Q., Zhao, Y., Wan, H., Wang, R., Zhang, X., & Cheng, Y. (2022). Quantitative extraction and evaluation of tomato fruit phenotypes based on image recognition. *Frontiers in Plant Science*, *13*, 859290. https://www.frontiersin.org/articles/10.3389/fpls.2022.859290/full.

Ziaratban, A., Azadbakht, M., & Ghasemnezhad, A. (2017). Modeling of volume and surface area of apple from their geometric characteristics and artificial neural network. *International Journal of Food Properties*, *20*, 762–768. https://www.tandfonline.com/doi/abs/10.1080/10942912.2016.1180533.