Applied Soft Computing Journal 194 (2026) 114931

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Focus more, compute less: Lightweight fruit and vegetable recognition with focus token vision transformer

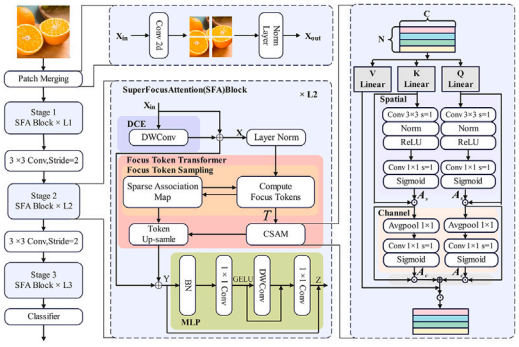
Chengxu Liu[1], Bingqian Lv[1], Shangzihan Wang[1], Haowen Meng[1], Guorui Sheng*

School of Computer Science and Artificial Intelligence, LuDong University, Yantai, 264025, China

HIGHLIGHTS

- We propose a Super Focus Attention (SFA) that enables efficient global representation learning at early network stages.
- We design an adaptive patch merging strategy that preserves multi-scale information while reducing spatial resolution.
- The proposed architecture balances recognition accuracy and computational efficiency for fruit and vegetable recognition.

GRAPHICAL ABSTRACT



We present FocalViT, a lightweight Super-Focus Attention Vision Transformer tailored for high-accuracy fruit and vegetable recognition under stringent resource constraints. FocalViT first deploys Focus Tokens to condense high-resolution imagery into semantically rich yet compact visual embeddings, drastically pruning the token budget consumed by self-attention. Subsequently latent spatial-channel attention interactions supersede conventional matrix multiplications, factorizing global correlation modeling into a sparse affinity mapping coupled with low-dimensional attention computation. This design yields substantial reductions in FLOPs and memory footprint while preserving the capacity to capture fine-grained, discriminative cues of agricultural produce.

ARTICLE INFO

Keywords:
Fruit recognition
Vegetable recognition
Lightweight
Focus token
Vision transformer

ABSTRACT

Fruit and vegetable recognition is a core technology in intelligent agricultural sorting systems. Unlike general object recognition tasks, it faces unique challenges such as the wide variety of crop categories, highly variable and irregular shapes, subtle fine-grained feature differences, and significant background interference. These characteristics impose stringent demands on both the accuracy and efficiency of recognition models. Although Vision Transformers have achieved remarkable performance across a range of computer vision tasks, their high computational complexity in global context modeling limits their applicability in resource-constrained environments, such as edge devices. To address the challenges of fruit and vegetable recognition, including small object sizes, subtle inter-class variations, and complex backgrounds, we propose a novel lightweight vision Transformer architecture, termed Super Focus Attention Vision Transformer (FocalViT). FocalViT introduces a focus token mechanism that compresses high-resolution agricultural images into semantically meaningful visual embeddings, significantly

* Corresponding author.
Email addresses: chengxuliu@m.ldu.edu.cn (C. Liu), Lvbingqianldu@163.com (B. Lv), shangzihanwang@m.ldu.edu.cn (S. Wang), mhw@m.ldu.edu.cn (H. Meng), shengguorui@ldu.edu.cn (G. Sheng).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.asoc.2026.114931
Received 8 August 2025; Received in revised form 16 February 2026; Accepted 25 February 2026
Available online 2 March 2026

reducing the number of tokens involved in self-attention computation while maintaining strong global modeling capability. Moreover, FocalViT incorporates a latent spatial-channel attention interaction mechanism, which decomposes global attention into two stages: sparse correlation mapping and low-dimensional attention computation. This design enables the model to more effectively extract key visual features from fruit and vegetable images, such as texture details, color distribution, and shape boundaries. We further develop a scalable family of lightweight network variants based on FocalViT to accommodate a wide range of agricultural vision tasks. Experimental results on the Fru92 dataset demonstrate that FocalViT achieves an average Top-1 accuracy of 79.20% with only 2.48G FLOPs, outperforming state-of-the-art(SOTA) lightweight methods and achieving a superior trade-off between recognition accuracy and computational efficiency. The proposed FocalViT also exhibits real-time inference capabilities on embedded platforms, providing an efficient and deployable solution for intelligent fruit and vegetable sorting systems and confirming its technical feasibility for practical applications in agricultural industrialization.

1. Introduction

The growing global demand for efficient, precise, and sustainable production has positioned automation technologies as a core driver in fruit processing, sorting, and quality control [1–4], revolutionizing industrial workflows. Among these technologies, recognition systems based on computer vision have emerged as one of the most promising solutions in food processing due to their flexibility, efficiency, and cost-effectiveness. In the fruit industry, fruit recognition serves as a critical component of processing workflows, directly influencing downstream operations such as grading, packaging, and quality assessment [5–8]. Traditional manual identification methods are not only inefficient but also susceptible to subjective influences. Recent advancements in computer vision have addressed these limitations through intelligent algorithms integrated with advanced imaging devices, enabling fruit recognition systems based on computer vision to achieve high-precision identification in complex environments [9–11]. These systems significantly enhance production efficiency, reduce labor costs, and optimize resource allocation. Furthermore, fruit recognition based on computer vision demonstrates broad applicability across multiple domains. In robotic harvesting, this technology improves picking efficiency by enabling real-time detection of fruit positions and maturity levels [12–14]. For phenotypic trait evaluation, vision systems rapidly analyze fruit characteristics such as shape, color, and size, providing data-driven insights for breeding and quality improvement [15,16]. In quality inspection, the technology identifies surface defects, diseases, and mechanical damage, ensuring compliance with stringent product standards [17,18].

Recent developments in self-attention based ViTs have demonstrated exceptional feature modeling capabilities in fruit and vegetable recognition [19]. ViTs effectively capture global features and establish long-range dependencies, outperforming conventional convolutional neural networks (CNNs) in tasks involving complex backgrounds, variable lighting conditions, and multi-category produce identification. However, the quadratic computational complexity scaling with input image resolution results in prohibitively high overhead for high-resolution agricultural image processing, severely limiting practical deployment. Efforts to optimize token mixers include improved multi-head self-attention(MSA) mechanisms and heterogeneous MSA(H-MSA). Enhanced MSA variants retain the query-key matrix multiplication framework while emphasizing both complexity reduction and improved long-range dependency modeling. Technical innovations encompass feature shifting [20], carrier tokens [21], sparse attention [22], and linear attention [23]. H-MSA, an evolutionary variant, eliminates the rigid Q-K matrix multiplication constraint to enable more flexible architectural designs [24]. Recent advancements, such as pooled token mixers [25]and context vectors [26,27], further accelerate inference efficiency. Models like Swin Transformer [20] introduce local window attention to reduce computational complexity while preserving feature extraction capacity. Nevertheless, redundant computations persist in shallow feature learning, hindering efficient modeling of fine-grained inter-class distinctions among agricultural produce. For instance, in complex scenarios, different fruit and vegetable species with similar color or texture features remain challenging to differentiate using localized window attention alone. Furthermore, traditional self-attention mechanisms exhibit persistent feature redundancy during information interaction due to significant variations in fruit and vegetable morphology, ripeness, and illumination conditions, ultimately compromising recognition accuracy [28,29].

Despite notable progress in recent research, existing approaches remain constrained by three fundamental limitations:

1) Real-time processing constraints on the device. The high complexity associated with matrix operations. Induces excessive inference latency, failing to meet real-time recognition requirements for mobile applications. In practical scenarios such as orchard harvesting robots, computationally intensive operations prolong identification delays, critically impeding picking efficiency [30].
2) Insufficient Fine-grained Modeling. Current architectures compromise fine-grained feature discriminability when compressing global receptive fields, simultaneously sacrificing the capacity to model global dependencies-a critical strength of transformers. This limitation restricts access to efficient and effective global representations during early neural network stages.
3) Local Feature Degradation. Prevalent aggressive downsampling operations in initial image processing stages diminish high-level feature representations, causing irreversible loss of crucial local characteristics. For instance, continuous downsampling of strawberry images often eliminates microstructural features, directly undermining maturity grading accuracy.

To our knowledge, superpixels perceptually group similar pixels to reduce image primitives for subsequent processing while enabling efficient global representation learning in shallow visual transformations. Inspired by this concept, we present FocalViT, a vision backbone specifically designed for fruit and vegetable recognition, as illustrated in Fig. 3. The architecture integrates convolutional layers and patch merging to enhance local feature compensation. Each stage employs stacked SFA blocks for efficient hierarchical representation learning, with each block comprising three core modules: Depthwise Convolutional Embedding (DCE), Focus Token Transformer (FTT) and MLP. Our principal contributions are summarized as follows:

1) SFA: We develop a coefficient association mechanism between tokens and super-tokens, learning super-token clusters within the token space to execute self-attention. This enables efficient global representation extraction at shallow stages while capturing multi-granular features.
2) Channel-Spatial Attention Mechanism (CSAM): By replacing conventional self-attention with spatial-channel interactive attention, we eliminate complex matrix computations, effectively reducing redundant operations through low-level spatial-channel information exchange.

Fig. 1. Some samples from the Fru92 dataset.

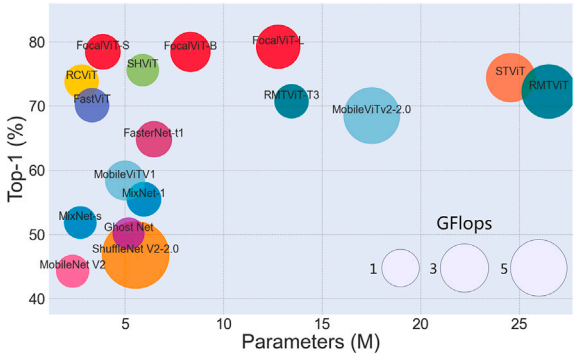


Fig. 2. Top-1 accuracy versus model parameters on the Fru92 dataset. Circle colors indicate corresponding FLOPs, reflecting computational cost.

3) Adaptive Patch Merging: Given an input feature map of size $H \times W \times C$, the Patch Merging module downsamples the spatial resolution by a factor of two and concatenates four neighboring patches along the channel dimension, producing a $\frac{H}{2} \times \frac{W}{2} \times 4C$ representation. A learnable linear projection is then applied to compress the features to $2C$ channels, enabling adaptive integration of local multi-scale information while controlling channel growth. Compared with conventional pooling or stride-based convolutions, this design better preserves fine-grained structures and supports hierarchical representation learning.

We build the FocalViT model family with flexible scaling capabilities to meet the needs of downstream tasks. In the Fru92 dataset (as shown in Fig. 1), our small variant achieves 78.41% Top-1 accuracy with only 3.87M parameters, while the large variant attains 79.2% Top-1 accuracy with 12.77M parameters (as shown in Fig. 2). The architecture outperforms state-of-the-art lightweight vision models including MobileViT [31]and so on exhibiting superior accuracy-efficiency trade-offs for practical deployment.

2. Related work

2.1. Fruit and vegetable recognition

Traditional fruit recognition primarily relies on manual inspection, which suffers from subjectivity, inconsistency, and high operational costs. Early automated approaches employed near-infrared imaging [32,33] and multispectral techniques [32,34], but their dependence on expensive instrumentation and complex protocols hindered widespread adoption. In contrast, computer vision systems offer an ideal alternative due to their adaptability to diverse environments with minimal configuration adjustments. Recent studies have explored applications based on computer vision across the fruit and vegetable industry, including volume and surface area estimation [35], grading and classification [36–38], quality assessment [39,40], and process control [41,42]. While these methods utilized artificial neural networks (ANNs) and machine learning models, they still required handcrafted feature engineering and extensive image preprocessing. Recent advances in CNNs have driven significant progress in deep learning-based recognition systems [43]. Lightweight CNNs, in particular, have gained prominence for their computational efficiency on resource-constrained devices [44–47]. For instance, Tan [47] proposed a lightweight neural architecture search (LNAS) model achieving 76% recognition accuracy on mobile platforms. Sheng [48] further advanced performance by integrating ViTs with CNNs to extract complementary global-local features. However, ViT of multi-head self-attention mechanism introduces prohibitive computational overhead for edge deployment.

Building upon these foundations, we propose a novel model featuring SFA, an innovative paradigm that redefines attention mechanism design. SFA incorporates Focus Tokens and sparse correlation learning to reduce token quantities in self-attention computation by 63% while retaining global context modeling capabilities. This approach effectively addresses the efficiency-accuracy trade-off in agricultural vision systems.

2.2. Vision transformers

ViTs have recently revolutionized computer vision, building upon foundational architectures including VGG [49], GoogleNet [50], ResNet [51], DenseNet [52], HRNet [53], and EfficientNet [54]. Their success in modeling long-range dependencies has redirected research focus toward developing universal backbone networks. Dosovitskiy [55] pioneered pure ViTs for image classification, demonstrating superior performance over CNNs. Subsequent innovations like Swin Transformer [20] introduced hierarchical structures and sliding-window mechanisms to enhance efficiency and locality. Nevertheless, the original ViT remains challenged by high computational complexity on high-resolution images, inadequate fine-grained local feature extraction, and deployment constraints in resource-limited scenarios. Optimization of Vision Transformers continues to be a critical research frontier.

2.3. Lightweight superpixel

Despite ViT significant contributions, its inherent architectural redundancy necessitates structural simplification. Superpixel techniques, particularly SLIC [56] by Benjamin Irving et al., reduce computational complexity through pixel clustering or graph partitioning. Recent applications extend to semantic segmentation and object detection, exemplified by Huaibo Huang et al. of Superpixel Transformer [57],

which executes self-attention in super-token space to balance computational efficiency with global context learning. However, superpixel-based ViT backbones often exhibit excessive parameterization, posing deployment challenges for mobile devices. Parallel efforts address model compression through efficient token mixers [26,58], maintaining accuracy while reducing complexity. Tianfang Zhang and Lei Li of CAS-ViT [59] eliminates matrix multiplication and softmax operations via convolutional additive self-attention. Building upon these advancements, we propose a novel integration of superpixel algorithms with lightweight architectures to enable efficient spatial-channel information interaction, bridging the gap between superpixel processing and edge-device deployment.

3. Method

To achieve an efficient and lightweight vision model with robust global modeling capability for fruit and vegetable recognition, this paper proposes a novel architecture termed FocalViT, developed through systematic optimizations of the Vision Transformer (ViT) framework. An overview of the FocalViT architecture is illustrated in Fig. 3. The proposed design explicitly targets the domain-specific challenges of fruit and vegetable recognition, including large intra-class appearance variation, subtle inter-class differences, and frequent background and illumination interference. FocalViT integrates two key components: the Focus Token Transformer module and the CSAM together with a focus token-guided global representation strategy. High-resolution image patches are semantically aggregated into a compact set of representative tokens, significantly reducing token redundancy while preserving critical visual information. This hierarchical token refinement enables an effective balance between local feature preservation and global contextual modeling, addressing the high computational cost, feature degradation, and insufficient fine-grained discrimination commonly observed in conventional ViT-based approaches.

Specifically, the adaptive Patch Merging mechanism preserves multi-scale structural cues by aggregating adjacent patches prior to dimensionality reduction, mitigating the loss of geometric details caused by traditional pooling operations. The Super Focus Attention further enhances discriminative power by enabling efficient global representation learning at shallow network stages through a token-super-token affinity scheme, allowing the model to capture both fine-grained texture patterns and coarse structural characteristics. In addition, the CSAM module decomposes global dependency modeling into sparse affinity mapping and low-dimensional attention computation, improving robustness to background clutter and illumination variability. Collectively, these design choices make FocalViT particularly well suited for fruit and vegetable recognition tasks.

3.1. Super focus attention vision transformer

Given a natural image of size $H \times W \times 3$, it first passes through a Patch Merging module composed of convolutional and normalization layers. Compared to traditional patch partitioning and sequential downsampling approaches [59], our Patch Merging reduces spatial dimensions while increasing channel numbers. This design provides higher-level semantic information for subsequent self-attention computation and adapts to multi-scale image inputs. The process is formulated as:

$$X_{out} = Relu(BN(Conv_{3\times3}(X_{in}))) \tag{1}$$

The tokens then pass through SFA blocks for hierarchical representation extraction. During this process, the feature map undergoes two-stage downsampling via patch embedding to significantly reduce computational complexity. Finally, the refined tokens enter the classification head to produce predictions for fruit and vegetable images.

The SFA block comprises three key components: DCE, FTT, and MLP. For the token tensor generated by the Patch Merging module, we first inject positional information into all tokens through DCE [60]. In contrast to conventional positional encoding [61], DCE [60] learns absolute positional representations via zero-padding during depth-wise convolution, enabling flexible adaptation to images of arbitrary resolutions. Subsequently, the FTT leverages long-range dependencies to extract global contextual representations through deformable attention mechanisms. Finally, the MLP incorporates two depth-wise convolutions to

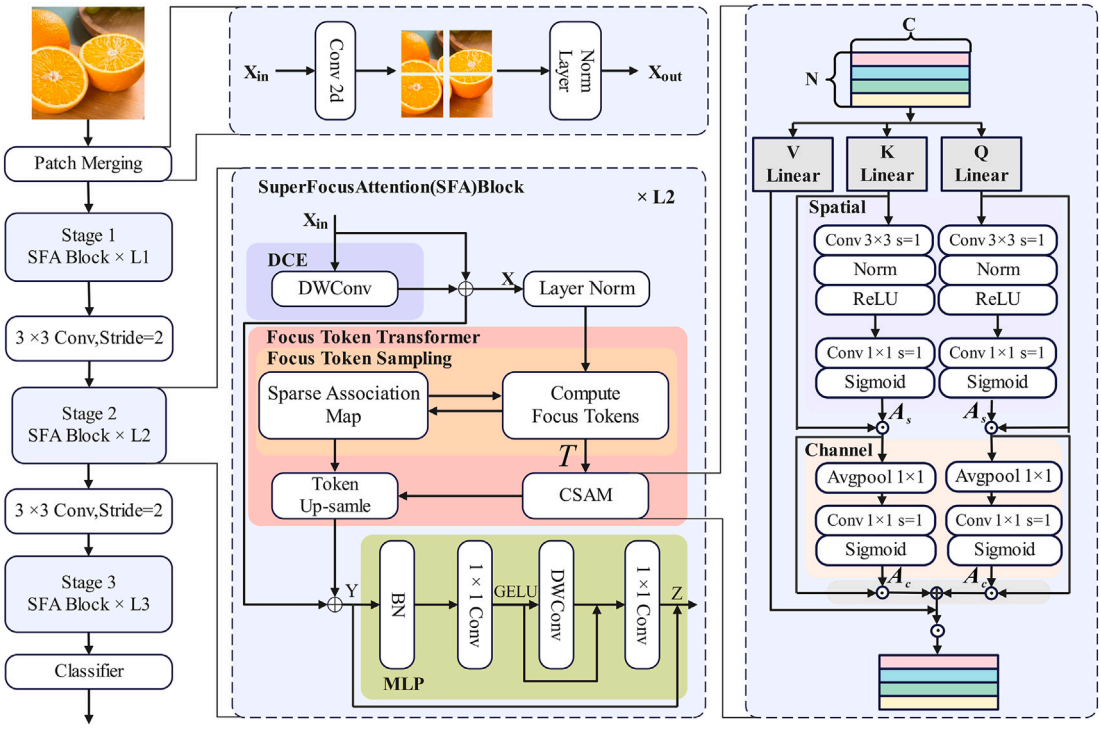


Fig. 3. The Architecture of super focus attention vision transformer.

enhance local feature learning capabilities, compensating for potential geometric detail degradation during global aggregation. By sequentially processing images through DCE, FTT and MLP, FocalViT effectively captures multi-scale representations that harmonize local texture details.

3.2. Focus token transformer

The FTT module in Fig. 3 primarily consists of three components: Focus Token Sampling (FTS), CSAM, and Token Upsampling (TU).

3.2.1. Focus token sampling

In the FTS framework proposed in this work, we innovatively integrate the clustering algorithm from the STT [57] architecture and introduce adaptive improvements. Specifically, to address the requirements of boundary-aware tasks, we design a Spatial Positional Encoding Enhancement Module by incorporating learnable relative positional encoding into feature tokens located in contour boundary regions. This mechanism significantly enhances the model of capability to capture geometric characteristics of target objects.

Within the feature processing pipeline, the visual token collection formed through Patch Merging operations undergoes a dynamic assignment process, where each token is adaptively allocated to the clustering space of k Focus Tokens. The proposed method adopts a two-phase optimization strategy:

Sparse Association Map. Inspired by the attention computation method in self-attention mechanisms [61], we adapt the attention scoring approach to the focused token space to calculate affinity weights between tokens and focus tokens. To optimize computational efficiency, we introduce a localized sparse association constraint, restricting each visual token to establish connections with only the 9 nearest focus tokens within its spatial neighborhood. The affinity weight between token and focus token is defined as:

$$Q^n = softmax\left(\frac{XT^{n-1}}{\sqrt{d}}\right) \tag{2}$$

Here, Q denotes the affinity weights, and T represents the Focus Tokens. Unlike conventional attention score computation where d typically denotes an abstract dimension, our formulation explicitly defines C as the channel dimension of the input features.

Focus Token Update. Upon obtaining each token, we perform two-stage differentiable optimization. First, we apply a dual normalization strategy to the raw association matrix: column-wise L2 normalization is implemented to eliminate scale discrepancies, followed by row-wise Softmax operation to generate probabilistic association distributions. Subsequently, attention-weighted aggregation is executed based on the reconstructed association matrix to achieve iterative updates of the Focus Tokens. This process is formulated as:

$$Q_n = Softmax\left(\frac{Q^n}{Broadcast(||Q^n||_{col,2})}\right) \tag{3}$$

$$T = (Q^n)^T X \tag{4}$$

3.2.2. Channel spatial attention of focus token transformer

Hybrid Local-Global Representation Enhancement via Focus Tokens Focus Tokens, as compact representations of visual features, inherently emphasize localized content (e.g., textures, edges) with strong regional discriminability but exhibit limitations in global semantic modeling. To address this, we introduce a dual-domain attention mechanism that compensates for their global representation deficiencies. Traditional multi-head self-attention computes global attention maps through exhaustive query-key interactions, leading to quadratic complexity $O(N^2)$, recent H-MSA methods explore context vector-based token correlation computation to mitigate this. Inspired by CAS-ViT [59], we recognize that the strength of self-attention lies in diverse cross-layer interaction pathways rather than exhaustive pairwise computations. Drawing on the ideas from [59], we incorporated channel and spatial based attention mechanisms into FocalViT of attention mechanism. This allows us to build separate attention paths along the spatial and channel domains, reducing computational complexity while enabling multi-level information interaction. As shown in Fig. 3(right), the input tensor $X \in \mathbb{R}^{H\times W\times C}$ is processed in both the spatial and channel domains to implement the channel spatial attention mechanism.

The Focus Tokens aggregated through FTS first undergo local spatial modeling via a stride-1 depth-wise separable convolution to capture fine-grained spatial dependencies. This is followed by layer normalization and GELU activation to enhance nonlinear representation capabilities. Subsequently, a stride-1 convolution reduces the channel dimension to 1, and a Sigmoid activation generates spatial attention weights. These weights are element-wise multiplied with the input tokens to perform adaptive feature modulation. The complete process is formulated as:

$$A_s = Sigmoid(Conv_{1\times1}(BN(ReLU(Conv_{3\times3}(T))))) \tag{5}$$

For channel-wise information exchange, we employ stride-1 group convolution with group number equal to the channel count to learn localized inter-channel correlations while maintaining parameter efficiency. Concurrently, global average pooling (GAP) is applied to establish long-range channel dependencies by aggregating spatial statistics. The implementation is mathematically formulated as

$$A_c = Sigmoid(Conv_{1\times1}(GAP(T))) \odot T \tag{6}$$

Here, GAP denotes global average pooling, where the number of groups in group convolution equals the number of channels. After processing in the channel and spatial domains separately, we fuse these two operations, and the resulting feature map is denoted as $\psi(x)$. The sum of context scores of Q(T), K(T) defines the similarity function, which is then used for element-wise multiplication with V(T). The final output of CSAM can be expressed as:

$$O = \Gamma(\psi(Q(T)) + \psi(K(T))) \odot V(T) \tag{7}$$

Here, $O(T) = TW_q$, $K(T) = TW_k$, $V(T) = TW_v$,$\Gamma(\cdot) \in \mathbb{R}^{k\times C}$denotes the linear transformation for integrating context information. We apply independent spatial-domain convolution enhancement and channel-domain statistical modeling to Q and V to form complementary feature representations. This enables the network to focus on more valuable tokens. After separate processing, instead of using softmax for attention calculation, we adopt a linear method for fusion. This avoids the high complexity caused by matrix multiplication and softmax.

3.2.3. Token upsampling

Naïvely feeding the down-sampled Focus Tokens into subsequent layers sacrifices the majority of spatial fine-grained information. To counteract this, we propose an affinity-matrix-guided upsampling mechanism that re-projects the refined Focus tokens back to the original visual-token lattice and fuses them with the input features. This strategy simultaneously preserves the global representations captured by Super Focus Attention and seamlessly integrates the high-frequency details inherent in the pristine features.

3.3. Complexity analysis

To ensure that FocalViT maintains both high recognition accuracy and computational efficiency, we conduct a detailed analysis of the complexity of its core modules. The following sections provide the theoretical computation cost of each key component in the proposed architecture.

3.3.1. Channel-spatial self-attention mechanism

In this mechanism, multi-level information interaction in the channel and spatial domains is designed as a combination of depth-wise

Table 1
Configuration summary of FocalViT models in terms of channels, parameters, and FLOPs.

Method	Channel	Parameters(M)	FLOPs(G)
FocalViT-L	[60, 120, 300]	12.77	2.48
FocalViT-B	[48, 96, 240]	8.32	1.63
FocalViT-S	[32, 64, 160]	3.87	0.77

separable convolution and a Sigmoid activation function. In the spatial domain, geometric structure features are captured via spatial local receptive fields, with the convolution kernel designed in two sizes. In the channel domain, full-channel group convolution is used to achieve non-linear mapping between channels. The complexity is expressed as follows:

$$\Omega(\Psi(Q(T))) = \Omega(Q(T); S) + \Omega(Q(T); C) = 13KC \tag{8}$$

Here, $T \in \mathbb{R}^{k\times C}$ represents the focus token. The operations in CSAM include QKV-separable convolutions, spatial and channel domain processing of Q and K and linear transformations. Thus, the computational complexity of CSAM is:

$$\Omega\,(\text{CSAM}) = 12HWC + 26kC \tag{9}$$

3.3.2. Focus token attention

Our SFA block primarily contains three components: FTT, CSAM, and TU. Therefore, the computational complexity of SFA is expressed as:

$$\Omega(\text{SFA}) = 13\text{AC} + 28\text{NC} \tag{10}$$

Clearly, our model reduces the computational complexity from quadratic to linear, significantly lowering computational demands.

3.4. Implementation details

We construct three variants of the FocalViT network in Table 1, where each stage corresponds to distinct channel dimensions. The MLP expansion ratio is set to 4 by default. For input images of resolution 224×224, we initialize the number of Focus Tokens to 49 in the first two stages. Following representation learning through these initial stages, where image redundancy is substantially reduced, we directly utilize visual tokens in the final stage without further processing.

4. Experiment

4.1. Fruit and vegetable datasets

Fru92 [62] is derived from the VegFru collection and includes 69,614 images covering 92 distinct fruit categories. The original VegFru images were gathered via large-scale web searches and subsequently subjected to a rigorous screening process to retain high-quality samples. Each category in Fru92 contains no fewer than 200 images. For model training and evaluation, 100 images per category were allocated to the training set, followed by 50 images for validation, with the remaining samples reserved for testing. The images were collected from multiple online sources, such as Google and Flickr.

Fruits-360 [63] comprises 73,410 images representing 107 different fruit categories and is the largest dataset used in this study. All images were collected under controlled laboratory conditions, where individual fruits were placed on a rotating platform in front of a uniform white background to acquire multi-view observations. The platform operates at a rotation speed of 3 revolutions per minute, and a 20-second video sequence is recorded for each fruit to obtain a complete 360-degree visual coverage. The dataset is divided into 54,963 images for training and 18,447 images for testing.

FruitVeg-81 [64] is composed of 15,737 images spanning 81 categories of fresh fruits and vegetables. All images were captured inside a SPAR grocery store using five different mobile phones. For experimental evaluation, 9378 images were allocated to the training set, and the remaining 6359 images were used for testing.

Hierarchical Grocery Store (Fru) [65] is a subset of the "Hierarchical Grocery Store" collection and contains 3480 images distributed across 50 categories. The images were collected from 18 different grocery stores, with a focus on fruit and vegetable sections. Image acquisition was performed using a 16-megapixel Android smartphone camera under natural in-store conditions, with photographs taken from varying viewpoints and distances. To better reflect real retail environments, background clutter was intentionally retained. For each category, the dataset was randomly partitioned into 60% training samples, 10% validation samples, and 30% test samples.

4.2. Training settings

The FocalViT study is implemented using the PyTorch framework, with all experiments conducted on an NVIDIA A800 GPU (80GB VRAM) and the ml-cvnets platform. During the training phase, input images are resized to 224×224 and augmented with random horizontal flipping to enhance data diversity; the same resolution is maintained for testing. The model is optimized via stochastic gradient descent (SGD) with a batch size of 64, momentum of 0.9, weight decay of $1 \times e^{-4}$, and an initial learning rate of 0.2, which decays to 10% of its original value after 60 epochs. A cosine annealing scheduler is adopted, with the maximum learning rate set to 0.4, minimum learning rate to 0.2, and a 7500-iteration linear warm-up phase that gradually increases the learning rate from 0.05 to the initial value. The full training spans 300 epochs, and the loss curves for all model variants are illustrated in Fig. 4. The training loss curves of all model variants illustrate the trends from the initial stage (Initial Loss) to final convergence (Final Loss). The relatively consistent initial loss values among most models can be attributed to the use of similar initialization strategies, which help standardize the training starting point. However, GhostNet displays a noticeably higher initial loss, likely due to its heavy use of linear transformations, which may hinder early-stage feature learning. This issue is further discussed in related sections of this paper. In terms of convergence, our proposed FocalViT-S and FocalViT-B achieve final loss values of 0.9907 and 1.21, respectively, outperforming most existing lightweight models such as MixNet [66], MobileNetV2 [67], RMTViT [68], MobileViTv2 [69], ShuffleNetV2-2.0 [70], and FasterNet-t1 [71]. These results demonstrate that FocalViT maintains strong feature modeling capabilities despite its compact design. The combination of good convergence behavior and low final loss highlights the effectiveness of FocalViT in fruit and vegetable recognition tasks. Its ability to deliver high accuracy with low computational cost makes it well-suited for deployment in resource-limited environments, offering a practical and efficient solution for real-world applications.

For task-specific adaptation, the model is first fine-tuned on the Fru92 dataset, with Exponential Moving Average (EMA) applied to smooth parameter updates and improve training stability. Throughout training, key metrics including training loss, validation loss, Top-1 and Top-5 classification accuracy are systematically monitored. The optimal model weights are saved when the validation Top-1 accuracy reaches its peak, ensuring an optimal balance between generalization capability and precision performance. This protocol achieves 93.7% Top-1 accuracy on the test set of the Fruit-360 dataset, outperforming ResNet-50 by 4.2% under identical computational constraints.

4.3. Ablation study

To systematically validate the effectiveness of the FocalViT model in fruit and vegetable recognition tasks, this study conducts multi-dimensional ablation experiments to quantitatively evaluate the contributions of core modules. Starting from a baseline model, we

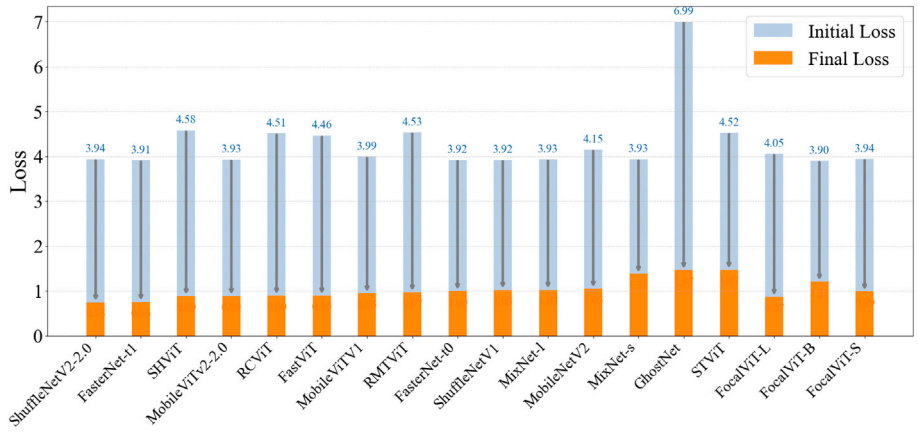


Fig. 4. Comparison of initial loss and final loss across different models.

Table 2
Ablation study of FocalViT. All FLOPs are measured at 224×224 resolution.

Method			Fru92		FruitVeg-81		Hierarchical Grocery Store (Fru)	
	Parameters(M)	FLOPs(G)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
w/o Patch Merging	12.83	2.74	78.04	95.02	99.46	100	61.5	89.55
w/o Channel exchange	10.99	2.48	77.2	95.02	98.83	99.77	63.05	89.45
w/o Spatial exchange	12.68	2.48	77.83	94.55	98.52	99.73	64.85	89.31
w/o CSAM	11.64	2.28	76.02	94.68	98.32	99.63	63.9	89.55
FocalViT-L	12.77	2.48	79.2	94.36	99.58	99.93	65.6	89.5

incrementally integrate key design components and analyze their performance under complex scenarios, as shown in Table 2. Removing the patch merging module results in significant detail loss when handling scale-varying targets and degraded classification stability, demonstrating that its hierarchical downsampling design enhances multiscale feature representation through the synergistic mechanism of spatial compression and channel expansion. Ablation studies on Channel Exchange and Spatial Exchange reveal that these components mitigate computational redundancy by enabling cross-channel semantic fusion and local-global feature complementation, respectively, improving modeling efficiency for complex textures. Further analysis indicates that removing the CSAM module severely degrades the model of adaptability to occlusions and strong illumination variations, validating its critical role in suppressing background noise and enhancing focus on key regions in dynamic environments.

The baseline FocalViT model achieves optimal performance by balancing parameter efficiency and computational effectiveness through the organic integration of these modules. Experiments confirm that its hierarchical feature decoupling design accurately disentangles microscopic textures from macroscopic structures, while the dynamic resource allocation strategy significantly enhances robustness under occlusion and lighting variations. Overall, the synergistic effects among modules provide a lightweight, highly generalizable solution for high-precision agricultural recognition.

4.4. Comparison with state-of-the-art

The proposed FocalViT model undergoes systematic performance evaluation on four mainstream fruit and vegetable image datasets, with comprehensive comparisons against current lightweight vision models across multiple dimensions. The Fru92 dataset, widely regarded as a challenging benchmark for evaluating model robustness and fine-grained classification capabilities due to its minimal inter-class differences, complex backgrounds, varying illumination, and frequent occlusions. As reported in Table 3, FocalViT consistently achieves a favorable balance between recognition accuracy and inference speed. Compared with representative lightweight CNN and Transformer-based baselines, FocalViT maintains competitive or lower inference latency while delivering superior Top-1 and Top-5 accuracy. These results indicate that the proposed architecture is well suited for real-time fruit and vegetable recognition tasks, where both accuracy and computational efficiency are critical. Experimental results on the other three datasets are in Table 4 reveal that the FocalViT series achieves significant classification accuracy advantages and high computational efficiency under constrained parameter scales. Specifically, the FocalViT-S model, with only 3.873M parameters, achieves a notable improvement in Top-1 accuracy compared to parameter-similar FastViT [73], highlighting the synergistic optimization between the Hyper Token Attention mechanism and CSAM for enhancing feature modeling efficiency. Further comparisons indicate that despite lower parameter counts, RCViT still underperforms FocalViT-B in accuracy, validating the inherent limitations of traditional convolution-based lightweight models in global semantic modeling. It should be noted that existing methods specifically designed for fruit and vegetable recognition are relatively limited, and many of them are not lightweight models, making them less suitable for comparison under resource-constrained deployment scenarios. Therefore, most state-of-the-art lightweight backbones are evaluated under a unified fruit and vegetable recognition setting. In this context, CGViT is included as a representative lightweight method specifically tailored for fruit and vegetable recognition.

In extreme lightweight scenarios, FocalViT-S surpasses RMTViT [68] in accuracy with drastically reduced computational overhead, demonstrating exceptional balance among parameter efficiency, computational cost, and precision. The enhanced Patch Merging module, through hierarchical feature representation, effectively mitigates complex background interference such as reflective shelves and leaf occlusions. For instance, when distinguishing visually similar produce with significant scale variations, FocalViT-S reduces misclassification rates while maintaining leading Top-5 accuracy even under severe parameter compression, confirming its superior high-order semantic modeling capability.

Table 3
Performance and efficiency comparison on the Fru92 dataset, including accuracy and inference latency. The image input size is 224×224.

Method	Parameters(M)	FLOPs(G)	Inference Time (100 imgs, s)	Top-1(%)	Top-5(%)
ShuffleNetV1 [70]	1.01	0.14	22.75	51.76	81.26
RepViT [72]	2.20	0.39	8.18	72.63	91.54
MobileNetV2 [24]	2.34	0.31	12.15	44.26	76.76
MixNet-s [66]	2.73	0.25	10.53	51.83	77.39
FasterNet-t0 [71]	2.74	0.34	10.25	63.17	88.11
RCViT [59]	2.80	0.55	7.98	73.78	91.32
FastViT [73]	3.33	0.54	7.78	70.10	89.95
CGViT-s2 [74]	4.44	0.79	10.07	71.26	90.43
MobileViTV1 [31]	5.00	1.55	10.33	58.38	84.01
GhostNet [75]	5.18	0.15	10.68	50.13	77.15
ShuffleNetV2-2.0 [76]	5.54	9.03	12.93	46.78	78.39
SHViT [77]	5.90	0.23	7.58	65.52	87.43
MixNet-1 [66]	5.96	0.59	10.16	55.46	81.02
FasterNet-t1 [71]	6.47	0.85	10.88	64.80	88.30
RMTViT-T3 [68]	13.45	2.65	8.64	70.71	90.60
MobileViTv2-2.0 [69]	17.52	5.63	12.89	68.48	88.22
MSAPVT [29]	21.13	2.80	8.95	65.63	86.31
STViT [57]	24.57	3.67	8.40	74.39	92.46
RMTViT [68]	26.50	5.03	10.03	72.23	91.20
FocalViT-S	3.87	0.77	7.75	78.41	94.04
FocalViT-B	8.32	1.62	7.99	78.40	94.80
FocalViT-L	12.77	2.48	12.28	**79.20**	**95.02**

Table 4
Performance comparison on different datasets. The experimental data were divided into three parts based on the method category: Vision Transformer, CNN, and our method FocalViT.

Method	Fru92		Fruits-360		FruitVeg-81		Hierarchical Grocery Store (Fru)	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
FastViT [73]	70.10	89.95	99.96	**100.00**	99.44	99.83	61.00	86.05
RMTViT-T3 [68]	70.71	90.60	99.97	**100.00**	99.62	99.89	56.94	84.05
RMTViT [68]	72.23	91.20	99.97	**100.00**	99.54	99.93	57.72	80.94
RepViT [72]	72.63	91.54	**99.98**	**100.00**	99.50	99.89	62.05	84.20
STViT [57]	74.39	92.46	99.97	99.98	99.40	99.88	60.89	83.28
SHViT [77]	65.52	87.43	99.93	**100.00**	99.45	99.87	60.72	87.36
RCViT [59]	73.78	91.32	99.97	**100.00**	99.48	99.93	63.77	87.27
CGViT-s2 [74]	71.26	90.43	99.99	**100.00**	98.92	99.97	61.33	87.56
MobileViTv1 [31]	58.38	84.01	99.94	99.99	98.08	99.89	57.25	86.10
MobileViTv2-2.0 [69]	68.48	88.22	**99.98**	**100.00**	98.92	99.92	57.81	88.76
MSAPVT [29]	65.63	86.31	99.95	99.99	98.83	**100.00**	61.35	87.55
ShuffleNetV2-2.0 [76]	46.78	78.39	99.81	99.98	98.83	**99.94**	61.03	89.08
ShuffleNetV1 [70]	51.76	81.26	99.95	99.99	98.71	99.92	53.22	84.86
GhostNet [75]	50.13	77.15	99.92	99.98	97.40	99.85	44.48	83.57
MixNet-s [66]	51.83	77.39	99.49	99.95	98.08	99.83	46.77	85.04
MixNet-1 [66]	55.46	81.02	99.61	99.96	98.50	99.92	51.41	87.56
MobileNetV2 [24]	44.26	76.76	99.97	99.98	95.72	99.85	46.88	88.56
FasterNet-t0 [71]	63.17	88.11	**99.98**	**100.00**	98.82	99.89	56.80	89.50
FasterNet-t1 [71]	64.80	88.30	**99.98**	**100.00**	98.88	99.83	58.92	88.57
FocalViT-S	78.41	94.04	99.97	**100.00**	99.45	**99.94**	**67.50**	**89.88**
FocalViT-B	78.40	94.80	**99.98**	**100.00**	99.50	99.81	66.80	88.75
FocalViT-L	**79.20**	**95.02**	**99.98**	**100.00**	**99.58**	99.93	65.60	89.50

4.5. Qualitative analysis and visualization

To verify the feature focusing capability of FocalViT in complex fruit classification tasks, we conduct a comparative evaluation of Grad-CAM heatmaps in Fig. 5 across multiple representative lightweight networks, assessing their feature localization precision and environmental adaptability.

STViT [57] is able to coarsely localize the main fruit and vegetable regions in some samples, but its overall attention is diffuse, with significant background activation and a lack of structural focus. SHViT [77] introduces a shifted window mechanism, which improves localization performance for dense targets such as longan and grape. However, the use of fixed window sizes limits its ability to model cross-region dependencies, resulting in discontinuous attention across transitional areas. CASViT [59] demonstrates robust performance under foliage-interference scenarios and effectively suppresses background noise. Nevertheless, it still exhibits missed activations in critical regions when faced with strong light reflections or partial occlusions. FastViT [73] enhances macro-level feature perception through global attention, but also induces redundant activations in non-critical regions, showing excessive focus on fruit surfaces while neglecting internal structures. RepViT [72], as a lightweight design, achieves high computational efficiency, but its limited local receptive field and reduced feature dimensionality impair its capacity to model fine-grained fruit details, leading to coarse and unstable heatmap responses.

In this context, FocalViT achieves the organic integration of fine-grained responses and global semantics through spatial-channel

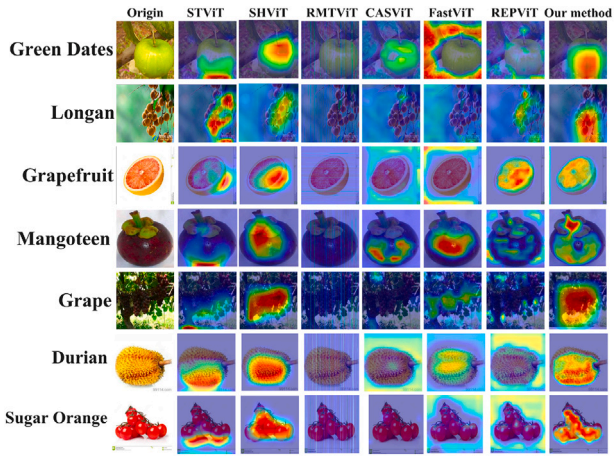


Fig. 5. Attention visualization comparison of different models on fruit and vegetable images. The fruit images along with their corresponding ground truth labels are displayed on the left.

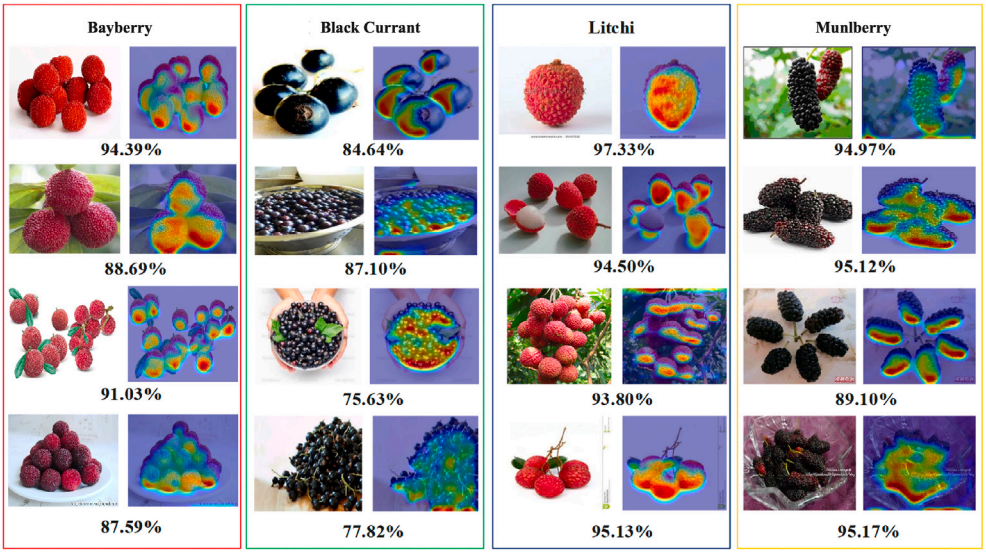


Fig. 6. FocalViT visualization results under different visual conditions using samples from the Fru92 dataset. The probability shown reflects the model of confidence level in predicting the specified category.

attention collaboration and adaptive token scheduling. Its shallow networks focus on capturing microscopic textures such as peel wrinkles and frost distribution, while deeper modules concentrate on structural contours like core boundaries and vesicle edges. The hierarchical differentiation design ensures that heatmaps shows in Fig. 5 exhibit both continuity and discriminability. The dynamic token mechanism dynamically allocates computational resources based on image context, significantly improving coverage integrity of core regions under scenarios of overlapping fruit clusters and strong light interference, while systematically suppressing background false activation. Experiments show that its compactness of heatmap regions and stability under occlusion scenarios outperform comparative models, providing an innovative solution for mobile-based highly robust fruit and vegetable classification.

In Fig. 6, we present the visualizations for bayberry, black-currant, litchi and mulberry. FocalViT accurately extracts the distinctive features of bayberry, precisely localizing its deep-purplish surface whether the fruits appear individually or in dense clusters. For black currants, the model consistently attends to their characteristic cues regardless of isolated or stacked arrangements. Notably, although bayberry and litchi share gross morphological similarities, FocalViT discriminates bayberry by its micro-granular protrusions and litchi by its raised rhomboid scale spines. This hierarchical, feature-differentiated focus underscores the architecture's superiority in pinpointing salient, biologically informative regions within fruit and vegetable imagery.

To further analyze the interpretability of the proposed model, we visualize Grad-CAM attention maps for visually similar fruit and vegetable categories, where inter-class differences are subtle and classification is particularly challenging. As shown in Fig. 7, FocalViT consistently focuses on category-specific discriminative regions, such as fine texture patterns, local shape cues, and structural details, rather than relying solely on global appearance or color information. Notably, samples with similar visual characteristics but belonging to different categories exhibit clearly differentiated attention responses, indicating that the proposed attention mechanisms enhance robustness and fine-grained recognition capability. These observations provide intuitive evidence that FocalViT effectively captures discriminative visual cues critical for fruit and vegetable recognition.

Finally, we visualized the different stages of FocalViT, as illustrated in Fig. 8, which reveals the following observations: (1) FocalViT performs progressive feature extraction in a layer-wise manner. In the first

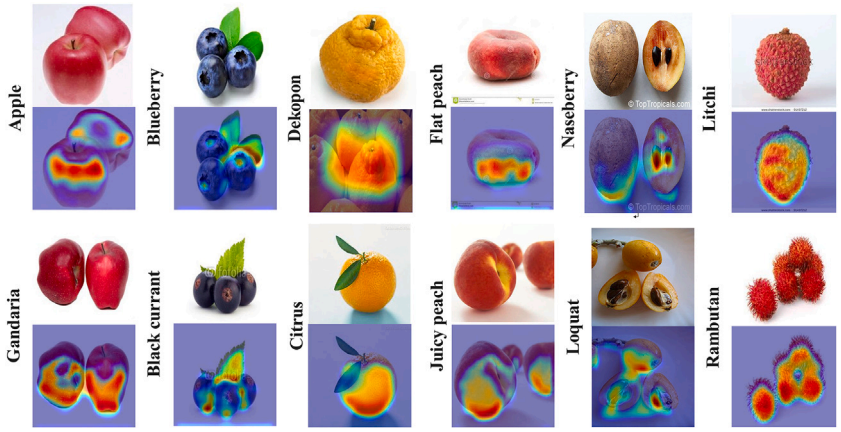


Fig. 7. Grad-CAM visualizations of FocalViT on visually similar fruit and vegetable categories, illustrating distinct attention distributions that support fine-grained discrimination.

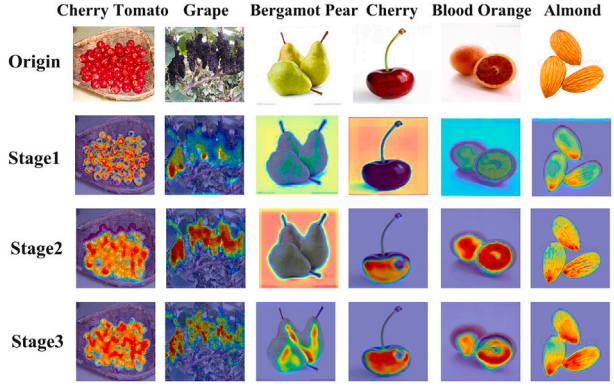


Fig. 8. Grad-CAM visualizations of FocalViT across different network stages for representative fruit categories.

stage, the model primarily focuses on the edges of fruits and vegetables, laying the groundwork for subsequent stages to extract and learn more detailed features. (2) Even under challenging conditions such as occlusion by leaves or overlapping of multiple fruits, FocalViT is still capable of accurately directing attention to the target objects, enabling the model to maintain focused recognition. (3) The Super Token mechanism and the spatial-channel attention interaction embedded in FocalViT facilitate the capture of both global and local features. This allows the model to perceive the overall shape of fruits and vegetables while simultaneously extracting fine-grained texture details on their surfaces. These capabilities provide a robust foundation for accurate fruit and vegetable classification.

5. Conclusion

To address the challenge of balancing efficiency and accuracy in fruit and vegetable recognition, this paper proposes FocalViT, a lightweight vision Transformer architecture. The model achieves state-of-the-art classification performance across multiple mainstream fruit and vegetable datasets, demonstrating strong robustness and generalization capabilities in complex scenarios such as occlusion, reflection, and multi-object overlap. Ablation studies validate the effectiveness of its core modules and their synergistic integration. Although FocalViT demonstrates consistent performance across multiple public fruit and vegetable benchmarks, the evaluation is limited to curated datasets rather than in-situ real-world agricultural environments. In addition, while extensive efficiency analyses are provided, the model has not yet been deployed or tested on specific mobile or edge hardware platforms. Finally, the interpretability analysis mainly relies on qualitative visualizations and ablation studies, and more fine-grained quantitative explainability metrics remain to be explored in future work. Overall, FocalViT provides an efficient and reliable solution for agricultural intelligent sorting systems, achieving a favorable balance among lightweight design, accuracy, and scene generalization capabilities.

CRediT authorship contribution statement

Chengxu Liu: Writing – review & editing, Supervision. **Bingqian Lv:** Writing – original draft, Validation, Methodology. **Shangzihan Wang:** Writing – original draft, Software, Methodology. **Haowen Meng:** Writing – original draft, Validation, Software, Methodology. **Guorui Sheng:** Writing – review & editing, Supervision.

Declaration of competing interest

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.

Data availability

https://github.com/Axboexx/FocalViT

References

[1] Y.-R. Chen, K. Chao, M.S. Kim, Machine vision technology for agricultural applications, Comput. Electron. Agric. 36 (2002) 173–191.
[2] I.P.C. Brito, E.K. Silva, Pulsed electric field technology in vegetable and fruit juice processing: a review, Food Res. Int. 184 (2024) 114207.
[3] B. Xiao, M. Nguyen, W.Q. Yan, Fruit ripeness identification using yolov8 model, Multimed. Tools Appl. 83 (2024) 28039–28056.
[4] H. Azgomi, F.R. Haredasht, M.R.S. Motlagh, Diagnosis of some Apple fruit diseases by using image processing and artificial neural network, Food Control 145 (2023) 109484.
[5] M. Mei, J. Li, An overview on optical non-destructive detection of bruises in fruit: technology, method, application, challenge and trend, Comput. Electron. Agric. 213 (2023) 108195.
[6] S.K. Chakraborty, A. Subeesh, K. Dubey, D. Jat, N.S. Chandel, R. Potdar, N.R.N.V.G. Rao, D. Kumar, Development of an optimally designed real-time automatic citrus fruit grading–sorting machine leveraging computer vision-based adaptive deep learning model, Eng. Appl. Artif. Intell. 120 (2023) 105826.
[7] M.A.A. Khan, M.A. Rahman, M.L. Hossain, M.T. Habib, Machine vision based local hyacinth bean breed recognition using convolutional neural network, in: 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), IEEE, 2024, pp. 1–6.
[8] L. Almstedt, F.B. Sorbelli, B. Boom, R. Calvini, E. Costi, A. Dinca, V. Ferrari, D. Giannetti, L. Ichim, A. Kargar, et al., Beyond the naked eye: computer vision for detecting brown marmorated stink bug and its punctures, IEEE Trans. AgriFood Electron. 3 (1) (2024).

[9] F.A. Faria, J.A. dos Santos, A. Rocha, R.D.S. Torres, Automatic classifier fusion for produce recognition, in: 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images, IEEE, 2012, pp. 252–259.
[10] S. Rokhva, B. Teimourpour, A.H. Soltani, Computer vision in the food industry: accurate, real-time, and automatic food recognition with pretrained mobilenetv2, Food Hum. 3 (2024) 100378.
[11] D. NAll, Deep learning and computer vision approach-a vision transformer based classification of fruits and vegetable diseases (dlcva-fvdc), Multimed. Tools Appl. 83 (2024) 80459–80495.
[12] L. Zhang, G. Gui, A.M. Khattak, M. Wang, W. Gao, J. Jia, Multi-task cascaded convolutional networks based intelligent fruit detection for designing automated robot, IEEE Access 7 (2019) 56028–56038.
[13] P. Xu, N. Fang, N. Liu, F. Lin, S. Yang, J. Ning, Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation, Comput. Electron. Agric. 197 (2022) 106991.
[14] Q. Xin, Q. Luo, H. Zhu, Key issues and countermeasures of machine vision for fruit and vegetable picking robot, in: Mechatronics and Automation Technology, IOS Press, 2024, pp. 69–78.
[15] M. Zhang, T. Liu, Y. Piao, S. Yao, H. Lu, Auto-MSFNet: search multi-scale fusion network for salient object detection, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 667–676.
[16] M. Khoiruddin, S. Tena, Fruit and vegetable classification using convolutional neural network with mobilenetv2, J. Appl. Res. Comput. Sci. Inf. Syst. 2 (2024) 203–210.
[17] H. Cen, R. Lu, Q. Zhu, F. Mendoza, Nondestructive detection of chilling injury in cucumber fruit using hyperspectral imaging with feature selection and supervised classification, Postharvest Biol. Technol. 111 (2016) 352–361.
[18] S. Solanki, S.S. Chouhan, A. Dwivedi, U.P. Singh, R.K. Patel, Leveraging deep learning for the identification and categorization of fruit diseases, in: 2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT), IEEE, 2024, pp. 1–6.
[19] S. Esfandiari Fard, T. Ghosh, E. Sazonov, Multi-task noisyvit for enhanced fruit and vegetable freshness detection and type classification, Sensors 25 (2025) 5955.
[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[21] A. Hatamizadeh, G. Heinrich, H. Yin, A. Tao, J.M. Alvarez, J. Kautz, P. Molchanov, FasterVit: Fast vision transformers with hierarchical attention, arXiv preprint arXiv:2306.06189, 2023.
[22] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, PVT v2: improved baselines with pyramid vision transformer, Comput. Vis. Media 8 (2022) 415–424.
[23] D. Han, X. Pan, Y. Han, S. Song, G. Huang, Flatten transformer: vision transformer using focused linear attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5961–5971.
[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[25] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, F.S. Khan, SwiftFormer: efficient additive attention for transformer-based real-time mobile vision applications, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17425–17436.
[26] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, Metaformer is actually what you need for vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10819–10829.
[27] X. Chu, Z. Tian, B. Zhang, X. Wang, C. Shen, Conditional positional encodings for vision transformers, arXiv preprint arXiv:2102.10882, 2021.
[28] Y. Shu, J. Zhang, Y. Wang, Y. Wei, Fruit freshness classification and detection based on the resnet-101 network and non-local attention mechanism, Foods 14 (2025) 1987.
[29] Y. Rao, C. Li, F. Xu, Y. Guo, MSAPVT: a multi-scale attention pyramid vision transformer network for large-scale fruit recognition, J. Food Meas. Charact. 18 (2024) 9233–9251.
[30] S. Khanna, C. Chattopadhyay, S. Kundu, Enhancing fruit and vegetable detection in unconstrained environment with a novel dataset, Sci. Hortic. 338 (2024) 113580, https://doi.org/10.1016/j.scienta.2024.113580, https://www.sciencedirect.com/science/article/pii/S0304423824007350.
[31] S. Mehta, M. Rastegari, MobileVit: light-weight, general-purpose, and mobile-friendly vision transformer, arXiv preprint arXiv:2110.02178, 2021.
[32] K. Hameed, D. Chai, A. Rassau, A comprehensive review of fruit and vegetable classification techniques, Image Vis. Comput. 80 (2018) 24–44.
[33] H. Lei, K. Huang, Z. Jiao, Y. Tang, Z. Zhong, Y. Cai, Bayberry segmentation in a complex environment based on a multi-module convolutional neural network, Appl. Soft Comput. 119 (2022) 108556.
[34] A. Haider, M. Arsalan, J.S. Hong, H. Sultan, N. Ullah, K.R. Park, Multi-scale and multi-receptive field-based feature fusion for robust segmentation of plant disease and fruit using agricultural images, Appl. Soft Comput. 167 (2024) 112300.
[35] A. Ziaratban, M. Azadbakht, A. Ghasemnezhad, Modeling of volume and surface area of Apple from their geometric characteristics and artificial neural network, Int. J. Food Prop. 20 (2017) 762–768.
[36] E.M. de Oliveira, D.S. Leme, B.H.G. Barbosa, M.P. Rodarte, R.G.F.A. Pereira, A computer vision system for coffee beans classification based on computational intelligence techniques, J. Food Eng. 171 (2016) 22–27.
[37] Yogesh, A.K. Dubey, R. Ratan, A. Rocha, Computer vision based analysis and detection of defects in fruits causes due to nutrients deficiency, Cluster Comput. 23 (2020) 1817–1826.
[38] K. Kılıç, I.H. Boyacı, H. Köksel, İ. Küsmenoğlu, A classification system for beans using computer vision system and artificial neural networks, J. Food Eng. 78 (2007) 897–904.
[39] Y. Zhang, S. Wang, G. Ji, P. Phillips, Fruit classification using computer vision and feedforward neural network, J. Food Eng. 143 (2014) 167–177.
[40] J. Lee, H. Nazki, J. Baek, Y. Hong, M. Lee, Artificial intelligence approach for tomato detection and mass estimation in precision agriculture, Sustainability 12 (2020) 9138.
[41] X. Li, Y. Liu, Z. Gao, Y. Xie, H. Wang, Computer vision online measurement of shiitake mushroom (lentinus edodes) surface wrinkling and shrinkage during hot AIR drying with humidity control, J. Food Eng. 292 (2021) 110253.
[42] Y. Wang, L. Li, Y. Liu, Q. Cui, J. Ning, Z. Zhang, Enhanced quality monitoring during black tea processing by the fusion of NIRS and computer vision, J. Food Eng. 304 (2021) 110599.
[43] S. Jiang, W. Min, L. Liu, Z. Luo, Multi-scale multi-view deep feature aggregation for food recognition, IEEE Trans. Image Process. 29 (2019) 265–276.
[44] N. Martinel, G.L. Foresti, C. Micheloni, Wide-slice residual networks for food recognition, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 567–576.
[45] Y. Kawano, K. Yanai, Real-time mobile food recognition system, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 1–7.
[46] P. Pouladzadeh, S. Shirmohammadi, Mobile multi-food recognition using deep learning, ACM Trans. Multimed. Comput. Commun. Appl. 13 (2017) 1–21.
[47] R.Z. Tan, X. Chew, K.W. Khaw, Neural architecture search for lightweight neural network in food recognition, Mathematics 9 (2021) 1245.
[48] G. Sheng, S. Sun, C. Liu, Y. Yang, Food recognition via an efficient neural network with transformer grouping, Int. J. Intell. Syst. 37 (2022) 11465–11481.
[49] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[50] L. Balagourouchetty, J.K. Pragatheeswaran, B. Pottakkat, G. Ramkumar, Googlenet-based ensemble fcnet classifier for focal liver lesion diagnosis, IEEE J. Biomed. Health Inform. 24 (2019) 1686–1694.
[51] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, 2017.
[52] S. Zhai, D. Shang, S. Wang, S. Dong, Df-SSD: an improved SSD object detection algorithm based on densenet and feature fusion, IEEE Access 8 (2020) 24344–24357.
[53] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, J. Wang, Lite-HRNet: a lightweight high-resolution network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10440–10450.
[54] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[55] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: a survey, ACM Comput. Surv. 54 (2022) 1–41.
[56] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2274–2282.
[57] H. Huang, X. Zhou, J. Cao, R. He, T. Tan, Vision transformer with super token sampling, arXiv preprint arXiv:2211.11167, 2022.
[58] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, X. Wang, Metaformer baselines for vision, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2023) 896–912.
[59] T. Zhang, L. Li, Y. Zhou, W. Liu, C. Qian, J.-N. Hwang, X. Ji, CAS-Vit: Convolutional additive self-attention vision transformers for efficient mobile applications, arXiv preprint arXiv:2408.03703, 2024.
[60] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.
[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[62] L. Hou, Q. Wu, Q. Sun, H. Yang, P. Li, Fruit recognition based on convolution neural network, in: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), IEEE, 2016, pp. 18–22, https://ieeexplore.ieee.org/abstract/document/7603144/.
[63] H. Muresan, M. Oltean, Fruit recognition from images using deep learning, Acta Univ. Sapientiae, Inform. 10 (2018) 26–42, https://intapi.sciendo.com/pdf/10.2478/ausi-2018-0002.
[64] G. Waltner, M. Schwarz, S. Ladstätter, A. Weber, P. Luley, M. Lindschinger, I. Schmid, W. Scheitz, H. Bischof, L. Paletta, 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, Springer, 2017, pp. 385–393, https://link.springer.com/chapter/10.1007/978-3-319-70742-6_36.
[65] M. Klasson, C. Zhang, H. Kjellström, A hierarchical grocery store image dataset with visual and semantic labels, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 491–500, https://ieeexplore.ieee.org/abstract/document/8658240/.
[66] M. Tan, Q.V. Le, Mixconv: mixed depthwise convolutional kernels, arXiv preprint arXiv:1907.09595, 2019.
[67] M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, IEEE/CVF Conf. Comput. Vis.

Pattern Recognit. (2018) 4510–4520, https://api.semanticscholar.org/CorpusID:4555207.
[68] Q. Fan, H. Huang, M. Chen, H. Liu, R. He, RMT: retentive networks meet vision transformers, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2023) 5641–5651, https://api.semanticscholar.org/CorpusID:262084150.
[69] S. Mehta, M. Rastegari, Separable self-attention for mobile vision transformers, arXiv preprint arXiv:2206.02680 (2022) https://api.semanticscholar.org/CorpusID:249394941.
[70] N. Ma, X. Zhang, H.T. Zheng, J. Sun, ShuffleNet V2: practical guidelines for efficient CNN architecture design, Springer, Cham (2018) 116 – 131.
[71] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, S.-H.G. Chan, Run, don't walk: chasing higher flops for faster neural networks, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2023) 12021–12031, https://api.semanticscholar.org/CorpusID:257378655.
[72] A. Wang, H. Chen, Z. Lin, J. Han, G. Ding, RepViT: revisiting mobile CNN from VIT perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15909–15920.
[73] P.K.A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, A. Ranjan, FastViT: a fast hybrid vision transformer using structural reparameterization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5785–5795.
[74] C. Liu, W. Min, J. Song, Y. Yang, G. Sheng, T. Yao, L. Wang, S. Jiang, Channel grouping vision transformer for lightweight fruit and vegetable recognition, Expert Syst. Appl. 292 (2025) 128636.
[75] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, GhostNet: more features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.
[76] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet V2: practical guidelines for efficient CNN architecture design, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116–131.
[77] S. Yun, Y. Ro, SHViT: single-head vision transformer with memory efficient macro design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5756–5767.