

Article

Low-Cost, Nondestructive Cultivar Identification of Dried Goji Berries Using RGB Images and a Lightweight LSH-CoAtNet Model

Lei Shi ^{1,2}, Zhaocong Lyu ¹, Yansong Li ¹, Jing Guo ¹, Zhenyang Chen ¹, Cheng Qian ², Zhuo Bai ¹ and Helong Yu ^{1,2,*} 

¹ Institute of Smart Agriculture, Jilin Agricultural University, Changchun 130118, China

² College of Information Technology, Jilin Agricultural University, Changchun 130118, China

* Correspondence: yuhelong@aliyun.com

Abstract

Accurate cultivar identification of commercial dried goji berries is essential for raw material sorting, batch consistency assessment, and quality control during processing and distribution. Conventional approaches based on manual judgment or physicochemical analysis are often subjective, labor-intensive, time-consuming, and costly, making them unsuitable for rapid commercial sorting and quality inspection. To develop a rapid, low-cost, and nondestructive method for dried goji berry cultivar identification, this study proposes a visual recognition framework that integrates RGB imaging with lightweight deep learning. A dataset comprising 25,899 RGB images from five cultivars of commercial dried goji berry samples, namely Ningqi No. 7, Linqi No. 5, Ningqi No. 1, Keqi 6082, and Jingqi No. 1, was constructed. Given the pronounced surface shrinkage, complex texture, and subtle inter-cultivar appearance differences of dried goji berries, an image quality enhancement method was designed to strengthen the representation of color gradation, textural details, and edge information. For model development, CoAtNet was selected as the baseline network and redesigned for lightweight deployment. By integrating an improved feature extraction module and an information-preserving downsampling module, the proposed LSH-CoAtNet model enhances fine-grained feature representation while reducing computational cost. On the quality-enhanced image dataset, the proposed method achieved an accuracy of 98.80%, a precision of 98.81%, a recall of 98.80%, and an F1-score of 98.80%. The model contained only 6.41 M parameters and required 1.60 GFLOPs, outperforming the baseline model in both classification performance and computational efficiency. Ablation experiments and five-fold cross-validation further confirmed the effectiveness of the image quality enhancement strategy, the contribution of each improved module, and the stability of the model. Overall, the proposed method, which combines RGB image quality enhancement with LSH-CoAtNet, provides a low-cost, nondestructive, and efficient technical solution for rapid cultivar identification, raw material sorting, batch consistency assessment, and quality control of commercial dried goji berries during processing and distribution. It may also serve as a reference for intelligent classification and quality inspection of other specialty dried horticultural products.



Academic Editors: Luís Pádua and Anabela A. Fernandes-Silva

Received: 15 May 2026

Revised: 19 June 2026

Accepted: 24 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: dried goji berries; cultivar identification; processed horticultural products; RGB imaging; lightweight deep learning; nondestructive detection

1. Introduction

Goji berry (*Lycium barbarum* L.) is an important specialty horticultural crop in China with substantial edible, medicinal, and economic value. Rich in bioactive compounds such as polysaccharides, carotenoids, flavonoids, and phenolic compounds, it has attracted considerable attention in functional foods, products used as both food and medicine, and specialty agricultural products [1,2]. In production and distribution, goji berries are usually processed, stored, and sold as dried commodities. Different cultivars vary in nutritional quality, bioactive composition, aroma characteristics, and commercial-use traits; moreover, drying affects the color, phenotypic traits, and chemical quality of dried goji berries [3–5]. Therefore, rapid, accurate, low-cost, and nondestructive cultivar identification methods are increasingly needed for commercial inspection and quality control.

At present, cultivar identification and quality evaluation of commercial dried goji berries are still commonly performed using empirical manual judgment and conventional physicochemical analysis. For cultivar identification, manual assessment is usually based on visible traits such as fruit color, size, shape, surface shrinkage, and sensory characteristics. However, this process is easily affected by operator experience, subjective judgment, and environmental conditions, resulting in limited repeatability and consistency [6]. In fruit quality evaluation, physicochemical analytical methods can provide relatively reliable information on nutritional and bioactive components, but they generally require sample pretreatment, specialized instruments, trained personnel, and relatively long analysis time [7]. These limitations restrict their application in rapid, standardized, and large-scale inspection during raw material acceptance, sorting, processing, batch evaluation, and distribution supervision. Therefore, a rapid, nondestructive, and low-cost visual identification method is needed to support standardized cultivar identification and quality control of commercial dried goji berries.

In recent years, near-infrared spectroscopy, hyperspectral imaging, and multimodal detection technologies have been applied to geographical traceability, component analysis, quality evaluation, and classification of goji berry and related *Lycium* products, achieving favorable performance [8–12]. These methods can characterize differences in chemical composition, spectral response, and spatial information, and therefore provide effective tools for origin identification, quality assessment, and product classification. However, despite their high analytical accuracy, these methods generally depend on expensive specialized equipment, strict acquisition conditions, and complex preprocessing and modeling procedures. These requirements may limit their low-cost deployment and rapid adoption in small- and medium-sized processing enterprises, storage and sorting facilities, and market distribution scenarios. By contrast, RGB imaging combined with machine vision has the advantages of low equipment cost, convenient image acquisition, high detection efficiency, and straightforward system integration, and has been widely explored in agricultural product recognition, quality inspection, and intelligent sorting [13]. RGB image-based visual recognition therefore provides a practical and cost-effective pathway for rapid and nondestructive cultivar identification of commercial dried goji berries.

However, several research gaps remain. First, although RGB imaging and deep learning have been widely used in agricultural product recognition, their application to cultivar identification of commercial dried goji berries remains limited. After drying, cultivar-related visual differences become subtle and are mainly reflected in color gradation, shrinkage morphology, surface wrinkle texture, contour characteristics, and local brightness patterns. This indicates that dried goji berry cultivar identification has characteristics of a fine-grained visual recognition task rather than a simple appearance classification problem [14,15]. Second, existing general image preprocessing and enhancement methods are not necessarily designed for the visual characteristics of dried goji berries. Drying-

induced shrinkage, uneven color distribution, and complex surface texture may weaken cultivar-related visual cues, indicating the need for a task-oriented image quality enhancement method that emphasizes color gradation, wrinkle texture, and edge information. Previous studies on colorfulness measurement and local adaptive contrast enhancement have shown that color distribution and local contrast information can be quantitatively characterized and enhanced while preserving visual appearance [16,17]. Third, many deep learning models can achieve high classification accuracy, but their computational cost may limit deployment on sorting equipment, mobile terminals, or edge computing platforms. Therefore, it remains necessary to develop a lightweight model that can balance recognition accuracy, parameter size, computational complexity, and potential edge-device deployment feasibility [18–22].

Based on the above research gaps, this study investigated whether cultivar-specific visual differences in commercial dried goji berries could be better represented by enhancing color gradation, wrinkle texture, and contour information in RGB images, and whether a lightweight CoAtNet-based model could achieve accurate cultivar identification with reduced computational cost. Specifically, this study aimed to: (1) construct an RGB image dataset of five representative commercial dried goji berry cultivars, including Ningqi No. 7, Linqi No. 5, Ningqi No. 1, Keqi 6082, and Jingqi No. 1; (2) develop an image quality enhancement method to improve the representation of color, texture, and edge features; (3) design a lightweight LSH-CoAtNet model by introducing an improved feature extraction module and an information-preserving downsampling module; and (4) evaluate the proposed method in terms of classification performance, model compactness, cross-validation stability, and computational efficiency. This study aims to provide a feasible RGB image-based method for rapid, low-cost, and nondestructive cultivar identification of commercial dried goji berries and to offer a reference for intelligent classification and quality control of dried horticultural products.

2. Materials and Methods

2.1. Dataset Construction and Preprocessing

2.1.1. Image Acquisition and Dataset Construction

The dataset used in this study consisted of dried goji berry samples from five cultivars, namely Ningqi No. 7, Linqi No. 5, Ningqi No. 1, Keqi 6082, and Jingqi No. 1. To reduce the interference of irrelevant background factors while increasing the diversity of imaging conditions, all images were captured in a controlled indoor environment with a fixed shooting distance, a uniform background, and variable illumination settings. The uniform background and fixed shooting distance were used to reduce variations caused by imaging geometry and background interference, whereas the variable illumination settings were used to enrich the brightness and color appearance of the samples in the dataset. This image acquisition protocol was intended to help the model learn cultivar-related discriminative visual features under diverse but controllable imaging conditions.

To construct a single-berry image dataset suitable for subsequent cultivar classification, a three-stage image processing workflow was designed based on the original group images, as shown in Figure 1. The first stage involved image acquisition and preprocessing, including original image input, grayscale conversion, and Gaussian denoising, to reduce noise interference and improve the stability of subsequent segmentation. The second stage involved edge extraction, contour analysis, and target screening. Otsu thresholding [23] with morphological optimization, Canny edge detection [24] with edge enhancement, and contour extraction were sequentially used to locate candidate target regions, while non-target regions were removed according to area and shape constraints. The third stage involved target segmentation and output. Target masks were constructed based on the screening

results to extract individual goji berries, and each target was cropped using its bounding rectangle and saved independently, yielding single-berry images with a black background.

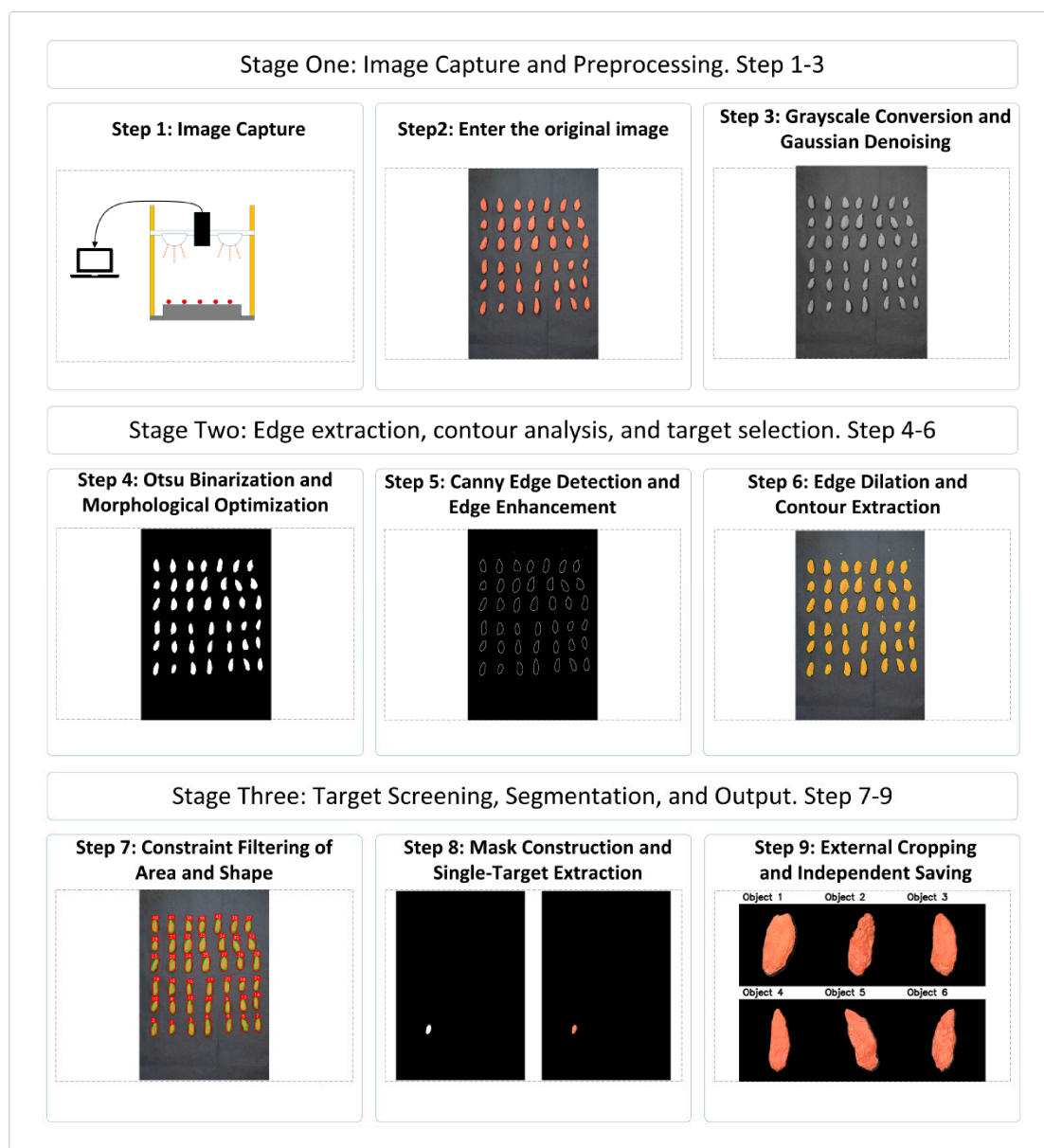


Figure 1. Workflow for extracting individual dried goji berry images from original multi-berry images using image preprocessing, edge detection, contour filtering, mask-based segmentation, and independent cropping.

During preprocessing, each goji berry was treated as an independent target region. The cropped single-berry images were uniformly resized to 224×224 pixels and saved as PNG files in the corresponding folders. Finally, the original samples were divided into training, validation, and test sets at a ratio of 7:2:1.

As shown in Figure 2, five data augmentation operations were applied in this study: vertical flipping, horizontal flipping, rotation, scaling, and translation. Transforming the original images in this way is a practical strategy when the sample size is limited. It increases dataset diversity and helps the model learn more robust features, thereby improving generalization ability [25]. For deep learning models, more training samples provide more learning opportunities and can improve both accuracy and stability. Details of the dataset are shown in Table 1.

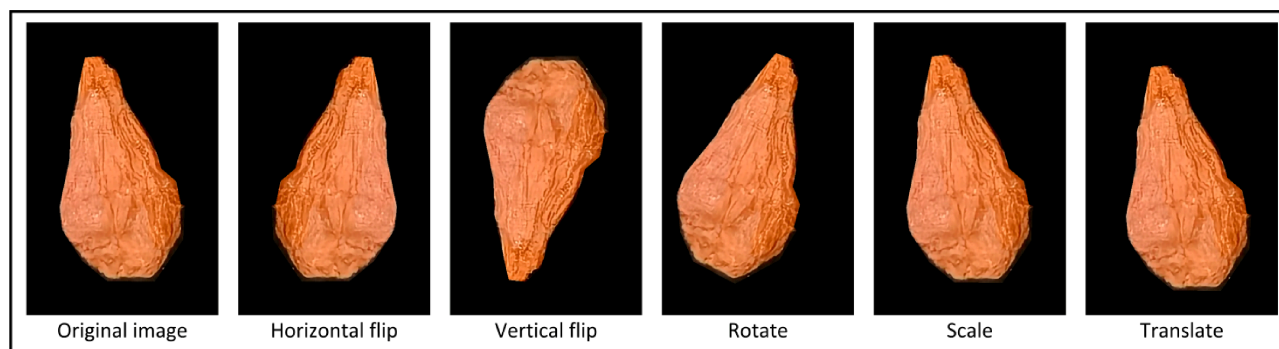


Figure 2. Representative examples of data augmentation operations applied to dried goji berry images, including horizontal flipping, vertical flipping, rotation, scaling, and translation.

Table 1. Distribution of images for the five dried goji berry cultivars, including the total number of images and the numbers in the training, validation, and test sets.

Cultivar	Number of Images	Training	Validation	Test
Ningqi No. 7	5346	3742	1069	535
Linqi No. 5	5178	3624	1035	519
Ningqi No. 1	5340	3737	1068	535
Keqi 6082	5211	3647	1042	522
Jingqi No. 1	4824	3376	964	484
Total	25,899	18,126	5178	2595

2.1.2. Image Quality Enhancement Method

Given the high similarity among cultivars in this dataset and the complexity of the illumination environment, an image quality enhancement method was introduced to extract more informative visual cues. The procedure used in this study is described below:

(1) Color space conversion: Images were first loaded with OpenCV and converted from the default BGR space to RGB to standardize the color channels for subsequent processing. The RGB images were then transformed into Lab space, in which the L channel represents luminance, the a channel represents the red-green component, and the b channel represents the yellow-blue component. By decoupling luminance from chromaticity, this step provides the basis for targeted enhancement of the a and b channels while reducing interference from luminance variation [26].

(2) Adaptive chromaticity enhancement based on wavelet transform: For the a and b channels in Lab space, wavelet transform was used to perform fine-grained enhancement in the frequency domain. First, a two-dimensional discrete wavelet transform (DWT) with the Haar basis was applied separately to the a and b channels, decomposing each channel into a low-frequency sub-band that captures smooth-region information and high-frequency sub-bands that retain edge and detail information. Second, Canny edge detection was applied to the low-frequency sub-bands, and Gaussian blurring was used to generate an edge-strength map describing the spatial distribution of edges. Gain factors were then computed adaptively according to edge strength. Specifically, low gain was assigned to strong-edge regions to avoid edge blurring, whereas higher gain was assigned to smoother regions to enhance subtle chromatic detail. This design preserves the boundary features of goji berries while strengthening discriminative fine details [27]. The gain factor was calculated as follows:

$$\text{gain} = 1.0 + 0.5 \times (1 - \text{edge_strength}) \quad (1)$$

where, for regions with high edge strength (i.e., edge_strength close to 1), $1 - \text{edge_strength}$ approaches 0 and the gain factor approaches 1.0, meaning that only limited enhancement is applied to the high-frequency sub-bands. This helps prevent blur or artifacts caused by over-enhancement near edges. For smoother regions with low edge strength (i.e., edge strength close to 0), $1 - \text{edge_strength}$ approaches 1 and the gain factor approaches 1.5, meaning that stronger enhancement is applied to the high-frequency sub-bands to emphasize subtle chromatic variations and fine details.

(3) Learning-based color fine-tuning using a 3D-LUT: To further enhance color gradation while preserving the natural appearance of goji berry images, a three-dimensional lookup table (3D-LUT) was introduced for nonlinear color mapping. First, a $17 \times 17 \times 17$ color grid covering the RGB space in the range [0,1] was constructed. For each sampling point in the grid, the color was converted to HSV space, the saturation channel (S) was randomly perturbed within -10% to $+30\%$, and the result was converted back to RGB to generate a dynamic LUT that simulates subtle color variation under real acquisition conditions. Second, after the image was transformed from Lab back to RGB, the enhanced image was normalized to [0,1] and mapped to the LUT grid according to pixel values, enabling nonlinear color fine-tuning. This data-driven mapping increases color gradation while maintaining the overall visual tone of the image [28].

(4) No-reference colorfulness control: To avoid oversaturation caused by excessive enhancement, the Hasler and Suesstrunk colorfulness metric was used for quantitative evaluation and correction. The image was decomposed into RGB channels and the colorfulness value was calculated [16]. The corresponding formula is as follows:

$$C = \sqrt{\sigma_{RG}^2 + \sigma_{YB}^2} + 0.3\sqrt{\mu_{RG}^2 + \mu_{YB}^2} \quad (2)$$

$$RG = R - G \quad (3)$$

$$YB = 0.5 \times (R + G) - B \quad (4)$$

where RG denotes the red-green component, YB denotes the yellow-blue component, μ denotes the mean value, and σ denotes the standard deviation. The colorfulness metric proposed by Hasler and Suesstrunk was used as a no-reference quantitative measure of the overall perceived colorfulness of enhanced images [16]. In this study, the colorfulness threshold was set to 100 as a conservative upper bound to avoid excessive saturation while reducing distortion of the original color appearance of dried goji berries. If the calculated colorfulness value C exceeded this threshold, the image was converted to HSV space, and the saturation channel (S) was proportionally reduced by a factor of $\text{max_colorfulness}/C$ until the colorfulness value satisfied the threshold requirement. This operation was used to maintain visual naturalness and reduce the risk of unrealistically saturated colors after enhancement.

(5) Local contrast post-processing: To further emphasize surface contours and wrinkle details, local contrast stretching was applied to the enhanced image in RGB space. First, the local mean and local standard deviation were calculated separately for each channel [29]. The local mean was computed as follows:

$$\mu(x, y) = \frac{1}{Z} \sum_{(i,j) \in \Omega_{(xy)}} G(i - x, j - y; \sigma) \cdot I(i, j) \quad (5)$$

$$G(\Delta x, \Delta y; \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\Delta x^2 + \Delta y^2}{2\sigma^2}\right) \quad (6)$$

$$Z = \sum_{(i,j) \in \Omega_{(xy)}} G(i-x, j-y; \sigma) \quad (7)$$

where $\Delta x = i - x$, $\Delta y = j - y$ denote the coordinate offsets of neighboring pixels relative to the central pixel (x, y) , ensuring that the Gaussian kernel is symmetrically distributed around the center. Z denotes the sum of all Gaussian weights within the neighborhood and is used to eliminate the influence of the total weight on the mean value. The formula for local standard deviation is as follows:

$$\sigma_{\text{local}}(x, y) = \sqrt{\frac{1}{Z} \sum_{(i,j) \in \Omega_{(xy)}} G(i-x, j-y; \sigma) \cdot [I(i, j) - \mu(x, y)]^2 + \epsilon} \quad (8)$$

where $[I(i, j) - \mu(x, y)]^2$ represents the squared deviation of a neighboring pixel from the local mean and reflects the degree to which the pixel departs from the local intensity trend. The Gaussian kernel $G(\cdot)$ was used to weight the squared deviations, so that pixels closer to the center contributed more strongly to the calculation of the local standard deviation. In this way, the local standard deviation could better characterize local variation around the central pixel. The parameter $\epsilon = 10^{-8}$ is a small constant introduced to avoid a zero standard deviation when pixel values within the local neighborhood are identical, thereby preventing numerical instability in the subsequent gain calculation. The normalization factor (Z) was identical to that used for the local mean so that the weighting scheme remained consistent.

Next, the local gain factor was calculated by combining the global standard deviation with the local standard deviation. Previous studies on local adaptive contrast enhancement have shown that local gain functions based on variance or standard deviation can enhance local image details, but excessively large gain values may amplify noise or introduce artifacts in low-variation regions [17,29,30]. Therefore, the gain factor (g) was constrained to the range of 0.1–5.0 in this study. The lower bound of 0.1 was used to avoid excessive suppression of local contrast, whereas the upper bound of 5.0 was used to prevent over-enhancement and noise amplification in regions with very small local standard deviation. This bounded gain strategy was adopted to improve local texture and contour visibility while maintaining image stability. The gain factor was calculated as follows:

$$g = \sigma_{\text{global}} / \sigma_{\text{local}} \quad (9)$$

$$\sigma_{\text{global}} = \sqrt{\frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W [I(x, y) - \mu_{\text{global}}]^2} \quad (10)$$

$$\mu_{\text{global}} = \frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W I(x, y) \quad (11)$$

where H and W denote the image height and width. Finally, local contrast enhancement was performed according to $I_{\text{out}} = \mu_{\text{local}} + (I_{\text{in}} - \mu_{\text{local}}) \cdot g$, after which outlier handling and value clipping were applied to obtain the final enhanced image.

Through the above procedure, this study established a complete image quality enhancement method that balances stronger color and detail representation with visual fidelity, and the entire dataset was processed accordingly.

2.2. Proposed Framework: LSH-CoAtNet

To address the subtle inter-cultivar differences in appearance, the similarity in texture and color, and the demand for computational efficiency in practical deployment, an

improved lightweight image classification framework was constructed on the basis of CoAtNet. The overall architecture is shown in Figure 3.

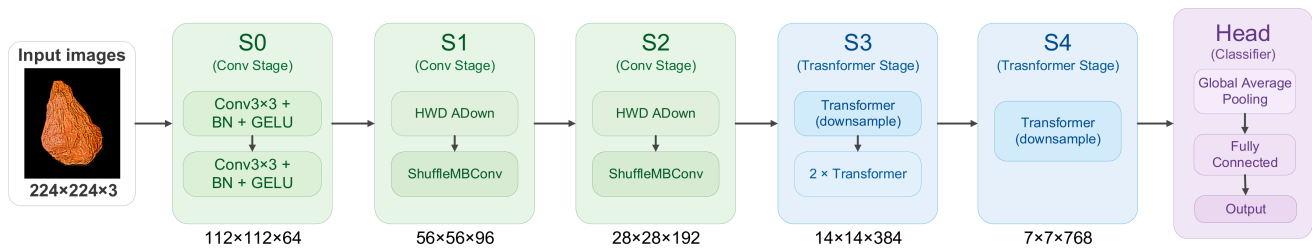


Figure 3. Overall architecture of the proposed LSH-CoAtNet model, including convolutional stages, HWD ADown and ShuffleMBConv modules, Transformer stages, and the final classifier head for dried goji berry cultivar identification.

In this study, CoAtNet was used as the baseline backbone and selectively redesigned within its hierarchical architecture [31]. The hybrid modeling advantage of convolution and attention in CoAtNet was retained, while lightweight attention, an enhanced low-level feature-extraction module, and an information-preserving downsampling module were introduced to better capture subtle inter-cultivar differences under lightweight constraints. Compared with the baseline CoAtNet, the final model improved accuracy on the quality-enhanced dataset from 94.83% to 98.80%.

As shown in the architecture, the proposed improvements can be summarized in three aspects:

(1) Lightweight reconstruction.

To reduce model complexity and improve deployment efficiency, the Transformer component and overall network depth of the baseline CoAtNet were redesigned. Specifically, the original Softmax-based self-attention was replaced with LinearAttention, the relative position bias parameters were removed, and both the number of attention heads and the per-head dimension were reduced to lower the computational and memory cost of attention. At the network level, the stacking depth was pruned by reducing num_blocks from [2, 2, 3, 5, 2] to [2, 2, 2, 3, 1], thereby removing redundant computation. This redesign preserves the hybrid modeling strengths of CoAtNet while providing a lighter backbone for efficient feature extraction and practical deployment.

(2) ShuffleMBConv.

Compared with the original MBConv, ShuffleMBConv first splits the input channels into two branches. One branch uses a lightweight transformation to preserve and transmit basic features, whereas the other passes through an MBConv-based main branch for deeper local modeling. The two branches are then concatenated and channel-shuffled to enable cross-branch information exchange. This design retains the local representation capability and channel-attention advantages of MBConv while reducing the proportion of channels that must pass through the full MBConv path, thereby lowering both computational complexity and parameter cost. In addition, the dual-branch structure and channel shuffle improve feature diversity and fusion efficiency, making the module better suited to high-resolution stages in lightweight networks.

(3) HWD ADown.

Conventional downsampling often reduces resolution at the expense of texture detail. In this study, HWD ADown was used to replace the downsampling operations in stages S1 and S2 of the backbone so that critical texture and contour information could be preserved while spatial dimensions were reduced. This is particularly important for fine-grained goji berry cultivar identification, where inter-class differences are subtle.

By combining these three improvements, the proposed model achieved 98.80% accuracy on the quality-enhanced dataset while reducing the parameter count and computational cost to 6.41 M and 1.60 GFLOPs, respectively, compared with 16.99 M parameters and 3.35 GFLOPs for the baseline CoAtNet. These results indicate a favorable balance between recognition performance and deployment efficiency.

2.3. Lightweight Reconstruction

To reduce model complexity and improve deployment efficiency, the Transformer component and overall network depth were redesigned for lightweight deployment while retaining the hybrid modeling framework of CoAtNet [32,33]. The original CoAtNet uses a Softmax-based multi-head self-attention mechanism [34] and deep hierarchical stacking to enhance feature modeling capability [31], but these designs also introduce high parameter counts and computational overhead [35]. Considering the strong demand for model efficiency in processing, sorting, and distribution inspection of commercial dried goji berries, the baseline model was compressed from two aspects, namely attention computation and network depth, to balance classification performance with computational cost [32,33,36].

2.3.1. Linearized Attention Mechanism

The Transformer module of the original CoAtNet adopts a Softmax-based self-attention mechanism and introduces a relative positional bias table to enhance spatial relationship modeling [37]. However, Softmax-based self-attention requires explicit computation of pairwise similarities between Query and Key on high-resolution feature maps, and its time and storage costs increase rapidly with the number of tokens, thereby increasing deployment cost. Therefore, LinearAttention [38] was used in this study to replace the original attention structure. In addition, parameters related to relative positional bias were removed to further reduce the parameter count and simplify attention computation. As shown in Figure 4, in the present implementation, LinearAttention first generates Query, Key, and Value through linear layers, then uses ReLU as the feature-mapping function to nonlinearly map Query and Key, and replaces Softmax normalization with linear weighted computation. This avoids explicit construction of a full attention matrix and reduces the complexity of the attention module [35,38]. With reference to the single-head dimension setting in the Transformer block of the original CoAtNet, the number of attention heads and the dimension of each head were further reduced to heads = 4 and dim_head = 16, respectively, to decrease the parameters and computation of the attention branch [31]. Experimental results showed that this modification improved inference efficiency while maintaining classification performance, making it more suitable for cultivar identification of commercial dried goji berries in resource-constrained scenarios.

2.3.2. Network Depth Reduction

In addition to the attention module, the hierarchical depth of the baseline network was pruned to reduce redundant computation caused by repeated stacking [32,39]. As shown in Figure 5, the num_blocks configuration of the original CoAtNet-0 is [2, 2, 3, 5, 2], whereas it was reduced to [2, 2, 2, 3, 1] in this study to decrease overall model complexity while keeping the number of channels in each stage unchanged. The core idea of this design is that, for fine-grained visual recognition tasks such as cultivar identification of commercial dried goji berries, shallow and middle-level texture, color, and contour details play important roles [40]. Although excessively deep network stacking can enhance representation capability, it also introduces additional computational burden. Therefore, this study sought a better balance between accuracy and efficiency by reducing the number of deep modules [32,33,39].

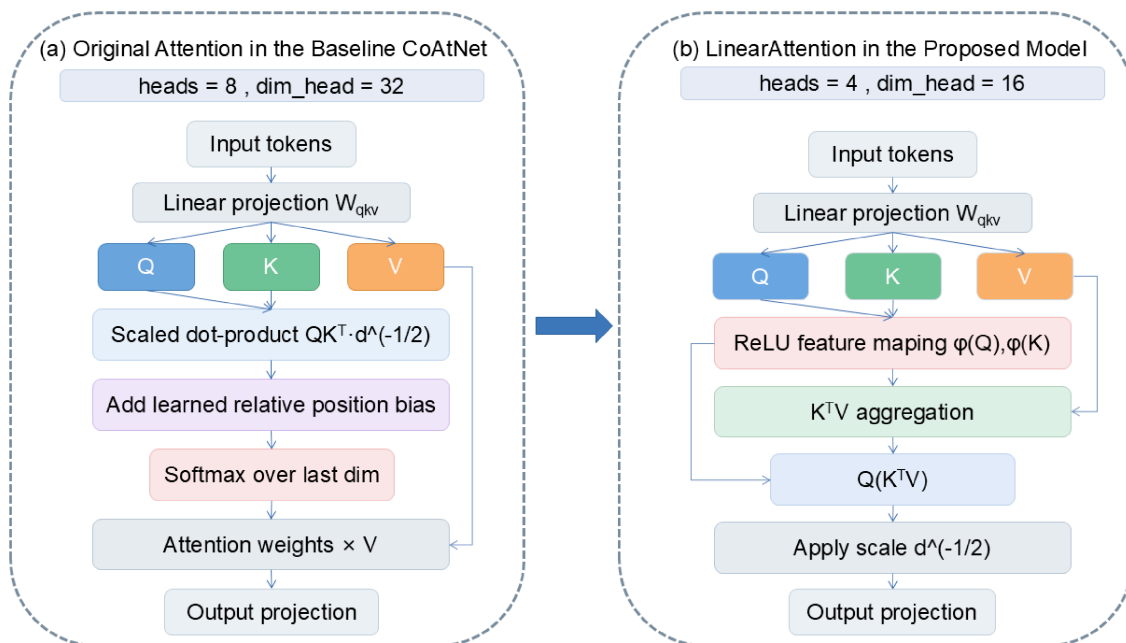


Figure 4. Comparison between the original Softmax-based self-attention in the baseline CoAtNet and the ReLU-based LinearAttention module in the proposed model. (a) Original Softmax-based attention; (b) LinearAttention with ReLU feature mapping.

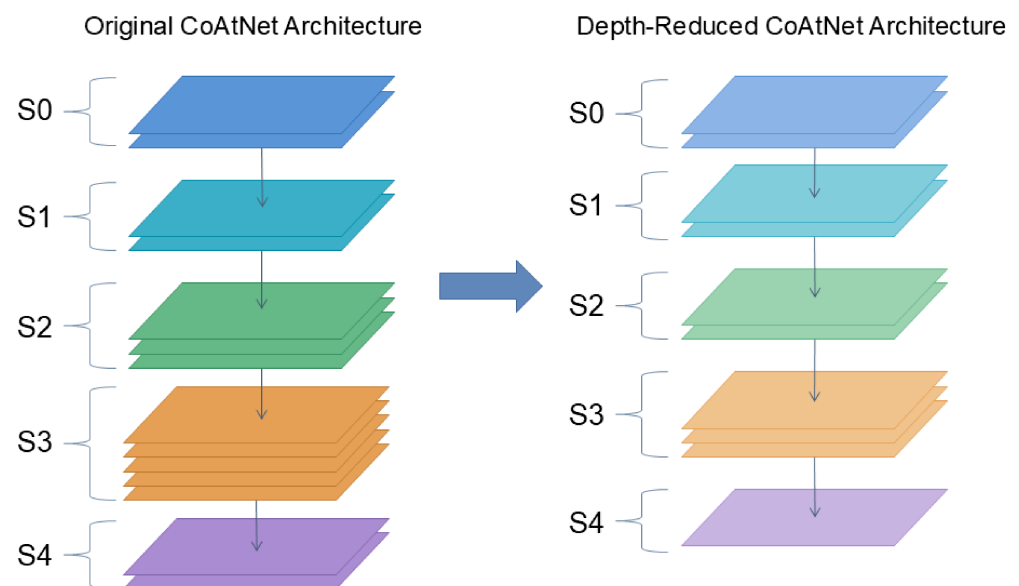


Figure 5. Stage-wise depth configuration of CoAtNet and the proposed LSH-CoAtNet. The number of repeated blocks in stages S0–S4 was reduced from [2, 2, 3, 5, 2] in the original CoAtNet to [2, 2, 2, 3, 1] in LSH-CoAtNet.

2.4. ShuffleMBConv

The improvement of CoAtNet aimed to further reduce computational complexity while maintaining classification accuracy. To this end, this study proposed a ShuffleMBConv module, which integrates the channel split and channel shuffle mechanism of ShuffleNet [41] with the MBConv-style lightweight convolutional block used in CoAtNet. This design is intended to enhance local feature representation and cross-channel information interaction while keeping the computational cost low.

The structure of the ShuffleMBConv module used in this study is shown in Figure 6. Specifically, for an input feature map with a spatial size of $H \times W$, the feature channels are first evenly split into two branches along the channel dimension. The first branch adopts a

lightweight pointwise transformation consisting of 1×1 convolution, batch normalization, and GELU activation, which provides low-cost feature transformation while preserving part of the original information flow. The second branch uses an MBConv-style residual transformation to strengthen local feature extraction. In this branch, a 1×1 convolution is first used for channel projection, followed by batch normalization and GELU activation. A 3×3 depthwise convolution is then applied to capture local spatial information, followed by another batch normalization and GELU activation. Subsequently, an SE module is introduced to recalibrate channel-wise feature responses. Finally, a 1×1 convolution and batch normalization are used to restore the channel representation, and the transformed feature is added to the branch input to form a residual connection. The outputs of the two branches are concatenated along the channel dimension and then processed by a channel shuffle operation to produce the final output.

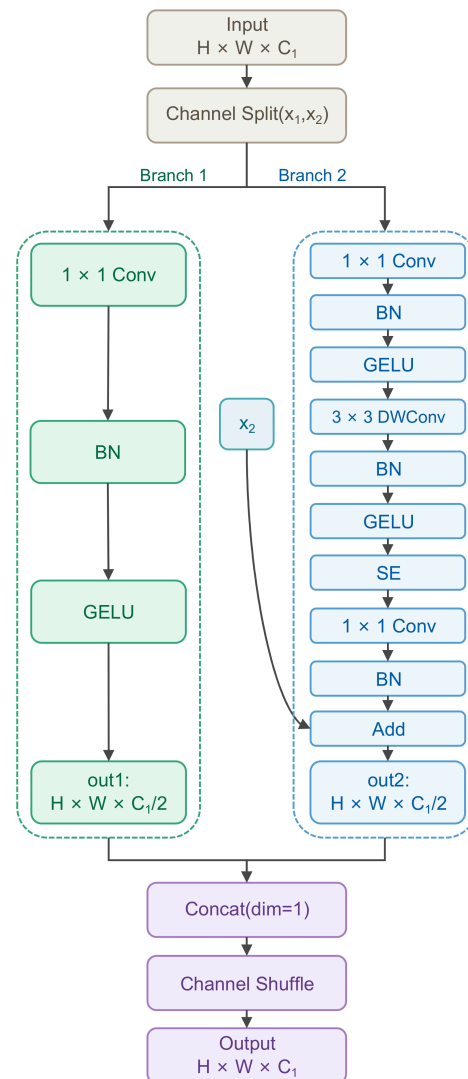


Figure 6. Structure of the proposed ShuffleMBConv module, including channel split, lightweight branch transformation, MBConv-based residual feature extraction, channel concatenation, and channel shuffle.

After the two branches are processed, their outputs are concatenated along the channel dimension. Channel shuffle is then applied to promote information exchange between the two branches and reduce the information isolation caused by fixed channel splitting. Through this design, ShuffleMBConv maintains the spatial resolution of the feature map while improving local representation capability and channel interaction efficiency with relatively low computational overhead.

2.5. HWD ADown

Conventional downsampling methods change feature map size by controlling convolution kernel size and stride. However, images with complex features require more convolution kernels, which reduces efficiency and substantially increases the number of parameters and computational cost [41]. To reduce the spatial dimension of feature maps while preserving key discriminative information, the HWD ADown module was used in this study to replace the downsampling operations in the S1 and S2 layers of the original model.

As shown in Figure 7, the HWD ADown module consists of two parallel branches and a feature fusion layer. The input feature map first undergoes average pooling with a stride of 2 to preliminarily compress the spatial dimension and smooth noise, thereby accommodating slight texture fluctuations in goji berry images. The processed feature map is then divided into two parts along the channel dimension. One branch is processed by the HWD submodule, which applies a one-level Haar wavelet transform to the input feature map to obtain one low-frequency component reflecting the overall contour information of goji berries and three high-frequency components corresponding to detail information in the horizontal, vertical, and diagonal directions. The four components are then concatenated along the channel dimension, and depthwise separable convolution is used to compress the number of channels to the target dimension, achieving feature dimensionality reduction and fusion [27]. The other branch first uses max pooling with a stride of 2 to further compress the spatial dimension and retain locally salient features, and then uses depthwise separable convolution to adjust the number of channels to the target dimension, ensuring consistency with the output dimension of the HWD branch. Finally, the outputs of the two branches are fused through channel concatenation to complete the downsampling process.

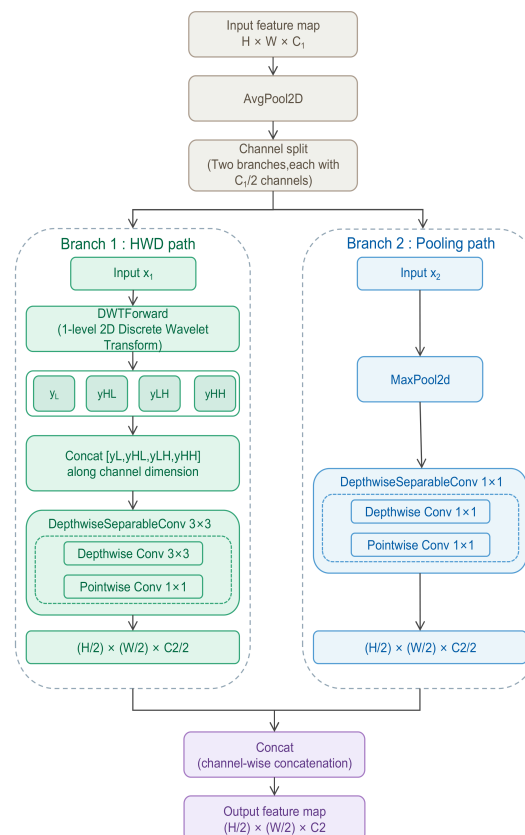


Figure 7. Structure of the proposed HWD ADown module, consisting of an HWD branch with one-level 2D discrete wavelet transform and a pooling branch with max pooling, followed by depthwise separable convolution and feature concatenation.

The HWD ADown module preserves key texture and edge information while reducing image resolution. It is designed to address the subtle differences in color, texture, and contour among commercial dried goji berry samples from five cultivars, improving the retention and recognition of fine-grained features and thereby enhancing classification accuracy.

2.6. Experimental Design and Software Environment

The experimental design of this study consisted of four main parts. First, several representative deep learning models were compared to evaluate their applicability to dried goji berry cultivar identification and to select an appropriate baseline network. Second, the effectiveness of the proposed image quality enhancement method was evaluated by comparing the classification performance of models trained using original images and enhanced images. Third, ablation experiments were conducted to analyze the individual contributions of the proposed lightweight feature extraction module and the information-preserving downsampling module. Finally, five-fold cross-validation was performed to evaluate the stability and reliability of the proposed LSH-CoAtNet model.

The software tools used for image preprocessing, data organization, performance evaluation, model complexity analysis, and result visualization are summarized in Table 2.

Table 2. Software tools and their corresponding versions and purposes used in this study.

Software/Tool	Version	Purpose
Python	3.8.18	Programming language
PyTorch	1.13.1	Model construction, training, and testing
torchvision	0.14.1	Image transformation and model-related utilities
CUDA	11.7	GPU-accelerated computation
OpenCV	4.8.1	Image preprocessing and image quality enhancement
NumPy	1.24.4	Numerical calculation and array processing
pandas	2.0.3	Dataset organization and experimental result recording
scikit-learn	1.3.2	Calculation of accuracy, precision, recall, and F1-score
THOP	0.1.1.post2209072238	Calculation of parameter size and floating-point operations
Matplotlib	3.7.5	Visualization of training curves and experimental results
OriginPro	2025b	Figure preparation and result visualization

3. Results

3.1. Experimental Setup

Models were constructed and improved using the PyTorch deep learning framework in the Anaconda3 environment, and model training and testing were conducted on a GPU under Windows 10. The GPU used was an NVIDIA GeForce RTX 8000.

The batch size was set to 16, the input image size was standardized to 224×224 pixels, the number of training epochs was set to 150, the learning rate was set to 0.00003,

and the random seed was set to 42. The Adam optimizer and cross-entropy loss function were used.

3.2. Evaluation Metrics

Several key evaluation metrics were selected for model assessment, namely accuracy, precision, recall, and F1-score. Their mathematical expressions are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (13)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN} \quad (15)$$

where P denotes precision, R denotes recall, TP (True Positive) denotes the number of samples predicted by the model as the positive class whose actual labels are also positive. TN (True Negative) denotes the number of samples predicted by the model as the negative class whose actual labels are also negative. FP (False Positive) denotes the number of samples predicted by the model as the positive class whose actual labels are negative. FN (False Negative) denotes the number of samples predicted by the model as the negative class whose actual labels are positive.

3.3. Network Model Comparison Experiments

To evaluate the suitability of the selected baseline model, eight deep learning models widely used in image classification in recent years were selected for comparative experiments, as shown in Figure 8. These models included CoAtNet [31], ResNet50 [42], ConvNeXt [43], EfficientNetV2 [44], GhostNet [45], MobileNetV4 [46], RepVGG [47], and ShuffleNetV2 [19]. Among them, CoAtNet achieved the best overall performance, with accuracy, precision, recall, and F1-score values of 94.83%, 94.87%, 94.83%, and 94.83%, respectively.

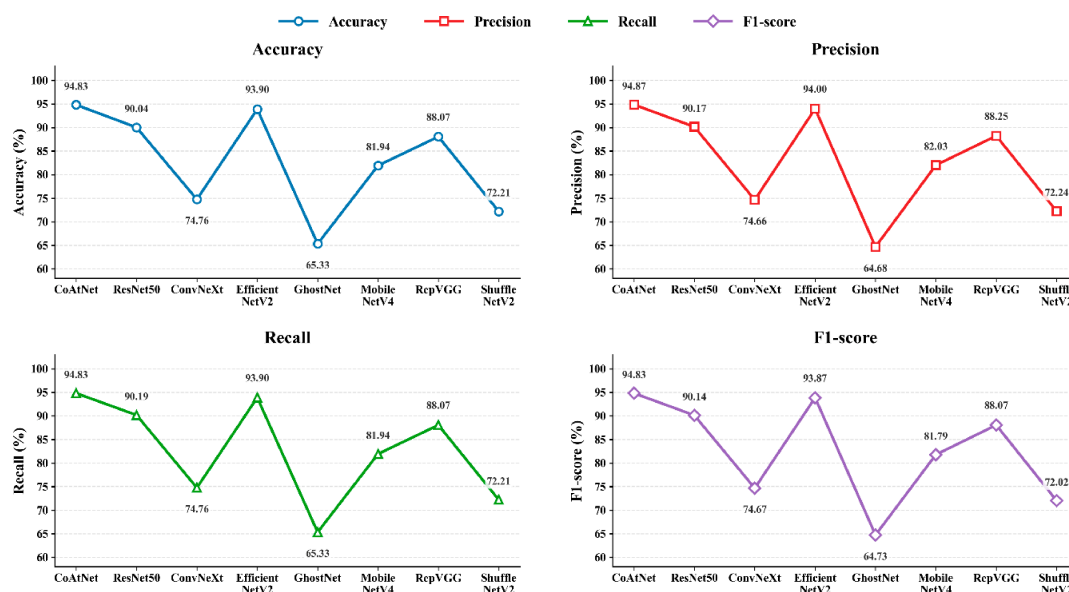


Figure 8. Overall classification performance comparison of different deep learning models on the test set in terms of accuracy, precision, recall, and F1-score.

To intuitively compare the class-level classification performance of different models, normalized confusion matrices for the eight models on the test set were plotted. In a normalized confusion matrix, the main diagonal elements represent the proportion of correctly classified samples in each true class, whereas the off-diagonal elements represent the proportion of samples misclassified into other classes. Therefore, higher normalized values along the main diagonal and lower off-diagonal values indicate stronger class discrimination ability. As shown in Figure 9, CoAtNet exhibited higher normalized values along the main diagonal, indicating good recognition performance for the five classes of commercial dried goji berries. In contrast, ResNet50, ConvNeXt, EfficientNetV2, GhostNet, MobileNetV4, RepVGG, and ShuffleNetV2 showed different degrees of off-diagonal misclassification in some classes, indicating weaker class-level discrimination than CoAtNet.

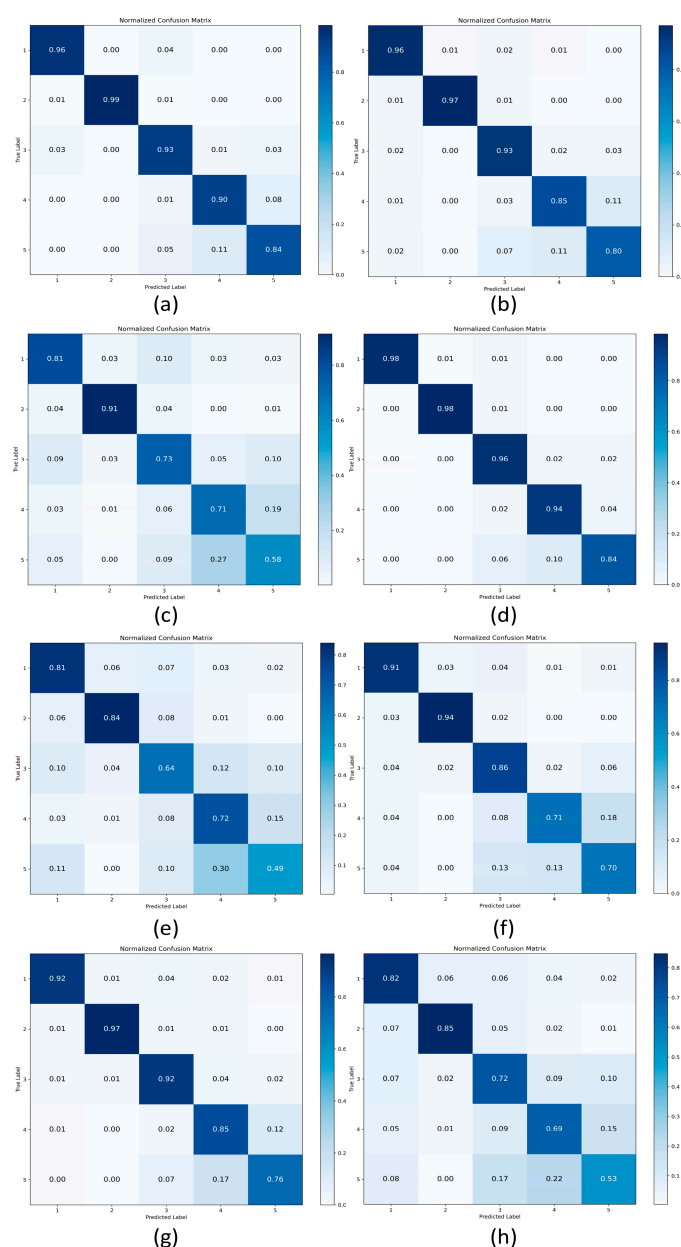


Figure 9. Normalized confusion matrices of different deep learning models on the test set. Panels (a–h) denote CoAtNet, ResNet50, ConvNeXt, EfficientNetV2, GhostNet, MobileNetV4, RepVGG, and ShuffleNetV2, respectively. Classes 1–5 represent Ningqi No. 7, Linqi No. 5, Ningqi No. 1, Keqi 6082, and Jingqi No. 1, respectively.

To further evaluate class-wise classification performance, the precision, recall, and F1-score of each network model across the five classes were calculated, as shown in Table 3. The results showed that CoAtNet achieved the highest values for the largest number of class-wise metrics among the compared models.

Table 3. Class-wise precision, recall, and F1-score of different deep learning models on the test set. All values are expressed as percentages. Bold values indicate the best result for each class and metric. Classes 1–5 correspond to Ningqi No. 7, Linqi No. 5, Ningqi No. 1, Keqi 6082, and Jingqi No. 1, respectively.

Model	Metric	1	2	3	4	5
CoAtNet	Precision (%)	97.99	99.16	94.37	95.19	94.17
	Recall (%)	98.53	99.16	97.01	92.95	93.45
	F1-score (%)	98.26	99.15	95.67	94.05	93.81
ResNet50	Precision (%)	93.90	99.14	86.62	86.72	84.49
	Recall (%)	96.14	97.05	93.01	84.99	79.77
	F1-score (%)	95.00	98.08	89.70	85.84	82.06
ConvNeXt	Precision (%)	80.00	91.93	70.36	68.47	64.06
	Recall (%)	80.88	91.35	73.45	71.07	58.38
	F1-score (%)	80.44	91.64	71.88	69.74	61.09
EfficientNetV2	Precision (%)	99.07	98.73	89.74	89.64	92.95
	Recall (%)	97.61	98.31	96.01	93.85	83.82
	F1-score (%)	98.33	98.52	92.77	91.70	88.15
GhostNet	Precision (%)	74.16	86.15	65.17	62.15	63.25
	Recall (%)	81.25	83.97	63.87	72.15	48.75
	F1-score (%)	77.54	85.04	64.52	66.78	55.06
MobileNetV4	Precision (%)	86.89	94.09	74.61	82.28	72.98
	Recall (%)	91.36	94.09	85.63	70.52	69.75
	F1-score (%)	89.07	94.09	79.74	75.95	71.33
RepVGG	Precision (%)	96.72	97.24	85.82	79.26	83.09
	Recall (%)	92.10	96.62	91.82	84.99	75.72
	F1-score (%)	94.35	96.93	88.72	82.02	79.23
ShuffleNetV2	Precision (%)	76.58	88.74	64.41	67.31	65.17
	Recall (%)	82.35	84.81	72.26	69.26	52.99
	F1-score (%)	79.36	86.73	68.11	68.27	58.45

As shown in Figure 10, CoAtNet and EfficientNetV2 showed relatively high distributions for precision, recall, and F1-score across the five classes. CoAtNet exhibited a narrower distribution range than most of the compared models, indicating smaller class-wise performance variation. In contrast, ConvNeXt, GhostNet, and ShuffleNetV2 showed lower metric distributions and wider ranges.

As shown in Table 4, the compared models exhibited clear differences in both classification performance and computational complexity. Lightweight models such as GhostNet and ShuffleNetV2 had relatively low parameter sizes and GFLOPs, but their classification performance was lower than that of CoAtNet. In contrast, CoAtNet achieved the best overall classification performance among the compared models, but it also required a relatively larger number of parameters and higher computational cost.

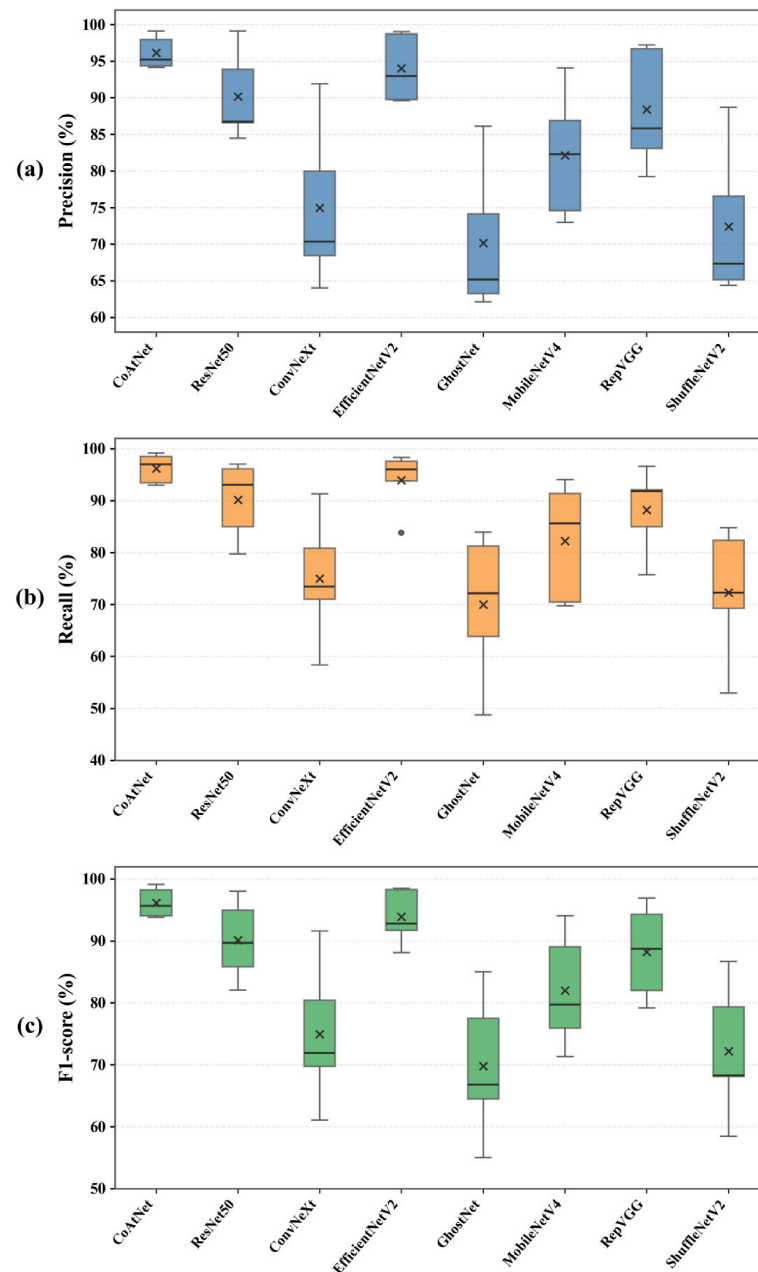


Figure 10. Class-wise distributions of precision, recall, and F1-score for different deep learning models on the test set. (a) Precision; (b) recall; (c) F1-score. Each boxplot summarizes the variation in model performance across the five dried goji berry cultivars.

Table 4. Model complexity comparison of different deep learning models in terms of parameters and GFLOPs.

Model	Params (M)	GFLOPs
CoAtNet	16.99	3.35
ResNet50	25.56	4.13
ConvNeXt	28.59	4.46
EfficientNetV2	20.31	25.54
GhostNet	5.18	0.16
MobileNetV4	9.71	0.91
RepVGG	7.83	1.53
ShuffleNetV2	2.28	0.15
LSH-CoAtNet	6.41	1.6

Based on the above results, CoAtNet was selected as the baseline network for subsequent model improvement. Image quality enhancement and lightweight structural optimization were then used to improve the model's representation capability for fine-grained class-related features while reducing computational overhead.

3.4. Verification of the Effectiveness of the Quality Enhancement Method

To verify the effectiveness of the proposed image quality enhancement method for the classification of commercial dried goji berries, comparative experiments were conducted using CoAtNet as the base model on the original image dataset and the quality-enhanced image dataset.

As shown in Figure 11, the quality-enhanced images showed clearer color gradation, surface texture, and edge contours than the original images. To further evaluate the influence of image quality enhancement on classification performance, the overall performance of CoAtNet on the two datasets was compared, as shown in Figure 12.



Figure 11. Visual comparison of representative dried goji berry images before and after image quality enhancement. The upper row shows the original images, and the lower row shows the corresponding quality-enhanced images for samples 1–6.

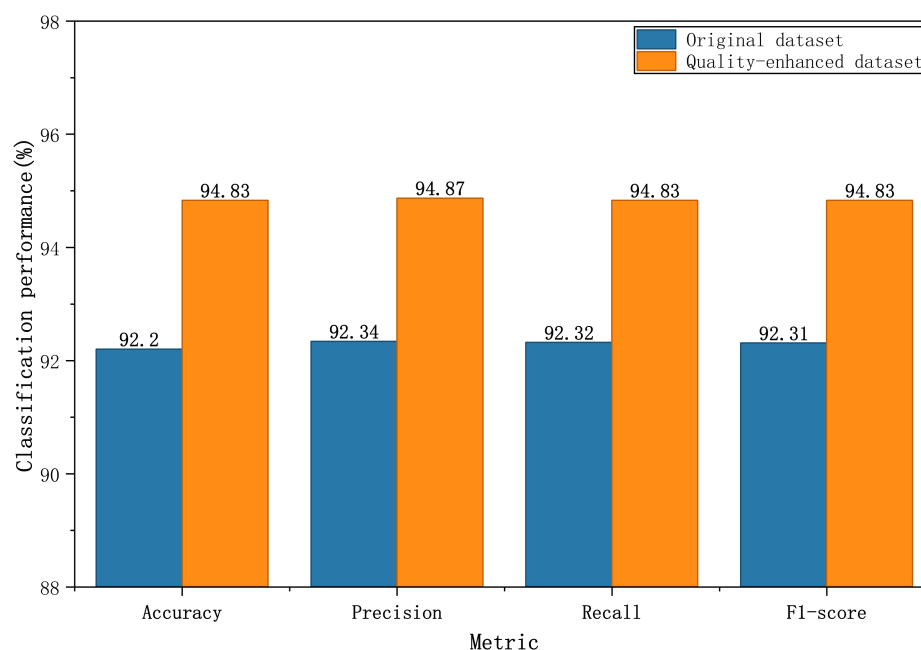


Figure 12. Classification performance comparison between the original dataset and the quality-enhanced dataset in terms of accuracy, precision, recall, and F1-score.

After image quality enhancement, the accuracy, precision, recall, and F1-score of CoAtNet increased from 92.20%, 92.34%, 92.32%, and 92.31% to 94.83%, 94.87%, 94.83%, and 94.83%, respectively. The corresponding improvements were 2.63, 2.53, 2.51, and 2.52%, respectively. These results indicate that the proposed image quality enhancement method improved the classification performance of CoAtNet on dried goji berry cultivar identification.

3.5. Results of Model Improvement

3.5.1. Ablation Study

Ablation experiments are an important means of verifying the effectiveness of each improved module [48]. To systematically analyze the effects of lightweight reconstruction, ShuffleMBConv, and HWD ADown on model performance and efficiency, ablation experiments were performed using CoAtNet as the baseline model. Accuracy, precision, recall, and F1-score were used as performance metrics, while parameter count and GFLOPs were used as efficiency metrics. The experimental results are shown in Table 5.

Table 5. Ablation results of the proposed LSH-CoAtNet components in terms of classification performance and model complexity. LR, SMB, and HAD denote lightweight reconstruction, ShuffleMBConv, and HWD ADown, respectively. A check mark indicates that the corresponding component is included.

Base	LR	SMB	HAD	Acc (%)	Prec (%)	Rec (%)	F1 (%)	Params (M)	GFLOPs
✓				94.83	94.87	94.83	94.83	16.99	3.35
✓	✓			93.67	93.69	93.67	93.66	6.41	1.68
✓		✓		96.41	96.41	96.41	96.41	17.28	3.06
✓			✓	97.18	97.19	97.18	97.18	17.29	3.62
✓	✓	✓		93.21	93.20	93.21	93.19	6.33	1.59
✓	✓		✓	96.99	97.01	96.99	96.99	6.34	1.57
✓		✓	✓	98.84	98.85	98.84	98.84	16.68	3.08
✓	✓	✓	✓	98.80	98.81	98.80	98.80	6.41	1.60

As shown in Table 5, lightweight reconstruction substantially reduced model complexity, although it resulted in a slight decrease in classification performance when used alone. Compared with the baseline CoAtNet, the model with only lightweight reconstruction decreased accuracy from 94.83% to 93.67%, precision from 94.87% to 93.69%, recall from 94.83% to 93.67%, and F1-score from 94.83% to 93.66%. Meanwhile, the number of parameters decreased from 16.99 M to 6.41 M, corresponding to a reduction of approximately 62.27%, and GFLOPs decreased from 3.35 to 1.68, corresponding to a reduction of approximately 49.85%.

When ShuffleMBConv was introduced alone, accuracy increased to 96.41%, which was 1.58% higher than the baseline. Precision, recall, and F1-score also increased to 96.41%, 96.41%, and 96.41%, respectively. The number of parameters was 17.28 M, which was only 0.29 M higher than that of the baseline, while GFLOPs decreased from 3.35 to 3.06.

When HWD ADown was introduced alone, the model achieved higher classification performance than the baseline. Accuracy increased to 97.18%, which was 2.35% higher than the baseline. Precision increased to 97.19%, while recall and F1-score both increased to 97.18%. Under this configuration, the number of parameters was 17.29 M, and GFLOPs increased from 3.35 to 3.62.

When lightweight reconstruction, ShuffleMBConv, and HWD ADown were introduced simultaneously, the model achieved an accuracy of 98.80%, precision of 98.81%, recall of 98.80%, and F1-score of 98.80%. Compared with the baseline CoAtNet, the final model improved accuracy by 3.97% while reducing the number of parameters from 16.99 M to 6.41 M and GFLOPs from 3.35 to 1.60. Compared with the ShuffleMBConv + HWD ADown combination, the final model showed only a slight decrease in accuracy, from 98.84% to 98.80%, but reduced the number of parameters from 16.68 M to 6.41 M and GFLOPs from 3.08 to 1.60.

Overall, the ablation results show that the final LSH-CoAtNet model achieved higher classification performance than the baseline CoAtNet while substantially reducing parameter size and computational complexity.

3.5.2. Comparison Before and After Model Improvement

To evaluate the influence of model structural improvement on classification of commercial dried goji berries, CoAtNet was used as the baseline model and compared with the proposed LSH-CoAtNet. The comparison results are shown in Figure 13. Figure 13a presents the loss and accuracy curves of the two models on the validation set; Figure 13b shows the confusion matrix of LSH-CoAtNet on the test set; Figure 13c presents the precision, recall, and F1-score of the two models across the five classes; and Figure 13d further compares the models before and after improvement in terms of classification performance and model efficiency.

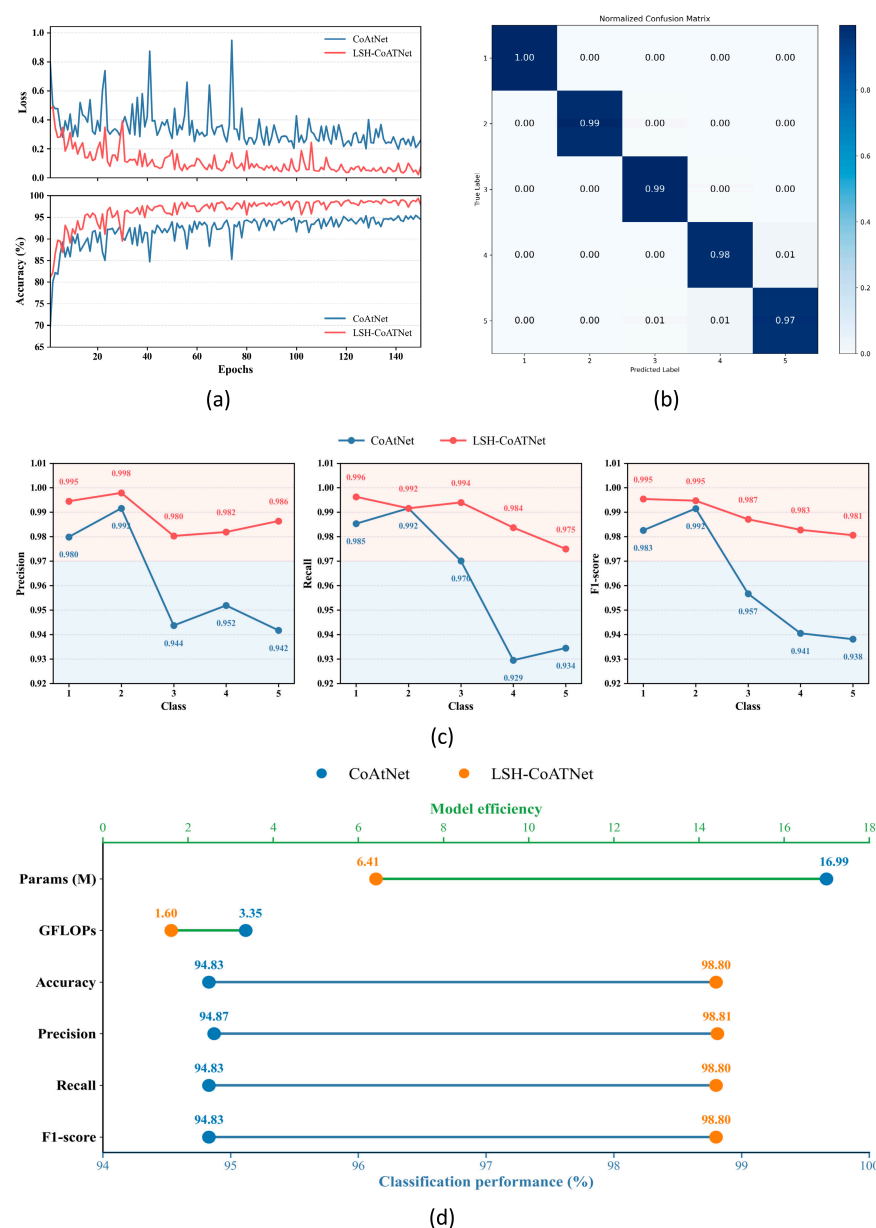


Figure 13. Performance comparison between CoAtNet and LSH-CoAtNet. (a) Validation loss and accuracy curves; (b) normalized confusion matrix of LSH-CoAtNet on the test set; (c) class-wise precision, recall, and F1-score comparison; (d) comparison of classification performance and model complexity in terms of accuracy, precision, recall, F1-score, Params, and GFLOPs. Classes 1–5 represent Ningqi No. 7, Linqi No. 5, Ningqi No. 1, Keqi 6082, and Jingqi No. 1, respectively.

As shown in Figure 13a, LSH-CoAtNet exhibited lower validation loss and higher validation accuracy during training, with relatively smaller fluctuations in the later stage. By contrast, the validation loss of the baseline CoAtNet fluctuated more markedly. Although its accuracy generally showed an upward trend, its stability was inferior to that of LSH-CoAtNet.

The confusion matrix in Figure 13b shows that the predictions of LSH-CoAtNet on the test set were mainly concentrated along the main diagonal. The numbers of correctly identified samples in the five classes were 542, 470, 498, 544, and 506, respectively, and the total number of misclassified samples was small.

Figure 13c further shows the class-wise performance differences between CoAtNet and LSH-CoAtNet. Compared with CoAtNet, LSH-CoAtNet achieved higher precision, recall, and F1-score values across most classes, with smaller class-wise fluctuations. In particular, for Classes 3, 4, and 5, where CoAtNet showed relatively lower performance, the improvement achieved by LSH-CoAtNet was more evident. For example, the F1-scores of LSH-CoAtNet for Classes 4 and 5 reached 98.28% and 98.06%, respectively, whereas those of CoAtNet were 94.05% and 93.81%, respectively.

As shown in Figure 13d, LSH-CoAtNet outperformed CoAtNet in both overall classification performance and model efficiency. Compared with the baseline model, the accuracy, precision, recall, and F1-score of LSH-CoAtNet increased from 94.83%, 94.87%, 94.83%, and 94.83% to 98.80%, 98.81%, 98.80%, and 98.80%, corresponding to increases of 3.97, 3.94, 3.97, and 3.97%, respectively. Meanwhile, the number of parameters decreased from 16.99 M to 6.41 M, and GFLOPs decreased from 3.35 to 1.60, representing reductions of 62.27% and 52.24%, respectively.

Overall, LSH-CoAtNet achieved higher classification performance with lower parameter size and computational complexity than the baseline CoAtNet.

To conduct paired comparisons of the two models and provide statistical evidence, McNemar's test is performed in this paper. For each test image, the prediction outputs of the two models are converted into binary results, i.e., correct classification or incorrect classification. A 2×2 contingency table is constructed based on the paired classification results (correct or incorrect) obtained from CoAtNet and LSH-CoAtNet. The exact McNemar's test is adopted to examine whether the difference in classification accuracy between the two models is statistically significant. The difference is regarded as statistically significant if the p -value is less than 0.05.

As shown in Table 6, 37 samples incorrectly classified by CoAtNet were correctly classified by LSH-CoAtNet, whereas only 3 samples correctly classified by CoAtNet were incorrectly classified by LSH-CoAtNet. The exact McNemar test showed a significant difference between the two models ($p = 1.95 \times 10^{-8}$, $p < 0.001$), indicating that the improvement achieved by LSH-CoAtNet was statistically supported.

Table 6. Paired prediction results and exact McNemar test between CoAtNet and LSH-CoAtNet on the same test set.

Paired Prediction Result	Number of Samples
Both models correctly classified	2550
CoAtNet correct, LSH-CoAtNet incorrect	3
CoAtNet incorrect, LSH-CoAtNet correct	37
Both models incorrectly classified	5
Exact McNemar test p -value	1.95×10^{-8}

3.6. K-Fold Cross-Validation

To further verify the robustness, stability, and generalization ability of the proposed model for classification of commercial dried goji berries, five-fold cross-validation was conducted on the quality-enhanced image dataset. The data were randomly shuffled, and the experimental results were then analyzed comprehensively. Five-fold cross-validation is a widely used and effective validation technique. Specifically, the dataset was randomly shuffled and divided into five mutually exclusive subsets. In each trial, one subset was selected as the validation set and the remaining four subsets were used as the training set. This process was repeated five times, and the results of all folds were statistically analyzed. This method helps minimize bias caused by data partitioning and thereby improves the reliability of model evaluation [49].

As shown in Table 7, the accuracies of the five folds were 98.82%, 99.23%, 99.21%, 99.00%, and 98.63%, respectively. The corresponding precision values were 98.82%, 99.23%, 99.20%, 98.98%, and 98.62%; recall values were 98.81%, 99.21%, 99.20%, 98.99%, and 98.60%; and F1-scores were 98.81%, 99.22%, 99.20%, 98.98%, and 98.61%. Across the five folds, the accuracy, precision, recall, and F1-score were 98.98%, 98.97%, 98.96%, and 98.96%, respectively, with standard deviations of 0.26, 0.26, 0.26, and 0.26%.

Table 7. Five-fold cross-validation results of LSH-CoAtNet in terms of accuracy, precision, recall, and F1-score.

Fold	Acc (%)	Prec (%)	Rec (%)	F1 (%)
1	98.82	98.82	98.81	98.81
2	99.23	99.23	99.21	99.22
3	99.21	99.20	99.20	99.20
4	99.00	98.98	98.99	98.98
5	98.63	98.62	98.60	98.61

The performance variation across the five folds was small. The accuracy ranged from 98.63% to 99.23%, with a range of 0.60%. Similar variation ranges were observed for precision, recall, and F1-score. These results indicate that LSH-CoAtNet maintained stable classification performance across different data partitions.

3.7. Edge-Device Inference Efficiency

To further evaluate the inference efficiency of the proposed model on resource-constrained hardware, CoAtNet and LSH-CoAtNet were tested on an NVIDIA Jetson Orin Nano platform. The trained PyTorch models were directly transferred to the edge device for inference testing, and no ONNX or TensorRT acceleration was used in this experiment. All input images were resized to 224×224 pixels, consistent with the input size used in the main experiments. The inference speed was evaluated using frames per second (FPS) and average latency under batch size 1 and batch size 16. The batch size 1 setting was used to approximate single-image online inference, whereas the batch size 16 setting was used to evaluate mini-batch inference throughput. A higher FPS and a lower latency indicate faster inference performance.

As shown in Table 8, LSH-CoAtNet achieved faster inference speed and lower latency than the baseline CoAtNet on the Jetson Orin Nano platform. Under batch size 1, LSH-CoAtNet achieved 50.18 FPS, which was higher than the 35.10 FPS of CoAtNet. The corresponding average latency decreased from 28.49 ms/image to 19.93 ms/image. Under batch size 16, LSH-CoAtNet achieved 98.31 FPS, whereas CoAtNet achieved 61.25 FPS. The average latency decreased from 16.33 ms/image to 10.17 ms/image.

Table 8. Inference efficiency of CoAtNet and LSH-CoAtNet on the Jetson Orin Nano platform in terms of FPS and latency under batch sizes of 1 and 16.

Model	FPS Batch = 1	Latency Batch = 1 (ms/Image)	FPS Batch = 16	Latency Batch = 16 (ms/Image)
CoAtNet	35.10	28.49	61.25	16.33
LSH-CoAtNet	50.18	19.93	98.31	10.17

Compared with CoAtNet, LSH-CoAtNet increased FPS by 42.96% and reduced average latency by 30.05% under batch size 1. Under batch size 16, FPS increased by 60.51%, and average latency decreased by 37.72%. These results indicate that the lightweight structural design of LSH-CoAtNet improved edge-device inference efficiency compared with the baseline CoAtNet.

3.8. Robustness Evaluation Under Perturbed Imaging Conditions

To further evaluate the robustness of the LSH-CoAtNet model, a perturbed test dataset was constructed based on the original test set in this work. Four categories of image perturbations were adopted for the experiment, including illumination variation, Gaussian noise, motion blur, and partial occlusion.

As shown in Table 9, LSH-CoAtNet achieved an accuracy of 92.92% on the perturbed test set. The macro-averaged precision, recall, and F1-score were 93.13%, 92.96%, and 92.89%, respectively, while the weighted precision, recall, and F1-score were 93.30%, 92.92%, and 92.96%, respectively. Although the classification performance decreased under perturbed imaging conditions compared with the original test condition, the model still maintained an accuracy above 90%. These results indicate that LSH-CoAtNet retained a certain degree of robustness under simulated illumination variation, image noise, motion blur, and partial occlusion conditions.

Table 9. Robustness performance of LSH-CoAtNet on the perturbed test set in terms of accuracy, macro-average metrics, and weighted-average metrics.

Metric	Value (%)
Acc	92.92
Macro Prec	93.13
Macro Rec	92.96
Macro F1	92.89
Weighted Prec	93.30
Weighted Rec	92.92
Weighted F1	92.96

These results indicate that LSH-CoAtNet retained a certain degree of robustness under simulated perturbation conditions such as illumination fluctuation, image noise, motion blur, and partial occlusion.

4. Discussion

The results of this study support the feasibility of RGB image-based cultivar identification for commercial dried goji berries. This study investigated whether cultivar-related visual differences in dried goji berries could be better represented by enhancing color, texture, and contour information in RGB images, and whether a lightweight CoAtNet-based model could achieve accurate identification while reducing computational cost. The experimental results showed that image quality enhancement improved the classification performance of the baseline CoAtNet, and the proposed LSH-CoAtNet further improved recognition accuracy while substantially reducing parameter size and GFLOPs. These find-

ings are consistent with previous studies showing that RGB machine vision has practical value in nondestructive recognition and quality inspection of agricultural products [13].

From a visual-feature perspective, dried goji berry cultivar identification can be regarded as a fine-grained visual classification problem. After drying, different cultivars may show similar overall appearance, and the discriminative information is mainly reflected in subtle differences in color gradation, surface wrinkle texture, shrinkage morphology, local brightness distribution, and contour characteristics. Therefore, directly using original RGB images may not fully highlight cultivar-related visual cues. The proposed image quality enhancement workflow, including color space conversion, wavelet enhancement, 3D-LUT color fine-tuning, colorfulness control, and local contrast enhancement, was designed to improve the visibility of color, texture, and edge-related information. This design is consistent with the principles of fine-grained image recognition, colorfulness measurement, and local adaptive contrast enhancement [14–17].

From the perspective of model design, the ablation results indicate that lightweight reconstruction, ShuffleMBConv, and HWD ADown contributed differently to the final model. Lightweight reconstruction substantially reduced parameter size and computational cost, but caused a slight decrease in accuracy when used alone. The introduction of ShuffleMBConv and HWD ADown compensated for this performance loss and improved the classification performance of the lightweight model. This suggests that, for fine-grained cultivar identification of dried goji berries, model performance depends not only on network depth or parameter size, but also on whether local texture, edge contour, and color-distribution information can be effectively preserved during feature extraction and downsampling. The McNemar test further showed that the difference in paired classification correctness between CoAtNet and LSH-CoAtNet was statistically significant, providing statistical support for the improvement achieved by the proposed model. The lightweight design of LSH-CoAtNet is also consistent with efficient network design principles that emphasize the balance between recognition performance and computational cost [18,19].

The perturbation-based robustness evaluation further showed that LSH-CoAtNet maintained an accuracy of 92.92% under simulated non-ideal imaging conditions, including illumination variation, Gaussian noise, motion blur, and partial occlusion. This result suggests that the proposed model retained a certain degree of robustness to common image disturbances. In addition, several comparison models showed relatively higher off-diagonal values for visually similar cultivars, such as Keqi 6082 and Jingqi No. 1. This may be related to the similarity of dried goji berry samples in red-orange color distribution, surface wrinkle texture, shrinkage morphology, and contour appearance. In contrast, LSH-CoAtNet substantially alleviated this class-level confusion, suggesting that the proposed model improved the representation of local texture and contour-related features.

Compared with near-infrared spectroscopy, hyperspectral imaging, and multimodal detection methods previously used for geographical traceability, quality evaluation, and classification of goji berry and related *Lycium* products [8,11,12], the RGB image-based method used in this study has lower equipment cost, simpler image acquisition procedures, and better potential for integration into low-cost visual inspection systems. However, spectral methods can provide richer chemical and spectral information, whereas RGB images mainly capture external visual characteristics. Therefore, the proposed RGB-based method should be regarded as a practical and cost-effective visual recognition approach rather than a complete replacement for spectral or physicochemical analysis. In this sense, LSH-CoAtNet further emphasizes the balance between fine-grained cultivar recognition and computational efficiency, which is important for potential application in mobile terminals, portable detection devices, and intelligent sorting systems.

Although this study achieved promising experimental results, several limitations remain. First, only five commercial dried goji berry cultivars were included, and the cultivar coverage remains limited. Future studies should include more cultivars, geographical origins, processing batches, drying degrees, and storage conditions to improve the generalization ability of the model. Second, the images were acquired mainly under controlled indoor conditions. Although the perturbation-based robustness experiment simulated illumination variation, noise, motion blur, and partial occlusion, further validation is still required under real sorting-line conditions, including complex backgrounds, fruit overlap, dynamic illumination, different imaging angles, and different imaging devices. Third, although preliminary offline edge-device inference testing was conducted on the Jetson Orin Nano platform, the current deployment experiment was still performed under a static image testing setting. The overall speed and stability of a real sorting system may also be affected by image acquisition, image transfer, preprocessing, continuous sample movement, dynamic illumination, and hardware integration. Therefore, further validation using actual commercial sorting equipment should be conducted to evaluate real-time inference efficiency and engineering adaptability under continuous sorting-line conditions. Finally, this study was based only on RGB image information and did not incorporate near-infrared, hyperspectral, or physicochemical indicators. Future work may explore multimodal fusion strategies to improve the comprehensive representation of cultivar differences.

5. Conclusions

To meet the demand for rapid, low-cost, and nondestructive cultivar identification of commercial dried goji berries, this study constructed an RGB image dataset comprising commercial dried goji berry samples from five cultivars and systematically compared the performance of several representative deep learning models in the fine-grained classification of dried goji berries. The experimental results showed that CoAtNet had clear overall advantages in recognition accuracy and feature representation capability. Therefore, CoAtNet was used as the baseline model, and LSH-CoAtNet, a lightweight network model suitable for cultivar identification of commercial dried goji berries, was further proposed.

To address the pronounced surface shrinkage, complex texture, subtle inter-cultivar appearance differences, and susceptibility of color and texture features to imaging conditions in dried goji berries, an image quality enhancement method was introduced to improve the representation of surface color gradation, textural details, and edge information. In terms of model structure, lightweight reconstruction was used to reduce the number of parameters and computational complexity; the ShuffleMBConv module was introduced to enhance local feature extraction and channel information interaction; and the HWD ADown module was adopted to preserve more texture, contour, and other fine-grained features related to cultivar identification during downsampling. Experimental results showed that LSH-CoAtNet achieved a classification accuracy of 98.80% on the quality-enhanced image dataset while requiring only 6.41 M parameters and 1.60 GFLOPs, achieving a favorable balance between recognition performance and model compactness.

In summary, the method proposed in this study, which combines RGB image enhancement with LSH-CoAtNet, provides a low-cost, nondestructive, and efficient visual detection solution for cultivar identification, raw material sorting, batch consistency assessment, and quality control of commercial dried goji berries during processing and distribution. This method may also provide a reference for intelligent classification, commercial sorting, and quality inspection of other specialty dried horticultural products.

Author Contributions: Conceptualization, Z.L., L.S. and H.Y.; Data curation, Z.L.; Formal analysis, Z.L.; Funding acquisition, L.S.; Investigation, Z.L., Y.L., Z.C., C.Q., J.G. and Z.B.; Methodology, Z.L.; Project administration, L.S. and H.Y.; Resources, L.S. and H.Y.; Software, Z.L.; Supervision, L.S. and

H.Y.; Validation, Z.L., Y.L., J.G., Z.C., C.Q. and Z.B.; Visualization, Z.L.; Writing—original draft, Z.L.; Writing—review and editing, L.S. and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Development Plan Project of Jilin Province, grant number 20250201059GX.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to ongoing research and planned further analyses and related publications based on the same dataset.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3D-LUT	Three-dimensional lookup table
CNN	Convolutional neural network
CoAtNet	Convolution and attention network
DWT	Discrete wavelet transform
F1-score	Harmonic mean of precision and recall
GFLOPs	Giga floating-point operations
GPU	Graphics processing unit
HSV	Hue, saturation, value
HWD	Haar wavelet downsampling
HWD ADown	Haar wavelet downsampling adaptive downsampling
Lab	CIELAB color space
LR	Lightweight reconstruction
LSH-CoAtNet	Lightweight CoAtNet incorporating ShuffleMBConv and HWD ADown modules
LUT	Lookup table
MBConv	Mobile inverted bottleneck convolution
RGB	Red, green, and blue
ShuffleMBConv	Shuffle mobile inverted bottleneck convolution
SMB	ShuffleMBConv

References

1. Shunkai, H.; Yiting, X.; Shadrack, S.M.; Jianghao, Z.; Lingxiao, X.; Yezhi, W.; Fei, W.; Chongjiang, C.; Xiao, X.; Biao, Y. *Lycium barbarum* (goji berry): A comprehensive review of chemical composition, bioactive compounds, health-promoting activities, and applications in functional foods and beyond. *Food Chem.* **2025**, *496*, 146588. [[CrossRef](#)] [[PubMed](#)]
2. Noreen, S.; Niazi, M.K.; Arshad, M.T.; Ikram, A.; Gnedeka, K.T. Goji berries: A review of their bioactive components and health-promoting properties. *Int. J. Food Sci. Technol.* **2025**, *60*, vvaf232. [[CrossRef](#)]
3. Solomando González, J.C.; Rodríguez Gómez, M.J.; Ramos García, M.; Nicolás Barroso, N.; Calvo Magro, P. Characterization and selection of *Lycium barbarum* cultivars based on physicochemical, bioactive, and aromatic properties. *Horticulturae* **2025**, *11*, 924. [[CrossRef](#)]
4. Huang, T.; Jia, N.; Zhu, L.; Jiang, W.; Tu, A.; Qin, K.; Yuan, X.; Li, J. Comparison of phenotypic and phytochemical profiles of 20 *Lycium barbarum* L. goji berry varieties during hot air-drying. *Food Chem. X* **2025**, *27*, 102436. [[CrossRef](#)] [[PubMed](#)]
5. Ju, Y.; Liu, H.; Niu, S.; Kang, L.; Ma, L.; Li, A.; Zhao, Y.; Yuan, Y.; Zhao, D. Optimizing geographical traceability models of Chinese *Lycium barbarum*: Investigating effects of region, cultivar, and harvest year on nutrients, bioactives, elements and stable isotope composition. *Food Chem.* **2025**, *467*, 142286. [[CrossRef](#)] [[PubMed](#)]
6. Patel, K.K.; Kar, A.; Jha, S.; Khan, M. Machine vision system: A tool for quality inspection of food and agricultural products. *J. Food Sci. Technol.* **2012**, *49*, 123–141. [[PubMed](#)]
7. Sabzi, S.; Nadimi, M.; Abbaspour-Gilandeh, Y.; Paliwal, J. Non-destructive estimation of physicochemical properties and detection of ripeness level of apples using machine vision. *Int. J. Fruit Sci.* **2022**, *22*, 628–645. [[CrossRef](#)]

8. Cui, J.; Li, K.; Hao, J.; Dong, F.; Wang, S.; Rodas-González, A.; Zhang, Z.; Li, H.; Wu, K. Identification of near geographical origin of wolfberries by a combination of hyperspectral imaging and multi-task residual fully convolutional network. *Foods* **2022**, *11*, 1936. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Mu, Q.; Kang, Z.; Guo, Y.; Chen, L.; Wang, S.; Zhao, Y. Hyperspectral image classification of wolfberry with different geographical origins based on three-dimensional convolutional neural network. *Int. J. Food Prop.* **2021**, *24*, 1705–1721. [\[CrossRef\]](#)
10. Yahui, L.; Xiaobo, Z.; Tingting, S.; Jiyong, S.; Jiewen, Z.; Holmes, M. Determination of geographical origin and anthocyanin content of black goji berry (*Lycium ruthenicum* Murr.) using near-infrared spectroscopy and chemometrics. *Food Anal. Methods* **2017**, *10*, 1034–1044.
11. He, C.; Shi, X.; Lin, H.; Li, Q.; Xia, F.; Shen, G.; Feng, J. The combination of HSI and NMR techniques with deep learning for identification of geographical origin and GI markers of *Lycium barbarum* L. *Food Chem.* **2024**, *461*, 140903. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Li, B.; Xia, R.; Li, J.; Zhang, J.; Zhang, Z.; Chen, J.; Chen, Y. Multimodal deep learning with hyperspectral imaging for accurate origin classification of wolfberries. *Food Chem. X* **2025**, *31*, 103166. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Lv, X.; Zhang, X.; Gao, H.; He, T.; Lv, Z.; Zhangzhong, L. When crops meet machine vision: A review and development framework for a low-cost nondestructive online monitoring technology in agricultural production. *Agric. Commun.* **2024**, *2*, 100029. [\[CrossRef\]](#)
14. Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; Belongie, S. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8927–8948. [\[CrossRef\]](#)
15. Yang, G.; He, Y.; Yang, Y.; Xu, B. Fine-grained image classification for crop disease based on attention mechanism. *Front. Plant Sci.* **2020**, *11*, 600854. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Hasler, D.; Suesstrunk, S.E. Measuring colorfulness in natural images. In *Proceedings of the Human Vision and Electronic Imaging VIII*; SPIE: Bellingham, WA, USA, 2003; pp. 87–95.
17. Dijk, J.; den Hollander, R.J.; Schavemaker, J.G.; Schutte, K. Local adaptive contrast enhancement for color images. In *Proceedings of the Visual Information Processing XVI*; SPIE: Bellingham, WA, USA, 2007; pp. 81–88.
18. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2019; pp. 1314–1324.
19. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 116–131.
20. Sun, H.; Wang, R.-F. BMDNet-YOLO: A Lightweight and Robust Model for High-Precision Real-Time Recognition of Blueberry Maturity. *Horticulturae* **2025**, *11*, 1202. [\[CrossRef\]](#)
21. Sun, H.; Xi, X.; Wu, A.-Q.; Wang, R.-F. ToRLNet: A Lightweight Deep Learning Model for Tomato Detection and Quality Assessment Across Ripeness Stages. *Horticulturae* **2025**, *11*, 1334. [\[CrossRef\]](#)
22. Wang, S.; Jiang, H.; Yang, J.; Ma, X.; Chen, J.; Li, Z.; Tang, X. Lightweight tomato ripeness detection algorithm based on the improved RT-DETR. *Front. Plant Sci.* **2024**, *15*, 1415297. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Otsu, N. A threshold selection method from gray-level histograms. *Automatica* **1975**, *11*, 285–296.
24. Canny, J. A computation approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 670–700.
25. Yang, S.; Xiao, W.; Zhang, M.; Guo, S.; Zhao, J.; Shen, F. Image data augmentation for deep learning: A survey. *arXiv* **2022**, arXiv:2204.08610.
26. Schwiegerling, J. *Field Guide to Visual and Ophthalmic Optics*; SPIE: Bellingham, WA, USA, 2004.
27. Li, Q.; Shen, L.; Guo, S.; Lai, Z. Wavelet integrated CNNs for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 7245–7254.
28. Zeng, H.; Cai, J.; Li, L.; Cao, Z.; Zhang, L. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2058–2073.
29. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **1987**, *39*, 355–368. [\[CrossRef\]](#)
30. Chang, D.-C.; Wu, W.-R. Image contrast enhancement based on a histogram transformation of local standard deviation. *IEEE Trans. Med. Imaging* **1998**, *17*, 518–531. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977. [\[CrossRef\]](#)
32. Yang, T.-J.; Howard, A.; Chen, B.; Zhang, X.; Go, A.; Sandler, M.; Sze, V.; Adam, H. Netadapt: Platform-aware neural network adaptation for mobile applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 285–300.
33. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 12934–12949. [\[CrossRef\]](#)
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

35. Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; Li, H. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 3531–3539.
36. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv* **2015**, arXiv:1510.00149.
37. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 464–468.
38. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2020; pp. 5156–5165.
39. Liu, Z.; Sun, M.; Zhou, T.; Huang, G.; Darrell, T. Rethinking the value of network pruning. *arXiv* **2018**, arXiv:1810.05270.
40. Liu, D.; Zhao, L.; Wang, Y.; Kato, J. Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognit.* **2023**, *140*, 109550. [[CrossRef](#)]
41. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 6848–6856.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778.
43. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 11976–11986.
44. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2021; pp. 10096–10106.
45. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 1580–1589.
46. Qin, D.; Lechner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B. MobileNetV4: Universal models for the mobile ecosystem. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 78–96.
47. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 13733–13742.
48. Lin, X.; Liao, D.; Du, Z.; Wen, B.; Wu, Z.; Tu, X.; Zhang, Y. An Improved Lightweight ConvNeXt for Peach Ripeness Classification. *Horticulturae* **2026**, *12*, 134. [[CrossRef](#)]
49. Chen, Z.; Yu, H.; Song, S.; Bi, C.; Guo, J.; Ling, X. J-Rice-ResNeXt: A deep learning-enhanced framework for high-accuracy Japonica rice varietal classification in precision agriculture. *Ind. Crops Prod.* **2025**, *237*, 122197. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.