Hindawi
Journal of Food Quality
Volume 2024, Article ID 6439900, 13 pages
https://doi.org/10.1155/2024/6439900

Research Article

CNFA: ConvNeXt Fusion Attention Module for Age Recognition of the Tangerine Peel

Fuqin Deng,[1,2,3] Junwei Li,[1] Lanhui Fu,[1] Chuanbo Qin,[1] Yikui Zhai,[1] Hongmin Wang,[1] Ningbo Yi,[4] Nannan Li,[5] and TinLun Lam[2,3]

[1]*School of Electronic and Information Engineering, Wuyi University, Jiangmen 529020, China*
[2]*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China*
[3]*The Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen 518100, China*
[4]*School of Textile Materials and Engineering, Wuyi University, Jiangmen 529020, China*
[5]*School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau 999078, China*

Correspondence should be addressed to Lanhui Fu; j002886@wyu.edu.cn and TinLun Lam; tllam@cuhk.edu.cn

Received 6 February 2024; Revised 22 April 2024; Accepted 2 May 2024; Published 14 May 2024

Academic Editor: Daniel Cozzolino

Xinhui tangerine peel has valuable medicinal value. The longer it is stored in an appropriate environment, the higher its flavonoid content, resulting in increased medicinal value. In order to correctly identify the age of the tangerine peel, previous studies have mostly used manual identification or physical and chemical analysis, which is a tedious and costly process. This work investigates the automatic age recognition of the tangerine peel based on deep learning and attention mechanisms. We proposed an effective ConvNeXt fusion attention module (CNFA), which consists of three parts, a ConvNeXt block for extracting low-level features' information and aggregating hierarchical features, a channel squeeze-and-excitation (cSE) block and a spatial squeeze-and-excitation (sSE) block for generating sufficient high-level feature information from both channel and spatial dimensions. To analyze the features of tangerine peel in different ages and evaluate the performance of CNFA module, we conducted comparative experiments using the CNFA-integrated network on the Xinhui tangerine peel dataset. The proposed algorithm is compared with related models of the proposed structure and other attention mechanisms. The experimental results showed that the proposed algorithm had an accuracy of 97.17%, precision of 96.18%, recall of 96.09%, and F1 score of 96.13% for age recognition of the tangerine peel, providing a visual solution for the intelligent development of the tangerine peel industry.

1. Introduction

Tangerine peel, derived from Citrus reticulata Blanco, is an agricultural product made from citrus peels that have been dried or dried for storage [1]. Tangerine peel from Xinhui (Jiangmen City, Guangdong Province, China) holds valuable economic value because of its geographical advantage, climatic advantage, and the unique production techniques of tangerine peel [2]. The value of the tangerine peel industry chain was 19 billion CNY in 2022, accounting for 20% of the total GDP of Xinhui. Xinhui tangerine peel is considered to be the quality and is rich in the flavonoid [3]. The flavonoid has great effects in anti-inflammatory, antiviral, and anti-atherosclerosis [4]. As shown in Table 1, as the storage time of tangerine peel increases, the higher the flavonoid contained in the tangerine peel [3], the higher the medicinal value. No relevant literature is found for 20-year tangerine peel. The age of the tangerine peel is one of the important criteria for measuring the quality of tangerine peel.

As the years increase, market value tends to increase. The prices of tangerine peel are shown in Table 1, with its market price increasing exponentially as the age increases. During the recovery stage of Covid-19 pandemic, tangerine peel had a mitigating effect on symptoms [5]. However, many

Table 1: The total flavonoid content percentage and the price of the tangerine peel in different ages.

Age (years)	Total flavonoid content percentage (%)	Price (CNY/kilogram)
1	4.97	540
5	6.17	1,300
10	6.44	1,960
15	6.91	3,360
>20	—	11,960

merchants have taken advantage of this gimmick by using young tangerine peel for craft processing to pass off as old tangerine peel as a way to gain more economic benefits, which not only harms the interests of consumers but also disrupts the market regulations to a certain extent [6]. Therefore, it is essential to develop a method that can flexibly, accurately identify the age of the tangerine peel.

Deep learning has been widely used in foods and agriculture in recent years [7, 8], and image recognition of plants has received a lot of attention from researchers. Following this trend, nondestructive age recognition of the tangerine peel contributes to the development of the intelligent tangerine peel industry. However, it still faces challenges as tangerine peels lack distinct shape differences and have similar colors. Therefore, the feature extraction of tangerine peels becomes more complex, leading to greater difficulty in recognition. Age recognition of the tangerine peel requires special attention to features such as oil bags, patterns, and color on the epidermis of the peel. Existing deep learning models struggle to capture these fine-grained details, and attention mechanisms are commonly used techniques to focus on such detailed features.

To effectively extract important features of tangerine peel, we designed a ConvNeXt fusion attention module (CNFA module) that uses a strategy to aggregate feature information extracted by ConvNeXt block and attention mechanisms. A high-level feature contains rich semantic information, which can be used for the localization of the tangerine peel. A low-level feature plays an important role in capturing crucial details of tangerine peel during feature extraction. In the CNFA module, the ConvNeXt block can effectively extract low-level feature information of images and aggregate hierarchical features. In addition, the cSE and sSE capture effective channel and spatial information adaptively in the image, including the local detailed feature of tangerine peel, and assign different attention weights to features of tangerine peel from different locations. The high-level feature generated by the attention module is utilized to guide the ConNeXt block in extracting the low-level feature. The CNFA module combines the obtained low-level feature information and high-level feature information, linking feature information to effectively extract features of tangerine peel images. We embedded the CNFA module into our network architecture, effectively extracting global contextual information and suppressing useless information. The main contributions of this work are as follows:

(1) We proposed a CNFA attention module aggregating low-level and high-level features in the network to improve the detection accuracy

(2) We validated the effectiveness of the CNFA module compared to other attention mechanisms through comparative experiments

The rest of this article is structured as follows. The second part reviews the related work. The third part introduces the network and implementation of age recognition of the tangerine peel. The fourth part introduces the experimental results and discussion. The fifth part includes the conclusion.

2. Related Work

The shapes and colors of tangerine peel are in different ages, which can be quite similar. It is difficult to recognize the age of tangerine peel for ordinary people. The main methods for identifying the age of the tangerine peel are the manual identification method and physical and chemical analysis method [9]. The former relies on experienced personnel to identify different ages of tangerine peel based on differences in color, shape, and odour. This is simple to operate, but it is susceptible to interference from subjective conditions and objective factors. The latter is mainly judged by detecting the content of components in the tangerine peel. Chen et al. used response surface methodology to optimize the process of microwave-assisted extraction of pectin polysaccharides from tangerine peel [10]. This method can analyze the age in terms of its material composition. Li et al. proposed a method to estimate the age of tangerine peel based on the trnL-trnF copy number [11]. The study explored the correlation between six DNA fragments and the age of tangerine peel. It was found that the trnL-trnF copy number showed a negative correlation with the age of the tangerine peel. Yue et al. extracted tangerine peel polysaccharides from five different-age tangerine peels and proposed the relationship between tangerine peel polysaccharides and their ages [12]. But these are tedious processes and destroy the sample, affecting secondary sales.

Age recognition of the tangerine peel is a novel research direction. Pan et al. used a handheld near-infrared spectrometer to scan the epidermis of tangerine peel and collected corresponding near-infrared diffuse reflection spectra [6]. After preprocessing, the data were used to identify the origin and age of tangerine peel using random forest, K-nearest neighbor, and linear discriminant analysis. Zhang et al. proposed a novel approach that combines near-infrared spectroscopy with machine learning to identify the age of the tangerine peel [13]. The method involves preprocessing the spectral data through Savitzky–Golay convolution smoothing, standard normal variate first-order derivatives, and principal component analysis (PCA) to yield characteristic spectral variables. The support vector machine (SVM) and K-nearest neighbor algorithms are employed for discrimination then. Pu et al. proposed a method for identifying the origins of tangerine peel using terahertz time-domain spectroscopy combined with CNN (convolutional neural network) [14]. Different spectral data were used to

construct 1D CNN and 2D CNN models. Additionally, an Add-CNN model was developed by combining both spectral and image data. However, these works require specialized equipment that lacks flexibility.

Deep learning [15] is a machine learning technique in which machines simulate the human brain to analyze data, with the computer vision [16] being one of the more prominent applications. In the past few years, the image classification of agricultural products represented by tangerine peel is emerging. Chu et al. introduced a method to increase the data volume of tangerine peel by utilizing traditional data augmentation, deep convolution generative adversarial networks (DCGAN), and Mosaic [17]. The data volume of the original sample was increased by 23 times. They also used the CBAM module in conjunction with CSPNet to extract the endocarp features and classify, which can effectively extract feature information such as color, size, and shape on the endocarp of tangerine peel. However, they ignored the low-level feature information on the epidermis of tangerine peel, such as the connection between the oil bag and the surrounding textures on the epidermis, and the typhoon scars that are produced in old tangerine peel.

Networks based on attention mechanisms [18] have become mainstream research, with Swim Transformer [19] gaining significant success on a variety of vision tasks and effectively solving the problem of large computational costs. However, it is very difficult to deploy the Swim Transformer since the calculation of the sliding window is very complex. To solve this problem, Liu et al. proposed ConvNeXt [20] by modifying the structure of ResNet [21]. Through a series of experimental comparisons, ConvNeXt has faster inference speed and a higher accuracy rate than Swim Transformer at the same FLOPS. Various attention mechanisms have been proposed to address the problem of difficult feature extraction. Hu et al. presented a squeeze and excitation module (SE) [22], where he added an attention mechanism to the channel dimension to obtain the importance of each channel of the feature map and assign weights to each feature by the importance level, thus allowing the network to learn the important features. Because previous attention mechanisms focus more on the analysis of the channel dimension, CBAM [23] implemented a sequential attention structure from both channel and spatial scopes. Spatial attention allows the neural network to focus more on the pixel regions in the image that play a decisive role in classification and ignore irrelevant regions, while channel attention is used to deal with the assignment relationship of the feature map channels, thus enhancing the effect of the attention mechanism on model performance. But these methods ignore the linkage between global feature information and local feature information, which can affect the fusion of features and the generation of accurate attention maps. Deng et al. built a csRSE module for occupancy grid map recognition [24], which contains a residual block for generating hierarchical features, followed by a channel SE block and a spatial SE block for adequate information extraction along the channel and space. To achieve more flexible computation allocation and content awareness, Zhu et al. introduced the content-independent sparsity into the attention mechanism and proposed the BiFormer which selectively attended to relevant tokens in an adaptive manner, without dispersing attention to other unrelated tokens [25].

3. Materials and Methods

3.1. Workflow. Figure 1 shows the workflow of the age recognition of the tangerine peel model. In this study, we used the CNFA-integrated network to identify the age of tangerine peel. First, we used a digital camera to capture the images of the tangerine peel. Then, we labeled the tangerine peel samples according to their age and created dataset. Finally, the model is trained and evaluated using the dataset. We input tangerine peel samples into the model. After the model training is completed, the model outputs the corresponding year of the sample.

3.2. Image Acquisition. There is currently no publicly available tangerine peel dataset to use, so it is necessary to collect images of tangerine peel. The tangerine peel sample was collected from Huicheng (Xinhui, Jiangmen, Guangdong Province, China, longitude 113.034 and latitude 22.4583). We collected sample with two batches. The sample information is given in Table 2. The original species of the tangerine peel samples is Dazhongyoushen. We used a Canon 760D camera (Canon Inc., Tokyo, Japan) with a resolution of 6000×4000 pixels to capture the epidermis of Xinhui tangerine peel under the same lighting conditions. The tangerine peels used for image acquisition are stored in the same warehouse with a consistent temperature, humidity level, and lighting condition. The temperature was 21°C, and the humidity was 60% in the warehouse. According to the experts, it was indicated that the samples underwent traceability detection before being sold. Therefore, the accuracy of the age can be ensured.

Since the tangerine peel sold online has a five-year interval between ages, the interval between each type of tangerine peel we collected is also five years. Under the premise of ensuring model performance and scientific sampling, we made efforts to achieve a concentrated distribution in our dataset. The 818 images of Xinhui tangerine peel were divided into five categories according to different ages. One sample corresponds to one image. The dataset was randomly divided into the training set, test set, and validation set in a ratio of 7 : 2 : 1. To decrease computational complexity and memory requirements, the training set images were uniformly converted to a resolution of 224×224 pixels. Figure 2 shows sample images of the dataset. The 1-year tangerine peel with bright orange skin and dense oil bags on the epidermis. The 5-year tangerine peel with reddish-brown skin and sparse oil bags on the epidermis. The 10-year tangerine peel with dark red skin and pig-bristle texture on the epidermis. The 15-year tangerine peel with black skin and more dense pig-bristle texture. The 20-year tangerine peel with typhoon scars on the epidermis.

Tangerine peel epidermis exhibits more age features. We used the tangerine peel dataset that distinguishes epidermis features in this experiment. The color is the most prominent

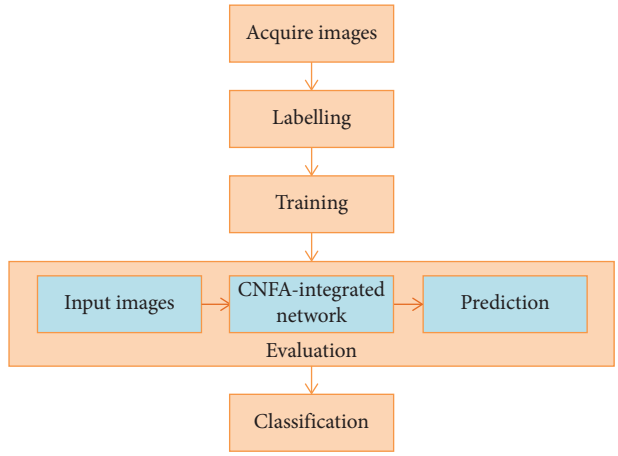

Figure 1: The workflow of the age recognition of the tangerine peel model.

Table 2: Sample information of 2 batches of different ages in the tangerine peel.

Location	Batch	Time	Category (year)	Sample size	Total sample size	The number of samples extracted	Total number of samples extracted
Huicheng, Xinhui	Batch 1	Mar 16, 2023	1	1,025	4,090	205	818
			15	940		188	
			20	465		93	
	Batch 2	Mar 17, 2023	5	875		175	
			10	785		157	

Figure 2: Samples images of the tangerine peel in different ages in the dataset.

low-level feature on the epidermis of tangerine peel, and the color of the epidermis will change with the increase of ages. However, the feature that oil bag and pig-bristle are the key to distinguishing the age of the tangerine peel. There are many sunken oil bags on the epidermis of tangerine peel, and the surrounding texture will combine with the oil bag, forming the pig-bristle texture of tangerine peel. With the increase of ages, the oil bags will dissipate, and the pig-bristle texture will become more obvious. In addition, old tangerine peel has increasingly obvious typhoon scars on the epidermis. Figure 3 shows the details of the oil bag, pig-bristle texture, and typhoon scar. The above features play a decisive role in the tangerine peel dataset.

3.3. CNFA-Integrated Network. Figure 4 shows the structure of the CNFA-integrated network, which consists of a stem convolution layer, a LN layer, CNFA stacked module with four stages, three downsampling layers, and a decision layer. The stem convolution layer has a kernel size of 4 and a stride of 4. The input image is processed by the stem convolution so that the continuous use of filters does not result in overlapping pixels. The use of the layer normalization (LN) layer [26] is to avoid the problems of gradient disappearance and gradient explosion. The proposed CNFA-integrated network is a multilevel design, with different feature map resolutions at each stage. The number of stacked modules of ResNet50 is (3, 4, 6, 3), with an approximate ratio of (1 : 1 : 2 : 1). In recent years, the number of stacked modules' ratio of most improved networks has been (1 : 1 : 3 : 1). To maintain the same network scale, the network is designed with the CNFA stacked module of four stages, with the number of modules stacked in each stage being (3, 3, 9, 3), and the input channel numbers being (96, 192, 384, 768), respectively. After passing through the CNFA module, the feature dimension of the feature map will change, leading to the loss of effective information. In order to ensure that effective information is retained, we added a separate downsampling layer between

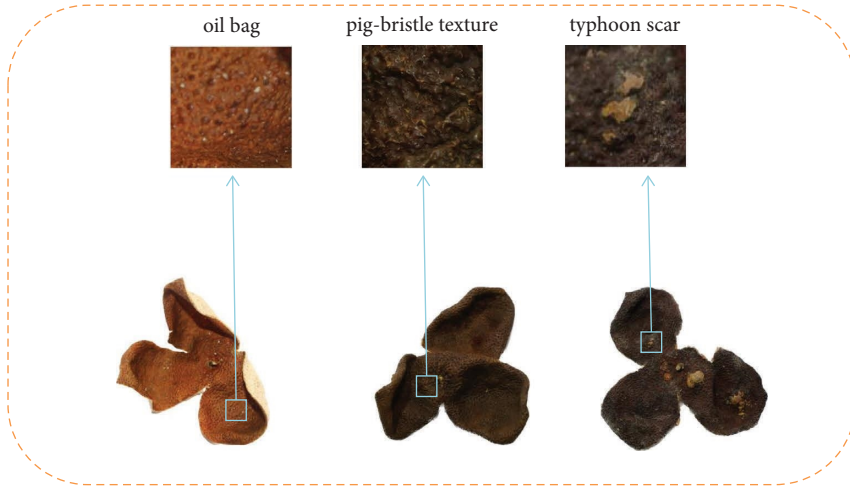

Figure 3: The details of the oil bag, pig-bristle texture, and typhoon scar of the tangerine peel.

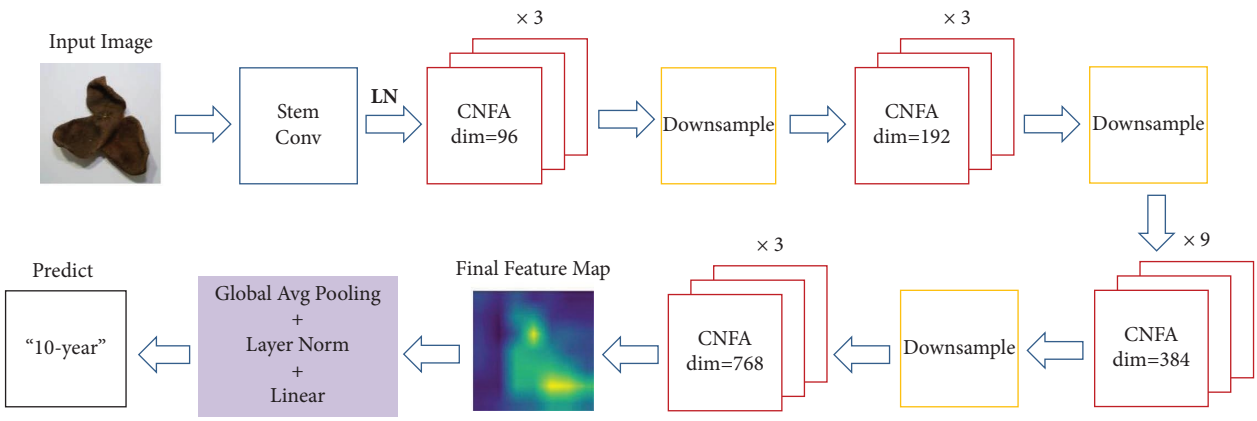

Figure 4: CNFA-integrated network architecture for age recognition of the tangerine peel. Dim is the number of channels input to the CNFA module.

CNFA modules. The downsampling layer includes an LN layer and a convolution layer with a kernel size of 2 and a stride of 2. After passing through the different convolution layers, the final feature map is output and the feature map is processed by the decision layer (consists of a global average pooling, a layer normalization layer, and a linear layer) to calculate the probability of predicting the category. GAP sums up all the pixel values of a feature map and calculates their average, resulting in a numerical value corresponding to the feature map. This technique reduces the number of parameters and computational workload. LN normalizes the values generated by GAP to improve the stability and training effectiveness of the network. Finally, the feature vector is fed into a linear layer to map it to the probability distribution of output classes.

For the task of identifying the age of the tangerine peel, the CNFA module is the core of feature extraction in the network and can effectively emphasize or suppress mapped feature information. As shown in Figure 5, the components of the CNFA module are a ConvNeXt block, a cSE block, and an sSE block. The ConvNeXt block is designed to capture the global low-level feature information and aggregate features at multiple levels. The sSE block determines the importance of specific positions in the input feature map and assigns corresponding weight parameters to highlight meaningful locations in spatial. Stacking the sSE block after cSE block can retain more intermediate layer spatial information. By placing the cSE block before the sSE block in sequential order, CNFA module models the correlations between channels and then further adjusts the spatial distribution of the feature map through the sSE. Through the calculations of the ConvNeXt block, the cSE block, and the sSE block, the ConvNeXt block extracts a low-level contextual feature and the attention module fusion generates a high-level contextual feature. The high-level features generated by the attention module are utilized to guide the ConNeXt block in extracting a low-level feature. These two features are aggregated into a global contextual feature. These features are weighted and averaged over all regions through an attention map.

The tangerine peel data input size is the same as the output size, and for the input feature map $X \in R^{H\times W\times C}$, the network computes the output feature map $Y \in R^{H\times W\times C}$. After the ConvNeXt block, the feature X_{CN} is computed:

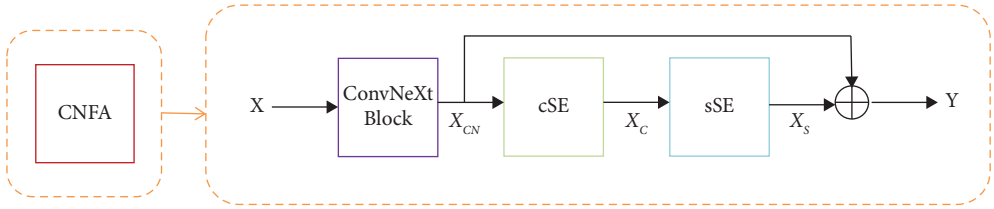

Figure 5: The component architecture of the CNFA module.

$$X_{\mathrm{CN}} = X \otimes F_{\mathrm{CN}}(X), \tag{1}$$

where $\otimes$ denotes element-wise multiplication.

The channel attention $F_C \in R^{1\times1\times C}$ is then calculated from the cSE block to obtain the output feature map of channel attention X_C:

$$X_C = X_{\mathrm{CN}} \otimes F_C(X_{\mathrm{CN}}). \tag{2}$$

The spatial attention $F_S \in R^{H\times W\times1}$ is then calculated from the sSE block to obtain the output feature map of spatial attention X_S:

$$X_S = X_C \otimes F_S(X_C). \tag{3}$$

The ConvNeXt block output feature map X_{CN} and the sSE block output feature map X_S are aggregated and then output to obtain the final refined output feature map Y:

$$Y = X_{\mathrm{CN}} + X_{\mathrm{CN}} \otimes X_S. \tag{4}$$

The following summary will describe the details of the ConvNeXt block, the cSE block, and the sSE block.

3.4. ConvNeXt Block. Figure 6 shows the structure of the ConvNeXt block, which consists of a 7×7 deep convolution layer, two 1×1 general convolution layers, an LN layer, a nonlinear Gaussian error linear unit (GeLU) activation layer [27], and a layer scale [28].

The 7×7 deep convolution layer mainly mixes spatial information, and a larger convolution kernel provides a larger receptive field to capture large-scale feature information. The large-kernel convolution operation is performed with a smaller number of channels to reduce the number of model parameters. The 1×1 general convolution layer expands and compresses the feature maps in the channel dimension to deepen the channels. This structure has a deep convolution layer as the front layer, and the subsequent general convolution layers are similar to the feed-forward block of TransFormer [29]. The reverse bottleneck structure is used to make the calculation of the ConvNeXt block more efficient. Therefore, the ConvNeXt block effectively and economically extracts global and local features.

In the ConvNeXt block, the use of the LN layer after the deep convolution layer is to avoid differences between training and inference. Considering that the variation in the output of one layer will cause strongly correlated changes in the total output of the next layer, the LN layer solves the covariate shift problem by setting the mean (μ) and variance (σ) of the summed inputs within each layer. For all hidden cells in the same layer, the LN layer is calculated as follows:

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(a_i^l - \mu^l\right)^2}, \tag{5}$$

where H indicates the number of hidden cells in the layer in which it is located.

To improve the nonlinearity and generalization ability of the model, a GeLU is used after the first general convolution layer. This activation function incorporates the idea of random regularization in the activation, which can achieve the effect of adaptive dropout and ensure the robustness of the model training. For input x, GeLU can be expressed as follows:

$$\mathrm{GeLU}(x) = \mathrm{xP}(X \le x) = x\varphi(x) = x \times \frac{1}{2}\left[1 + \mathrm{erf}\left(x/\sqrt{2}\right)\right]. \tag{6}$$

The purpose of the layer scale is to scale the input feature data, which allows for a more refined and precise representation of the features.

Thus, we input the x, and the output of the ConvNeXt block is calculated as follows:

$$F_{\mathrm{CN}}(X) = X + X \otimes f_3^{1\times1}\left(\mathrm{GeLU}\left(f_2^{1\times1}\left(\mu^l\left(f_1^{7\times7}\right)\right)\right)\right), \tag{7}$$

where *GeLU* is the GeLU function operation, μ^l is the LN layer operation, and $f^{7\times7}$ and $f^{1\times1}$ are the convolution layer operations with convolution kernel sizes of 7×7 and 1×1, respectively.

3.5. Channel Squeeze-and-Excitation Block (cSE). In the cSE block, we calibrate the correlation between image feature channels through spatial compression and channel excitation. As shown in Figure 7, the block first performs spatial compression. For the input feature vector $X_{\mathrm{CN}} \in R^{H\times W\times C}$, we use global average pooling to compress global spatial information and generate a unique channel vector $V_C \in R^{1\times1\times C}$ for each channel through the average value of global average pooling.

The block performs channel excitation to highlight channels with meaningful information. We take the dimensions obtained from the compression operation and run them through the multilayer perceptron (MLP) to count the weight values of the channels, which are then stimulated to the corresponding channels of the previous feature map for operation. The MLP consists of two fully connected layers,

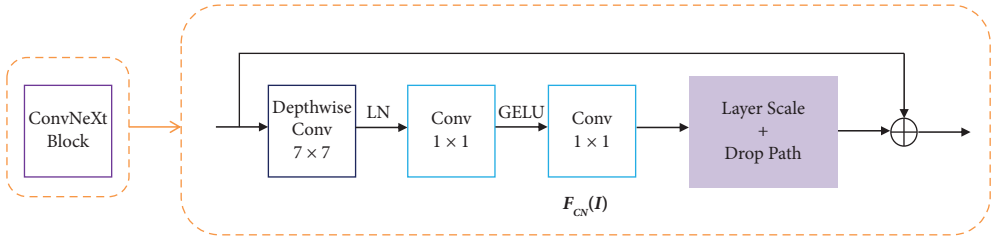

Figure 6: Component architecture of the ConvNeXt block: $F_{\mathrm{CN}}(I)$ denotes the ConvNeXt block operation, and I denotes the input.

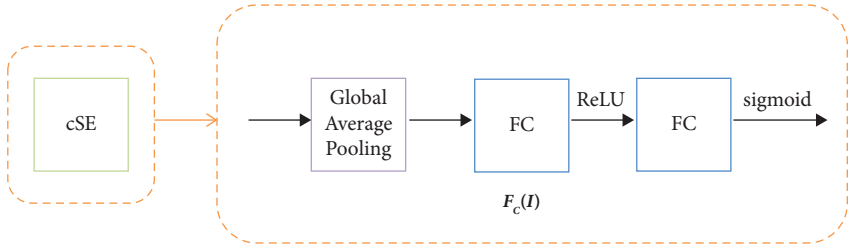

Figure 7: Component architecture of the cSE block: $F_C(I)$ denotes the cSE block operation, and I denotes the input.

a ReLU linear activation function and a sigmoid nonlinear activation function.

Thus, we input the X_{CN}, and the output of the cSE block is calculated as follows:

$$F_C(X_{\mathrm{CN}}) = \sigma(W_2\delta(W_1\mathrm{GAP}(X_{\mathrm{CN}}))), \tag{8}$$

where $W_1 \in R^{\overline{C}\times C}$, $W_2 \in R^{C\times\overline{C}}$ are the weights of the FC layers, δ is the ReLU activation function, and σ is the sigmoid activation function.

3.6. Spatial Squeeze-and-Excitation Block (sSE). In the sSE block, it is able to transform various deformation data in spatial and automatically capture important regional features. The setting of this block is to determine the importance of specific positions in the input feature map and highlight meaningful positions in spatial. As shown in Figure 8, the input feature vector X_C goes through several general convolution layers to generate an attention feature vector, which is then passed through a sigmoid function.

Thus, we input the X_C, and the output of the sSE block is calculated as follows:

$$F_S(X_C) = \sigma(W * X_C) = \sigma\left(f_4^{1\times1}\left(f_3^{3\times3}\left(f_2^{3\times3}\left(f_1^{1\times1}(X_C)\right)\right)\right)\right), \tag{9}$$

where σ is the sigmoid function, $*$ is the convolution operation, and $f^{1\times1}$ and $f^{3\times3}$ are the convolution layers with convolution kernel sizes of 1×1 and 3×3, respectively.

3.7. Parameter Selection and Model Training. We conducted experiments on the tangerine peel dataset for age recognition of the tangerine peel using the PyTorch framework. The network training was running on the NVIDIA RTX 3090 GPU. In the preset values for training, the learning rate was 0.0002, batch size was 16, weight decay was 0.0001, and the number of epoch was 200. The Adam optimizer was used to optimize the parameters, and the input and output image resolutions of the network were both 224×224. In this experiment, we used the cross-entropy (CE) loss function [30] to train the network. CE is shown as follows:

$$L = -\sum_{i=1}^{n} y_i \log(p_i), \tag{10}$$

where y_i is the true label and p_i is the predicted probability of the ith item.

To observe the training situation in real time, we validated the trained model on the validation set after each epoch of training. As shown in Figure 9, the training loss and validation loss of the CNFA-integrated network were unstable and high at the beginning stage, but they tended to stabilize between 25 and 50 training epochs as the number of training epochs increased. The stable training loss and validation loss in the later stage of training indicate that the CNFA-integrated network did not have overfitting. The model converged under the input data, and both loss values were less than 1 after stabilization, proving that the CNFA-integrated network can be used for age recognition of the tangerine peel. As shown in Figure 10, the training accuracy and validation accuracy of the CNFA-integrated network on the tangerine peel dataset tend to stabilize between 25 and 50 training epochs as the loss and the learning rate decrease. The training accuracy reached 100, and the validation accuracy reached about 96.88, proving that the CNFA-integrated network learned the important features of tangerine peel.

3.8. Model Evaluation Metrics. For age recognition of the tangerine peel, we use accuracy to evaluate the model. In order to evaluate the detection of each category of tangerine peel by the model, precision, recall, and F1 will also be used in the evaluation metrics. The formulas are as follows:

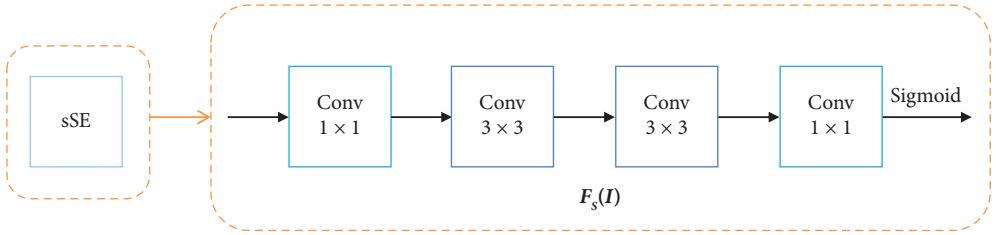

Figure 8: Component architecture of the sSE block: $F_{CN}(I)$ denotes the sSE block operation, and I denotes the input.

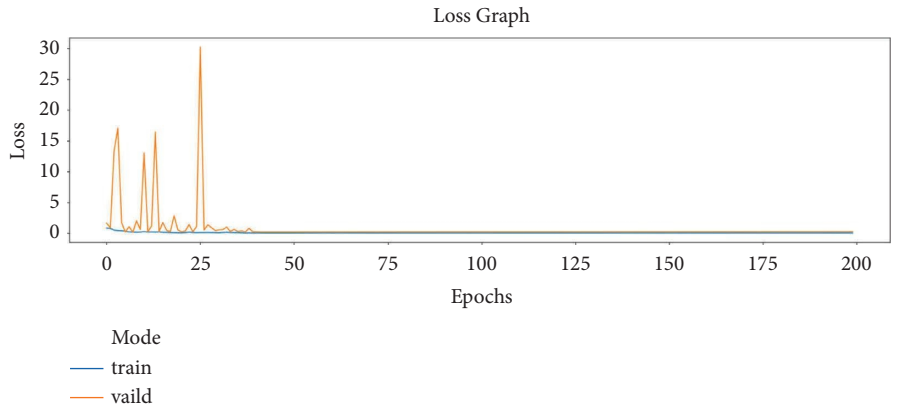

Figure 9: Loss curve of the CNFA-integrated network on the tangerine peel dataset.

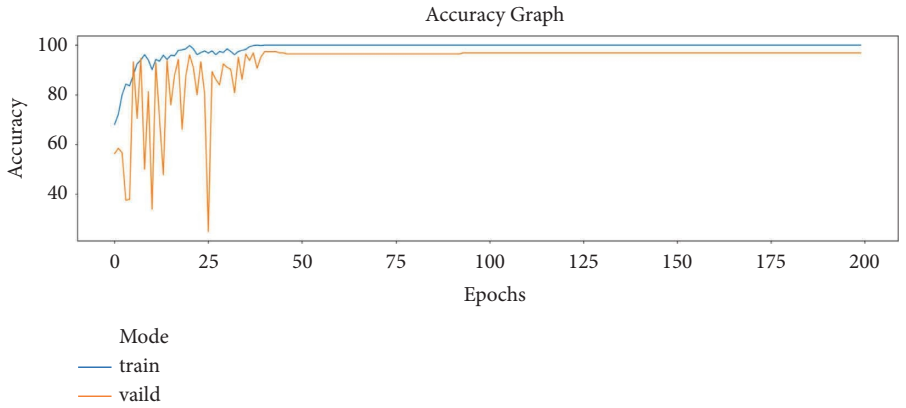

Figure 10: Accuracy curve of the CNFA-integrated network on the tangerine peel dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \times 100\%, \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \tag{13}$$

$$F1 = \frac{\text{TP}}{\text{TP} + 1/2\,(\text{FP} + \text{FN})} \times 100\%, \tag{14}$$

where TP is a true positive, FP is a false positive, TN is a true negative, and FN is a false negative.

After the model training is completed, the network model is tested on the test set. The test sets are traversed, and we predict the category of each image, then the prediction is analyzed according to the ground truth to determine whether it is correct. For a certain category of tangerine peel, if it is predicted correctly, it is TP and if it is predicted incorrectly, it is FN. Other categories of tangerine peel are negative; if they are predicted correctly, they are TN, and if they are predicted incorrectly, they are FP.

4. Results and Discussion

4.1. Implementation Details. We evaluated the performance of each model on the tangerine peel dataset and conducted subsequent experiments. Each experiment was implemented on a computer equipped with 32 GB RAM, Intel i9 CPU, NVIDIA GeForce RTX 3090 GPU, and Ubuntu16.04 operating system. Each model was trained using the same training set. Also, each model used the same training parameters as the CNFA-integrated network.

4.2. Manual Identification Results. Before validating the model's performance, we conducted a manual identification experiment. This experiment evaluated the effectiveness of manual identification based on the accuracy of three experts. The identification was performed on the test set of the tangerine peel dataset, which consists of a total of 164 images.

As shown in Table 3, the accuracy rates of the three experts were 90.12%, 74.07%, and 83.95%, with an average of 82.71%. Each expert had a different accuracy rate, and there was a significant difference in accuracy among them. This is because manual identification can be influenced by subjectivity, making the accuracy of manual detection less stable. Compared to manual identification methods, deep learning methods are more stable and have higher accuracy.

4.3. Model Evaluation Results. We demonstrated the effectiveness of the CNFA-integrated network by comparing it with other mainstream network models on the tangerine peel dataset. While ensuring accuracy above 95%, we maintained the model scale in our comparative experiments. In the experiments, we compared with CNN, ResNet50 and ResNet50 variants, and ConvNeXt to evaluate their performance for age recognition of the tangerine peel.

As shown in Table 4, the CNN achieved the accuracy of 82.59%, precision of 82.60%, recall of 81.65%, and F1 score of 82.12%. It was the worst performing model among all models evaluated, indicating poor feature aggregation performance of CNN in task. Compared with CNN, the accuracy of the ResNet50 was increased by 13.39%, the precision was increased by 12.05%, the recall was increased by 13.27%, and the F1 score was increased by 12.57%. It indicates that the residual structure network performs well for age recognition of the tangerine peel. After adding the attention module to ResNet50, the metrics are improved. Adding SE module and CBAM module improved accuracy by 0.25% and 0.29%, improved precision by 0.75% and 0.88%, improved recall by 0.54% and 0.2%, and improved F1 by 0.73% and 0.63%, respectively. It indicates that there is not much difference in performance between ResNet-SE and ResNet-CBAM. Both CBAM module and csRSE module are dual attention modules. Compared with the csRSE module that focuses on global features, ResNet-csRSE had a higher accuracy, precision, recall, and F1 by 0.29%, 1.14%, 0.54%, and 0.84% than ResNet-CBAM. ConvNeXt achieved accuracy of 96.70%, precision of 96.18%, recall of 96.09%, and F1 score of 96.13%. ConNeXt performed better than previous ResNet50 variant networks. BiFormer achieved accuracy of 96.38%, precision of 96.07%, recall of 95.91%, and F1 score of 95.99%. The performance of BiFormer performed slightly worse than ConvNeXt. The proposed CNFA-integrated network is a variant based on ConvNeXt. The accuracy was 97.17%, the precision was 96.71%, the recall was 96.86%, and the F1 score was 96.78%. CNFA-integrated had a higher accuracy, precision, recall, and F1 by 0.47%, 0.53%, 0.77%, and 0.65% than ConvNeXt. The metrics reached their maximum values, demonstrating the advantage of the CNFA-integrated network for age recognition of the tangerine peel. Through the comparative experiments results, the proposed CNFA-integrated network effectively captures global high-level and low-level information and aggregates information effectively through various modules. It validates the effectiveness of the CNFA module in detection accuracy.

Table 3: Results of manual identification experiments.

	Expert 1	Expert 2	Expert 3	Average
Accuracy (%)	90.12	74.07	83.95	82.71

We also conducted experiments on the processing speed. The detection time of the CNN was the longest, at 104.23 seconds. The CNN performed poorly in both performance and speed. Compared with CNN, the detection time of ResNet50 reduced by 10.54 seconds. Compared with ResNet50, adding the SE module, CBAM module, and csRSE module reduced the detection time by 1.67 seconds, 2.72 seconds, and 2.84 seconds, respectively. Among them, csRSE had the shortest detection time. Compared with ResNet50-CBAM and ResNet50-csRSE, ConNeXt had an increased detection time. This is because ConNeXt has a larger number of parameters, resulting in increased computational complexity. BiFormer had a longer detection time compared to other attention-based networks. The attention mechanism in BiFormer has a certain degree of sequentiality. The model's computations depend on previous information, resulting in a large computational workload and increased detection time. Our proposed CNFA-integrated network reduced the detection time, which added attention mechanisms on ConNeXt. The CNFA module helps the model focus on important information in the input data, thereby reducing computational load and processing time. The CNFA-integrated network achieved a shorter detection time while ensuring the highest accuracy, with a small difference compared to csRSE. It validates the effectiveness of the CNFA module in detection efficiency.

4.4. Ablation Experiments. To validate the effectiveness of the CNFA module, we conducted a series of ablation experiments. The ablation experiments included four control groups. The control groups consisted of ConvNeXt, ConvNeXt and cSE, ConvNeXt and sSE, and CNFA. These four network structures were trained on the tangerine peel dataset. Table 5 shows the accuracy, precision, recall, and F1 of the models tested. CNFA performed the best, followed by ConvNeXt, in terms of performance. However, when adding the cSE block and sSE block individually to ConvNeXt, the performance decreased. CNFA had a higher accuracy, precision, recall, and F1 by 0.62%, 1.1%,

Table 4: Results of model comparison experiments.

Network	Param (M)	FLOPS (G)	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Time (s)
CNN	25.4	4.217	82.59	82.60	81.65	82.12	104.23
ResNet50	25.6	4.158	95.98	94.65	94.92	94.69	93.69
ResNet50-SE	28.1	4.162	96.23	95.40	95.46	95.42	92.02
ResNet50-CBAM	28.1	4.168	96.27	95.53	95.12	95.32	90.97
ReNet50-csRSE	28.1	4.175	96.67	96.67	95.66	96.16	**90.85**
BiFormer	25.5	4.5	96.38	96.07	95.91	95.99	98.52
ConvNeXt	28.6	4.546	96.70	96.18	96.09	96.13	91.67
CNFA-integrated (ours)	30.1	4.551	**97.17**	**96.71**	**96.86**	**96.78**	91.02

Bold indicates the best performance.

Table 5: Ablation experiments result.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
ConvNeXt	96.70	96.18	96.09	96.13
ConvNeXt + cSE	96.55	95.61	95.9	95.75
ConvNeXt + sSE	95.26	94.15	94.91	94.53
CNFA (ours)	**97.17**	**96.71**	**96.86**	**96.78**

Bold indicates the best performance.

Table 6: Various classification metrics for tangerine peel in different ages.

Category (year)	Precision (%)	Recall (%)	F1 (%)
1	97.56	97.56	97.56
5	97.14	97.14	97.14
10	96.77	100	98.35
15	97.37	94.87	96.10
20	94.73	94.73	94.73

0.96%, and 1.03% than ConvNeXt and cSE. CNFA had a higher accuracy, precision, recall, and F1 by 1.91%, 2.02%, 1.95%, and 2.25% than ConvNeXt and sSE. This is because the exclusion of either channel or spatial information leads to the loss of important details and contextual in the tangerine peel dataset, resulting in incorrect data interpretation. Therefore, attention mechanisms should consider a balanced integration of both channel and spatial information.

4.5. The Result of the Classification Metrics. As shown in Table 6, it records the prediction metrics of each category in the test set. The CNFA-integrated network trained on the tangerine peel dataset and obtained the prediction metrics for each category of tangerine peel on the test set: precision, recall, and F1 score. The classification metrics of each category of tangerine peel were relatively high after feature learning with the CNFA-integrated network, indicating good recognition performance. Because there were fewer test data for 20-year tangerine peel, the displayed metrics were lower than other categories.

We further tested the performance of the CNFA-integrated network by using a confusion matrix. The confusion matrix with prediction in the columns and real label in the row exhibited the performance of the CNFA-integrated network. Figure 11 shows a confusion matrix of the CNFA-integrated network. The CNFA-integrated network achieved 97.17% accuracy in the test set. From the confusion matrix, only one sample is misclassified from each category. It indicates that the CNFA-integrated network has good generalization performance.

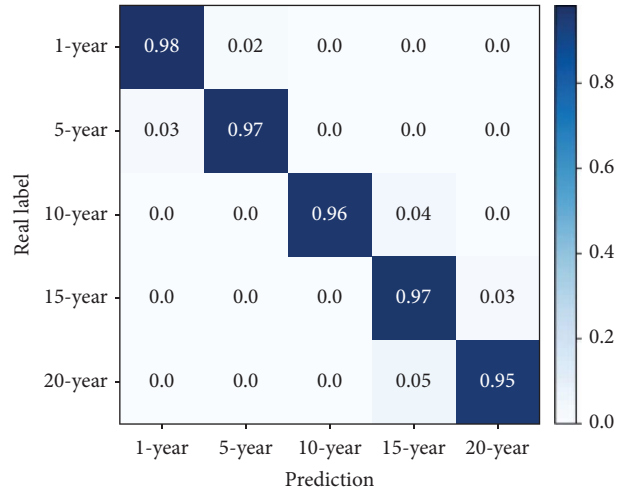

Figure 11: The confusion matrix of the CNFA-integrated network.

5. Discussion

The CNFA module can rapidly and accurately identify the age of tangerine peel. Through our manual identification experiments, we found that the accuracy fluctuates up to 16%. The instability of the manual detection method can impact the assessment process of tangerine peel quality. Deep learning-based age recognition of the tangerine peel avoids subjectivity and provides more stable results. In our comparative experiments, all models demonstrated good performance. By utilizing deep learning models, the accuracy of identifying age had essentially reached 90% above. Our proposed CNFA-integrated network achieved the highest accuracy, precision, recall, and F1 scores in the comparative experiments. Additionally, the CNFA-integrated network exhibited fast processing speed. We also showed the various classification metrics for tangerine peel in different ages, and each metric achieved about 95%. This indicates that the CNFA module exhibits strong classification capability for each category of tangerine peel.

To visually demonstrate the effectiveness of the CNFA module, the heat map visualization method we use is called Grad-CAM [31]. It generates a heat map by analyzing the

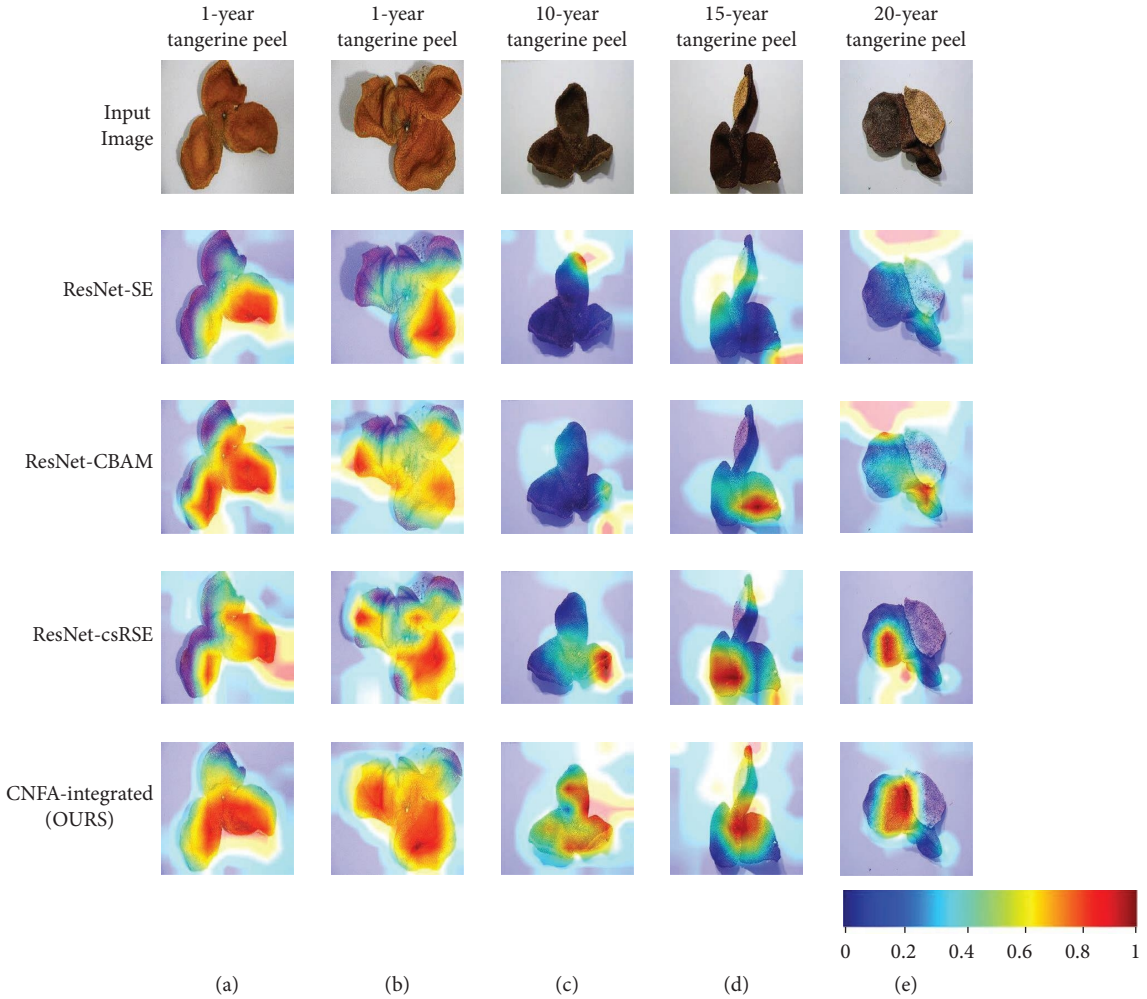

Figure 12: The visualization results of the heat maps on the tangerine peel dataset.

gradient information on the input image, visualizing the model's attention to different regions. As shown in Figure 10, the color scale of the heat map is in the bottom right corner. The values of the feature map are mapped to the range of [0, 1]. The heat map we generated shows red for high weight allocation positions and blue for low weight allocation positions. By overlaying the tangerine peel image with the result of Grad-CAM, we can effectively display the areas where the network has an impact on the task results.

As shown in Figure 12, the following are the input images of tangerine peel and their heat maps generated by different networks, including ResNet-SE, ResNet-CBAM, ResNet-csRSE, and CNFA-integrated network. The confidence score threshold was set to 0.5, and results with a confidence score less than 0.5 were considered to be wrong. These five images of tangerine peel were predicted incorrectly by ResNet-SE, ResNet-CBAM, and ResNet-csRSE. The samples were predicted successfully by the CNFA-integrated network, and the average confidence score is approximately 0.92. We will discuss the differences between the CNFA module and other attention mechanisms based on the features and heat maps of the output images.

Figure 12(a) shows a 1-year tangerine peel with a bright orange skin and dense oil bags on the epidermis. The sample was predicted to be five-year tangerine peel in ResNet variants because the colors of one-year tangerine peel and five-year tangerine peel are similar. In this sample, the dual-channel attention module performed well. However, the CBAM and csRSE modules still had issues with inaccurate localization of important features. The attention scope extended beyond the shape edges of the tangerine peel. The attention of the CNFA-integrated network was distributed in accurate regions, which can capture more areas of oil bags. Figure 12(b) shows a 1-year tangerine peel. Due to the high degree of curling in this sample, it is difficult to capture global features. The sample was predicted to be five-year tangerine peel in ResNet variants. The SE module failed to capture the shape of the sample, and the localization of the regions of interest was not comprehensive enough. Although the CBAM and csRSE modules extracted shape features more effectively, their ability to extract low-level features was insufficient, leading to incorrect prediction of the sample in Figure 12(b). The CNFA module located features of the age during decision-making, thereby improving the

prediction ability of the network. Figure 12(c) shows a 10-year tangerine peel with a dark red skin and pig-bristle texture on the epidermis. The heat maps generated by the SE, CBAM, and csRSE modules were not ideal, as the red regions only covered a local position of the tangerine peel, ignoring the overall shape and inaccurate localization of the features of the age. Three ResNet variants predicted the sample to be 15-year tangerine peel. The attention in the CNFA-integrated network covered the overall shape of the tangerine peel and accurately located the important feature areas. Figure 12(d) shows a 15-year tangerine peel with a black skin and a denser pig-bristle texture. In the heat maps generated by the three ResNet variants, red areas were mainly distributed in one section of the tangerine peel, ignoring the features of the age in other regions. Three ResNet variants predicted the sample to be 10-year tangerine peel or 20-year tangerine peel. The attention of the CNFA-integrated network was distributed in overall shape, which can capture more areas with a pig-bristle texture. Figure 12(e) shows a 20-year tangerine peel with typhoon scars on the epidermis. Three ResNet variants predicted the sample to be 15-year tangerine peel. The CNFA module accurately located the attention on the typhoon scars, while other attention modules ignore this important feature. This is mainly because the CNFA module can extract global feature information of tangerine peel. It helps the network accurately locate important feature positions, thus improving the accuracy of age recognition of the tangerine peel.

It can be seen that the CNFA proposed by us successfully recognizes the important features on the epidermis of tangerine peel in different ages. It aggregates low-level and high-level features on the epidermis of tangerine peel to provide more information for feature localization and can accurately locate the regions of interest to the important feature positions. By generating heat maps, we have effectively demonstrated that the CNFA module helps the network more accurately detect the appearance and details of different tangerine peels.

The method enables product quality control, traceability, and anticounterfeiting in the intelligent tangerine peel industry. By identifying the age of tangerine peel, producers can ensure that the products comply with standards and regulations. By utilizing age identification of the tangerine peel technology, it becomes possible to achieve anticounterfeiting and traceability for tangerine peel. Each tangerine peel can be associated with its unique age information, ensuring the authenticity and traceability of the age. By improving the database, this method can identify related varieties, counterfeit products, and artificially aged samples. Our algorithm has achieved a high level of accuracy. Additionally, the average detection time for detecting a single image is 0.55 seconds. Currently, the application is transmitting the data remotely to a server for processing and generating output results. The future work will focus on researching model lightweighting and efficiency improvement. This will facilitate the deployment of the model onto real-time terminals to achieve real-time detection.

6. Conclusions

This article proposes a method for age recognition of the tangerine peel based on the CNFA module to address the difficulty. Tangerine peel images are collected using a conventional digital camera and classified through the network model. This network can extract the global information of the tangerine peel and identify the important features that determine the age of the tangerine peel. The CNFA-integrated network had an accuracy of 97.17%, precision of 96.18%, recall of 96.09%, and F1 score of 96.13%, which did best in the comparison experiment. Furthermore, the CNFA module also exhibits fast processing speed. Finally, the validity of the model in the recognition task was verified through a visualization method based on heat maps, which concentrated the regions of interest on the important features of the tangerine peel and improved the detection accuracy of the age recognition. Therefore, this work has important application value for the identification of agricultural products represented by tangerine peel. Based on the excellent performance of this neural network, further exploration is needed. In future work, we will study multimodal network structures to improve detection accuracy and efficiency and achieve a lightweight structure to address the problem of extracting epidermis feature and endocarp feature.

Data Availability

The data will be made available upon reasonable request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Fuqin Deng performed writing of the original draft; Junwei Li conducted experiment, data analysis, modelling, and writing of the original draft; Lanhui Fu performed optimisation of the language and writing of the original draft; Chuanbo Qin carried out sample resource collection and processing; Yikui Zhai collected literature and performed optimisation of the language; Hongmin Wang proofread the study; Ningbo Yi proofread the study; Nannan LI proofread the study; TinLun Lam proofread the study. All the authors have read and agreed to the published version of the manuscript. Fuqin Deng and Junwei Li contributed equally to this work and shared the first authorship.

Acknowledgments

This work was supported in part by Wuyi University-Hong Kong-Macau Joint Funding Scheme (2022WGALH17, 2021WGALH18), the Science and Technology Development Fund(FDCT) of Macau under Grant no. 0071/2022/A, the funding (AC01202101103) from the Shenzhen Institute of Artificial Intelligence and Robotics for Society, and PhD Research start-up Fund of Wuyi University (No. BSQD2222).

References

[1] L. Yi, N. Dong, S. Liu, Z. Yi, and Y. Zhang, "Chemical features of pericarpium Citri reticulatae and pericarpium Citri reticulatae Viride revealed by GC–MS metabolomics analysis," *Food Chemistry*, vol. 186, pp. 192–199, 2015.

[2] X. Li, Y. Yang, Y. Zhu, A. Ben, and J. Qi, "A novel strategy for discriminating different cultivation and screening odor and taste flavor compounds in Xinhui tangerine peel using E-nose, E-tongue, and chemometrics," *Food Chemistry*, vol. 384, 2022.

[3] L. Lin, Z. X. Liu, and Y. Y. Mo, "Dynamic analysis of the total flavone and the hesperidin from different specific years in XinHui dried tangerine peel," *Lishizhen Medicine and Materia Medica Research*, vol. 19, no. 6, pp. 1432-1433, 2008.

[4] S. C. Ho and C. T. Kuo, "Hesperidin, nobiletin, and tangeretin are collectively responsible for the anti-neuroinflammatory capacity of tangerine peel (Citri reticulatae pericarpium)," *Food and Chemical Toxicology*, vol. 71, pp. 176–182, 2014.

[5] D. Y. Lee, Q. Y. Li, J. Liu, and T. Efferth, "Traditional Chinese herbal medicine at the forefront battle against COVID-19: clinical experience and scientific basis," *Phytomedicine*, vol. 80, 2021.

[6] S. Pan, X. Zhang, W. Xu, J. Yin, H. Gu, and X. Yu, "Rapid On-site identification of geographical origin and storage age of tangerine peel by Near-infrared spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 271, 2022.

[7] S. Li, B. Li, J. Li, and B. Liu, "Brown rice germ integrity identification based on deep learning network," *Journal of Food Quality*, vol. 2022, Article ID 6709787, 18 pages, 2022.

[8] L. Fu, F. Wu, X. Zou et al., "Fast detection of banana bunches and stalks in the natural environment based on deep learning," *Computers and Electronics in Agriculture*, vol. 194, 2022.

[9] H. Zhang, J. Cui, G. Tian et al., "Efficiency of four different dietary preparation methods in extracting functional compounds from dried tangerine peel," *Food Chemistry*, vol. 289, pp. 340–350, 2019.

[10] R. Chen, C. Jin, Z. Tong et al., "Optimization extraction, characterization and antioxidant activities of pectic polysaccharide from tangerine peels," *Carbohydrate Polymers*, vol. 136, pp. 187–197, 2016.

[11] F. Li, Y. Lu, C. Li et al., "trnL-trnF copy number is inversely correlated with storage time of Guang Chenpi, the aged sun-dried peels of Citrus reticulata 'Chachi," *Journal of Stored Products Research*, vol. 97, 2022.

[12] F. Yue, F. Zhang, Q. Qu et al., "Effects of ageing time on the properties of polysaccharide in tangerine peel and its bacterial community," *Food Chemistry*, vol. 417, 2023.

[13] X. Zhang, Z. Gao, Y. Yang, S. Pan, J. Yin, and X. Yu, "Rapid identification of the storage age of dried tangerine peel using a hand-held near infrared spectrometer and machine learning," *Journal of Near Infrared Spectroscopy*, vol. 30, no. 1, pp. 31–39, 2022.

[14] H. Pu, J. Yu, D. W. Sun, Q. Wei, and Q. Li, "Distinguishing pericarpium citri reticulatae of different origins using terahertz time-domain spectroscopy combined with convolutional neural networks," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 299, 2023.

[15] A. Helwan, M. K. Sallam Ma'aitah, R. H. Abiyev, S. Uzelaltinbulat, and B. Sonyel, "Deep learning based on residual networks for automatic sorting of bananas," *Journal of Food Quality*, vol. 2021, Article ID 5516368, 11 pages, 2021.

[16] F. Meng, J. Li, Y. Zhang, S. Qi, and Y. Tang, "Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 214, 2023.

[17] Z. Chu, F. Li, D. Wang, S. Xu, C. Gao, and H. Bai, "Research on identification method of tangerine peel year based on deep learning," *Food Science and Technology*, vol. 42, 2022.

[18] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *Journal of Neuroscience*, vol. 13, no. 11, pp. 4700–4719, 1993.

[19] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, Montreal, QC, Canada, October 2021.

[20] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, Seattle, WA, USA, June 2022.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[23] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, *Computer Vision – ECCV 2018*, pp. 3–19, Tel Aviv, Israel, October 2018.

[24] F. Deng, H. Feng, M. Liang et al., "Abnormal occupancy grid map recognition using attention network," in *Proceedings of the 2022 International Conference on Robotics and Automation (ICRA)*, pp. 8666–8672, Philadelphia, PA, USA, May 2022.

[25] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: vision transformer with bi-level routing attention," in *Proceedings of the conference on computer vision and pattern recognition (CVPR)*, pp. 10323–10333, Mancouver, BC, Canada, June 2023.

[26] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, https://arxiv.org/abs/1607.06450.

[27] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2016, https://arxiv.org/abs/1606.01540.

[28] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 32–42, Montreal, QC, Canada, October 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, Venice, Italy, October 2017.