DETECTION OF EARLY BRUISING IN 'HUANGGUAN' PEAR BASED ON MCC-DeepLabV3+ MODEL

/

基于 MCC-DeepLabV3+模型的皇冠梨早期瘀伤检测

Congkuan YAN, Haonan ZHAO, Dequan ZHU, Yuqing YANG, Ruixing XING, Qixing TANG, Juan LIAO*)
School of Engineering, Anhui Agricultural University, Hefei 230036, China
Tel: +8617755108602; Email: liaojuan@ahau.edu.cn
Corresponding author: Juan Liao
DOI: https://doi.org/10.35633/inmateh-76-23

Keywords: *early bruise detection, MCC-DeepLabV3+, near infrared imaging, semantic segmentation, 'Huangguan' pear*

ABSTRACT

Due to their delicate and thin skin, 'Huangguan' pears are very vulnerable to pressure and impact during picking, packing and transportation, which can cause bruising. Early detection of bruises allows for timely identification of affected fruits to reduce potential food safety risks. However, early bruises in 'Huangguan' pears, particularly those that occur within the 30 minutes, often show no visible differences in external features compared to healthy tissue, making conventional techniques such as manual and machine vision sorting ineffective. Accordingly, a near-infrared (NIR) camera imaging technique combined with deep learning segmentation algorithm for early bruise 'Huangguan' pears detection is proposed in this study. Firstly, a near-infrared camera imaging system is applied to collect early bruise images of 'Huangguan' pears, and then a lightweight segmentation model based on the DeepLabV3+ architecture, referred to as MCC-DeepLabV3+ is presented. In the MCC-DeepLabV3+ model, MobileNetV2 is used as the backbone network, reducing the parameter size and enhancing deployment efficiency. Additionally, the coordinate attention (CA) mechanism is integrated into the shallow feature extraction and ASPP modules to improve the extraction of positional information across various features, minimizing the discrepancy between segmented areas and the actual bruised regions. Furthermore, a cascade feature fusion (CFF) strategy is incorporated into the encoder to reduce segmentation edge discontinuities and ensure effective multi-level semantic fusion, improving segmentation accuracy. The experimental results show that the proposed model has achieved a mIoU of 95.68%, and mPrecision of 97.58% on the self-built dataset of early bruising in 'Huangguan' pears. Compared to benchmark models such as U-Net, SegNet, PSPNet and HRNet, the proposed model demonstrates superior segmentation performance, offering promising support for the development of nondestructive detection techniques for agricultural product quality.

摘要

由于黄冠梨的表皮细腻而娇嫩，在采摘、包装和运输过程中非常容易受到压力和冲击，这可能导致瘀伤。早期发现瘀伤可以及时识别受影响的水果，有助于减少潜在的食品安全风险。然而，皇冠梨的早期瘀伤，特别是那些在 30 分钟内发生的瘀伤，与健康组织相比，通常在外部特征上没有明显的差异，这使得人工和机器视觉分类等传统技术的效果不佳。因此，本研究提出了一种结合深度学习分割算法的近红外(NIR)相机成像技术用于早期皇冠梨瘀伤检测。首先，采用近红外相机成像系统采集黄冠梨的早期瘀伤图像，然后，提出了一种基于 DeepLabV3+架构的轻量级分割模型，称为 MCC-DeepLabV3+。该模型采用 MobileNetV2 作为骨干网络，减少了参数大小，提高了部署效率。此外，将坐标注意（CA）机制集成到浅层特征提取和 ASPP 模块中，提高了不同特征之间的位置信息提取，最大限度地减少了分割区域与实际损伤区域之间的差异。在编码器中引入级联特征融合（CFF）策略，减少了分割边缘不连续，保证了有效的多级语义融合，提高了分割精度。实验结果表明，在自建的黄冠梨早期瘀伤数据集上，该模型的 mIoU 和 mPrecision 分别达到了 95.68%和 97.58%。与 U-Net、SegNet、PSPNet 和 HRNet 等基准模型相比，该模型显示出优越的分割性能，为水果早期无损检测技术的发展提供了有力的支持。

INTRODUCTION

"Huangguan" pear is a high-quality pear variety widely grown in China, which has become an important cash crop in many areas because of its attractive appearance, delicate texture and high yield. However, the soft structure and high water content of 'Huangguan' pears make them susceptible to physical trauma or impact damage during harvesting, packing and transport, resulting in early tissue bruising. Bruising is usually defined as damage to the subcutaneous tissue of fresh produce that does not result in rupture of the epidermis *(Al-Dairi et al., 2024*). Early bruises in 'Huangguan' pears are cosmetically indistinguishable from healthy tissue and are difficult to detect with the naked eye, but they provide an ideal environment for microbial colonisation. If early damage is not detected and treated in a timely manner, the resulting microorganisms may spread and infect other healthy fruits, thus seriously affecting product quality and economic value *(Opara et al., 2014)*. Traditionally, testing the quality of 'Huangguan' pears in packinghouses has relied on manual inspection by trained professionals *(Patel et al., 2024)*, but this method is highly subjective and suffers from limitations such as time-consuming, labour-intensive and the possibility of secondary damage to the fruit *(Li et al., 2024)*. To cope with these shortcomings, machine vision technology based on RGB images is widely used *(Manavalan, 2020; Arumuga et al., 2023; Tian et al., 2024)*. The technology achieves a reduction in labour costs and the rate of misjudgement as well as an increase in detection speed during the inspection process by capturing RGB images of fruits and combining them with image processing algorithms *(Du et al., 2020)*. However, since the RGB image inspection method relies on surface visible light information, its detection effect is mainly limited to surface defects visible to the naked eye and restricted to early bruises or internal damage hidden under the fruit skin *(Li et al., 2021)*. Therefore, this method still has significant shortcomings in meeting the needs of early bruise detection.

It is well known that when fruit undergoes physical trauma, the pulp tissue is damaged and a series of complex physiological and biochemical reactions occur at the site of injury, altering the density and vibrational strength of the molecular bonds *(Li et al., 2018)*. These changes affect the sugar and water content, which leads to alterations in optical properties, making it possible to utilize optical properties for early identification of internal damage. Several studies have used hyperspectral imaging to extract spectral information to evaluate the intrinsic and extrinsic qualities of fruit. *Zhu et al., (2016)* applied hyperspectral imaging and chemo-metrics to predict the integrated quality of tomato non-destructively and determine the optimal harvesting period. *Wu et al., (2023)* used HSI combined with spectral and texture features to detect and classify early damage of Lingwu longdate based on 1D convolutional neural network (1D-CNN). *Liu et al., (2023)* used hyperspectral imaging, transfer learning and convolutional neural network modelling techniques to detect early mechanical damage in pears, and the resulting model achieved an accuracy of 96.61% on the test set, which was 3.64% higher than the pre-fine-tuning network. The above studies show that HSI performs well in fruit damage detection, but it faces challenges in practical applications such as high equipment cost, complex data processing and high computational resource requirements *(Mei et al., 2023)*. Compared with hyperspectral imaging and processing, NIR imaging offers simpler data processing, eliminating the need to analyse large amounts of spectral band information, which significantly simplifies data processing and reduces inspection costs. Due to the higher absorption of NIR light by bruised tissue, NIR cameras are able to acutely capture subcutaneous damage that cannot be identified by the naked eye and visible light, and visualise healthy tissue and damaged areas by combining spectral information with spatial information *(Ünal et al., 2024)*. In addition, the high sensitivity of NIR cameras to changes in key components such as moisture and sugar, as well as their small size and light weight, make them ideal for early bruise detection in fruit.

Based on near infrared imaging, traditional machine learning techniques are combined for fruit damage detection. *Nandi et al., (2016),* proposed an automatic mango ripeness grading method based on support vector machines (SVM) and fuzzy learning, achieving a grading accuracy of 87%. *Hu et al., (2018)* developed an algorithm for bruised apple identification using 3-D infrared imaging, featuring a vertex-based local binary pattern (vmLBP) for feature extraction and SVM for classification. The algorithm achieved a 91.83 ± 0.46% accuracy, outperforming traditional methods. However, the performance of machine learning models is highly dependent on feature selection, requiring manual elimination of redundant features or identification of the most representative information to minimize the effect of irrelevant input variables *(Li et al., 2022)*, which is not only time-consuming but also depends on expert knowledge. In contrast, due to its powerful feature learning capabilities, deep learning has become an alternative technology in several areas of agriculture *(Attri et al., 2023)*, such as pest and disease detection *(Shafik et al., 2024)*, crop yield estimation *(Barbosa et al., 2021)*, water management *(Chen et al., 2021)*, and soil analysis *(Zhong et al., 2021)*.

Deep learning automatically learns important features from images through multi-layer neural networks, significantly simplifying the feature extraction process, thereby eliminating the reliance on experts to manually select features *(Wu et al., 2020)*.This process not only improves classification accuracy but also reduces human intervention. In recent years, deep learning has demonstrated notable advantages in computer vision and agricultural applications. *Yuan et al., (2022),* developed an early bruise detection system for apples by combining near-infrared imaging with deep learning, achieving an accuracy rate exceeding 99%. The powerful learning and inference capabilities of deep learning give it a clear edge in fruit damage detection tasks, enabling more accurate and efficient automated detection.

Based on the analysis of the aforementioned detection methods and deep learning techniques, this paper proposes a method for early bruise detection in 'Huangguan' pears that integrates near-infrared imaging technology with deep learning algorithms. DeepLabV3+ *(Chen et al., 2018),* as a basic framework for segmentation model is employed for detecting damaged regions in near-infrared images of 'Huangguan' pears, with MobileNetV2 *(Sandler et al., 2018)* chosen as the backbone network to reduce the model's parameter size. Additionally, the coordinate attention (CA) *(Hou et al., 2021)* mechanism is introduced at both shallow and deep feature levels within the encoder to enhance the model's ability to detect bruise regions and improve segmentation accuracy. Furthermore, the cascade feature fusion (CFF) *(Zhao et al., 2018)* strategy is incorporated to better integrate shallow and deep semantic features, thereby enhancing the continuity of edge detection. Experimental results demonstrate that the improved model achieves early bruise detection in 'Huangguan' pears with a model size of only 6.103MB, while key metrics, such as mIoU and mPrecision, reach 95.68% and 97.58%, respectively, significantly improving both detection efficiency and accuracy.

MATERIALS AND METHODS

Experimental samples and damage device

The skin colour of 'Huangguan' pears closely resembles the colour of bruising in its early stages, making bruise identification particularly challenging. Bruise samples of 'Huangguan' pear were prepared as shown in Figure 1.

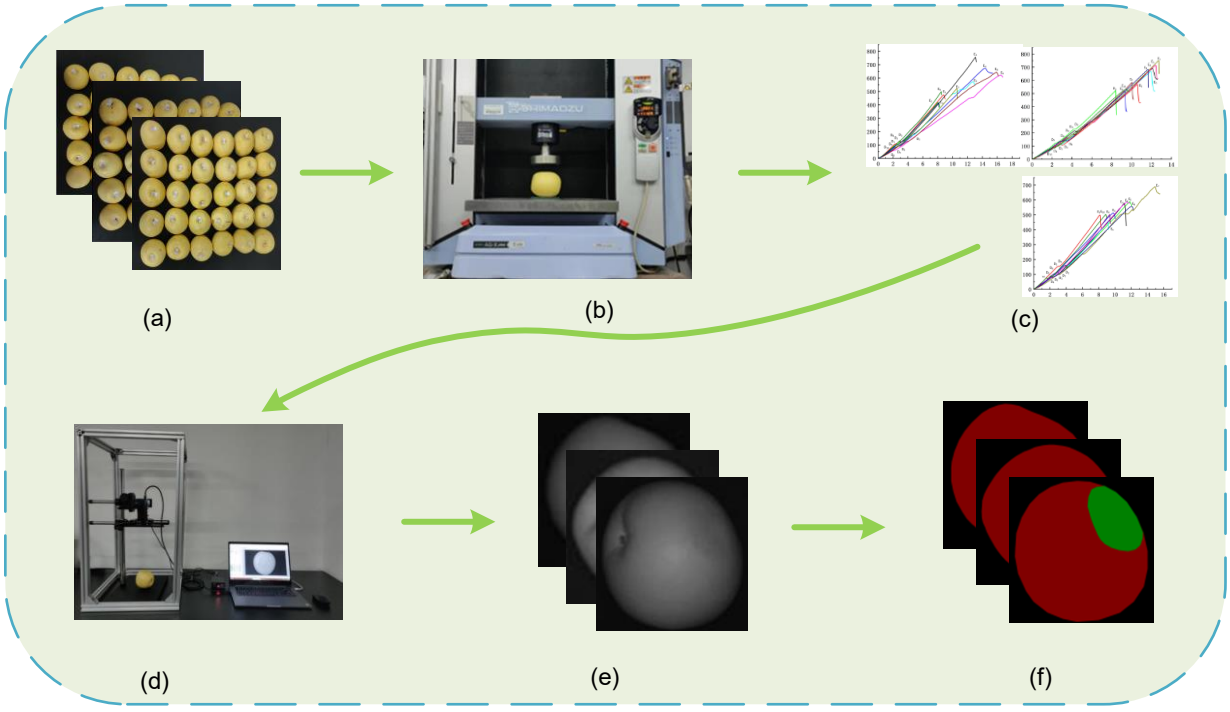


Fig. 1 - Process of bruise sample and data preparation

A total of 120 'Huangguan' pears, free of damage and disease, were purchased from a supermarket and stored at room temperature. These pears were categorized into three groups based on size: large, medium, and small, as shown in Figure 1(a). The diameters of the large pears ranged from 90 to 100 mm, medium pears from 80 to 90 mm, and small pears from 70 to 80 mm.

Since early-stage bruises on 'Huangguan' pears are not visible to the naked eye, it is difficult to obtain naturally bruised pears. Therefore, the bruised pear samples were artificially created. To simulate various levels of compressive bruising, a universal testing machine was used to perform mechanical compression tests, as shown in Figure 1(b). First, ten pears from each size group were randomly selected for preliminary compression tests to determine their biological yield and rupture points, providing a reference for pressure control during the bruise sampling process. The results, shown in Figure 1(c), indicated that the rupture force range for large pears was between 515.73 N and 749.91 N, for medium pears between 457.87 N and 755.73 N, and for small pears between 414.91 N and 689.29 N. Then, it was determined that the maximum lower bound of the rupture force range was 515.73 N, and the minimum upper bound was 689.29 N. For the subsequent bruise sampling, 90 pears were subjected to forces ranging from 520 to 680 N. To simulate varying degrees of internal bruising that 'Huangguan' pears may experience during harvesting and transport, a universal material testing machine applied pressure to the equatorial region of the pears. The pressure was initially set at 520 N and incrementally increased by 10 N until it reached 680 N.

Near infrared camera imaging system and image acquisition process

A NIR camera employed as the imaging device in this experiment is the IUA4100KPA model, manufactured by Hangzhou Jiecheng Instrument Co., Ltd., which features a Sony sensor with a resolution of 2688 × 1520 pixels and a maximum frame rate of 90 frames per second. The lens used is a C-M0418IR (3MP) high-definition industrial lens produced by FordTech, ensuring high-resolution, high-quality image capture suitable for precise detection of surface bruises on 'Huangguan' pears. The imaging system, shown in Figure 1(d), comprises a ring light source, a dark box, and a stand, providing stable and uniform illumination. The camera and lens are positioned at the centre of the ring light source, with the samples placed on a black background plate. The height and aperture of the lens were adjusted for optimal imaging conditions, and the camera's aperture and focal length were fine-tuned to ensure clarity and high resolution. This imaging system minimizes environmental interference and provides consistent, accurate image acquisition, laying a robust foundation for subsequent image processing and analysis.

Immediately after pressure application and bruise induction, NIR images were captured and subsequently collected at 10 min, 20 min, and 0.5 hour intervals to enrich the sample set. To ensure the diversity and comprehensiveness of the dataset, images were taken from multiple angles and positions. After removing unclear or substandard images, a final dataset consisting of 2700 high-quality NIR images of bruised 'Huangguan' pears was obtained, as shown in Figure 1(e). These images were annotated using LabelImg software, where the intact regions of the pears were marked in red, and the bruised regions were highlighted in green. A visualization of the labelled images is shown in Figure 1(f). The dataset was ultimately divided into training, validation, and test sets in an 8:1:1 ratio, comprising 2160, 270, and 270 images, respectively.

DESIGN OF BRUISE SEGMENTATION MODEL

DeepLabV3+ is a state-of-the-art deep learning-based semantic segmentation model designed to address the challenges of spatial information loss and low-resolution outputs commonly encountered in traditional convolutional neural networks for segmentation tasks. Its key innovations include the incorporation of atrous convolution and encoder-decoder architecture, which significantly enhance feature extraction capabilities and improve the resolution of segmentation results. In the model architecture, the input image is first processed by the backbone network, Xception, which extracts hierarchical features: shallow features that retain spatial details and deep features that encode rich contextual semantics. The deep features are then refined through the atrous spatial pyramid pooling (ASPP) module, which employs atrous convolutions with varying dilation rates and global pooling to effectively capture multi-scale contextual information and enrich feature representation. During the decoding stage, low-resolution deep feature maps are upsampled and fused with high-resolution shallow feature maps. This fusion process, followed by additional convolutional operations, restores the spatial resolution of the feature maps to match that of the input image, ultimately producing precise and high-quality segmentation outputs.

Although DeepLabV3+ demonstrates strong performance in multi-scale feature processing and output resolution enhancement, it exhibits certain limitations when applied to the segmentation of early bruise regions in near-infrared images of 'Huangguan' pear. The early bruise regions have minimal visual differences compared to healthy areas, particularly at the boundaries, making them challenging to distinguish, which leads to segmentation discontinuities and missed detections around these boundaries.

Furthermore, the high computational complexity of Xception as the backbone network restricts the model's scalability and efficiency in processing large-scale mass near-infrared datasets. To address these challenges, this paper proposes an improved model, MCC-DeepLabV3+. First, the original Xception backbone is replaced with MobileNetV2, which significantly reduces both the parameter count and computational complexity while retaining robust feature extraction capabilities. Second, a coordinate attention (CA) mechanism is integrated into the shallow features and ASPP modules of the backbone network. This enhancement improves the model's ability to capture spatial information at multiple feature levels, thereby increasing segmentation accuracy and mitigating missed detections. Finally, a cascaded feature fusion (CFF) strategy is introduced during the fusion of shallow and deep features. This strategy ensures a more effective integration of multi-level semantic information, addressing edge discontinuities and further improving the overall segmentation quality. The architecture of the improved model is illustrated in Figure 2.

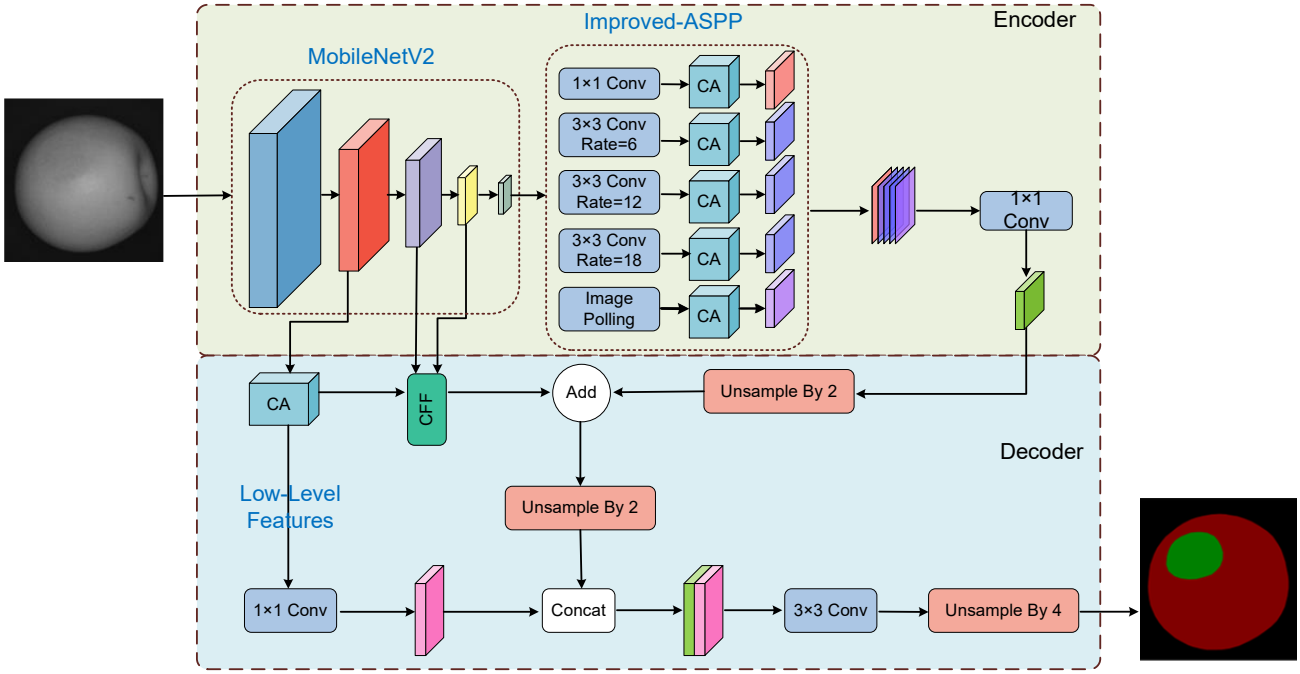
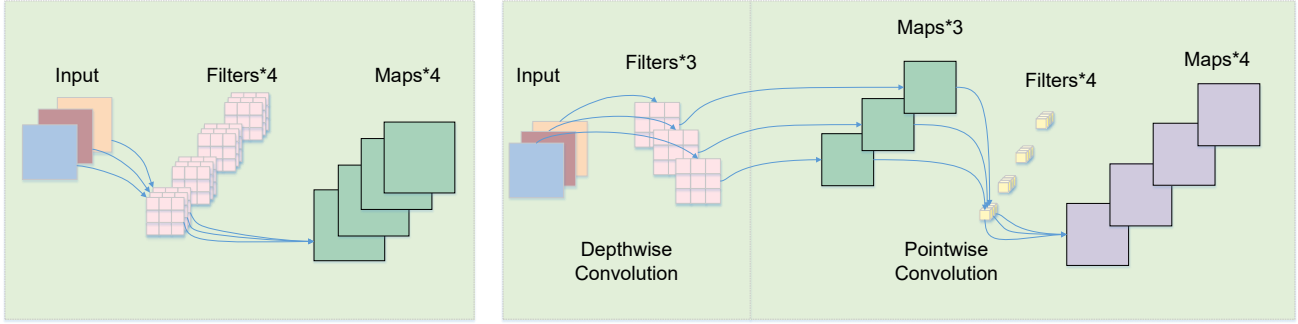


Fig. 2 - The structure of the improved DeepLabV3+ model for bruise 'Huangguan' pear detection

Lightweight backbone network

The Xception network, employed in DeepLabV3+, possesses robust feature extraction capabilities. However, its architecture, composed of numerous stacked convolutional layers, involves extensive multi-level computations, leading to a high parameter count and significant computational complexity. This computational burden presents a bottleneck for the early bruise segmentation task of 'Huangguan' pears, where rapid processing of large-scale near-infrared images is required. To address this issue, the MCC-DeepLabV3+ model replaces Xception with MobileNetV2. MobileNetV2 is a lightweight network architecture that leverages depthwise separable convolutions instead of standard convolutions, significantly reducing the parameter count and computational complexity. Additionally, its inverted residual structure minimizes feature redundancy while maintaining high representational power, enabling efficient feature extraction with a low computational cost.



(a) standard convolution *(b) depth separable convolution*

Fig. 3 - Structures of standard convolution and depth-separable convolution

Depthwise separable convolution is the core operation in MobileNetV2 for achieving model lightweighting. It decomposes the standard convolution operation into two independent steps: depthwise convolution (DW) and pointwise convolution (PW), effectively reducing the computational load of the model. As shown in Figure 3, compared with standard convolution, depthwise separable convolution is divided into two parts: first, depthwise convolution performs convolution independently on each input channel to extract spatial features; then, pointwise convolution applies 1×1 convolutions to perform point-by-point operations, linearly combining features from different channels and fusing spatial and channel information. This decomposition significantly reduces the computational complexity.

Specifically, as shown in Figure 3, depthwise convolution (DW) extracts spatial features by independently processing each input channel, while pointwise convolution (PW) fuses spatial and channel information through 1×1 convolutions. The computational complexity of depthwise separable convolution can be represented by the following two formulas:

$$C_{DW} = M \times D_F \times D_F \times D_K \times D_K \quad (1)$$

$$C_{PW} = M \times N \times D_F \times D_F \quad (2)$$

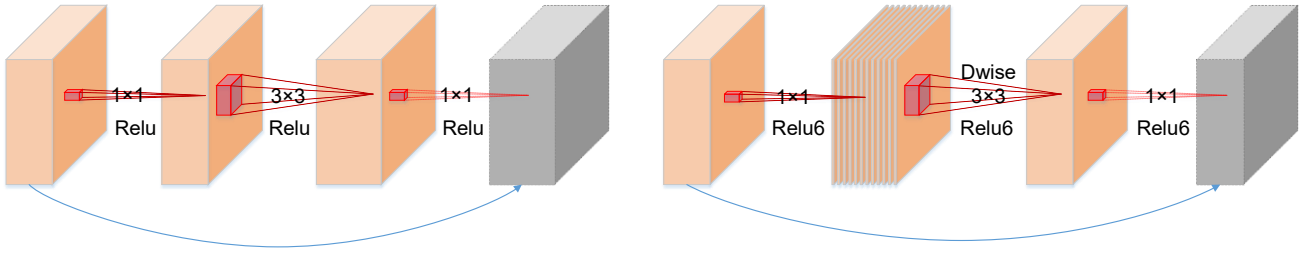
where D_F is the dimension of the feature map, D_K represents the size of the convolution kernel, M is the number of input channels, and N is the number of output channels.

Under equations (1) and (2), the total computation of the depth separable convolution can be expressed as Eq.(3). In contrast, the standard convolution is computationally intensive, which is calculated as Eq.(4). Comparing Eq. 3 and Eq. 4, it can be seen that the computational complexity of the depth-separable convolution is reduced to $\frac{1}{N}+\frac{1}{D_K^2}$ of the standard convolution, which allows MobileNetV2 to maintain a high feature extraction capability while reducing the computational effort.

$$C_{DS} = C_{DW} + C_{PW} = M \times D_F \times D_F \times D_K \times D_K + M \times N \times D_F \times D_F \quad (3)$$

$$C_{SC} = M \times N \times D_F \times D_F \times D_K \times D_K \quad (4)$$

In addition, MobileNetV2 introduces a novel inverted residual block, optimized based on traditional residual networks. While traditional residual networks mitigate the vanishing gradient problem through skip connections, the inverted residual block enhances computational efficiency and feature representation through a “dimensionality enhancement-extraction-dimensionality reduction” design, as illustrated in Figure 4. Specifically, the 1×1 convolution is used to reduce dimensions initially, followed by 3×3 depthwise separable convolution for feature extraction, and finally, another 1×1 convolution to restore the feature dimensions. The use of a linear activation function Relu6 helps to minimize feature loss during dimensionality reduction and improves the efficiency of information flow. Thus, by introducing deeply separable convolutions and inverted residual structures, MobileNetV2 reduces the model’s parameter count while maintaining strong feature extraction capabilities.



(a) Residual network **(b) Inverted residual network**

Fig. 4 - The structure of inverted residual network

Introduction of CA attention mechanism

In semantic segmentation tasks, the traditional DeepLabV3+ model relies on shallow features to provide spatial information and utilizes the ASPP module to extract multi-scale contextual information. However, this architecture exhibits certain limitations. Firstly, shallow features are different to capture the global positional information of bruise regions due to their small receptive fields. Secondly, the ASPP module demonstrates relatively weak capability in capturing spatial location information, although it focuses on aggregating cross-scale semantic features.

These challenges are exacerbated in the segmentation of early bruises in 'Huangguan' pears, where the subtle visual differences between healthy and bruised regions, combined with blurred boundaries, significantly increase the segmentation difficulty. To address these issues, this paper introduces the coordinate attention (CA) mechanism into the shallow feature and ASPP modules of the MCC-DeepLabV3+ model to enhance their spatial localization capabilities. In the shallow feature extraction stage, the CA mechanism strengthens the representation of local spatial details in bruise regions by leveraging cross-channel information modeling and orientation-awareness properties. In the ASPP module, the CA mechanism compensates for its inherent limitations in capturing spatial location information, enabling the model to more accurately perceive the boundary positions of bruise regions. This enhancement effectively improves the spatial localization capability of the model, thereby increasing the accuracy and continuity of boundary segmentation for early bruise regions in 'Huangguan' pears. Therefore, the introduction of the coordinate attention (CA) mechanism enables the encoding of channel relationships and remote dependencies based on precise location information, and the structure is shown in Figure 5.

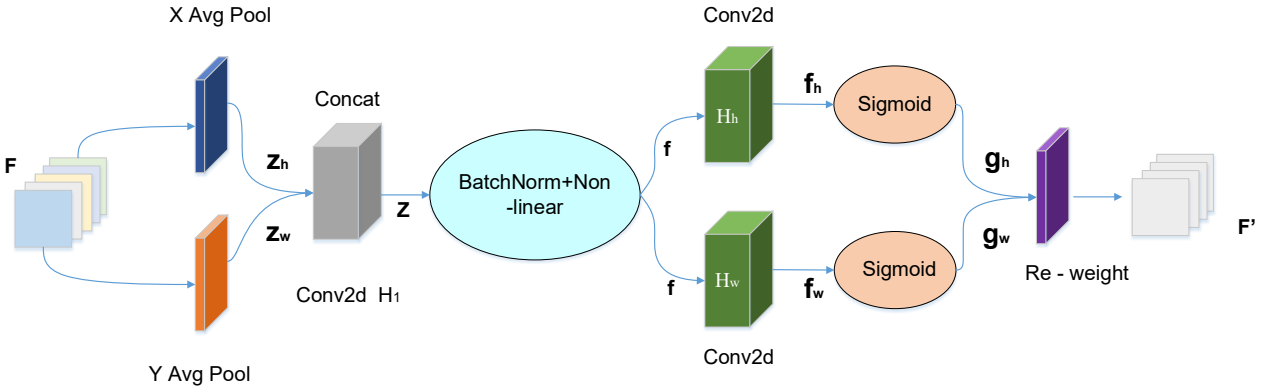


Fig. 5 - Structure of the coordinate attention mechanism module

From Figure 5, to globally encode the height H and width W directions of the input feature map **F**, the CA mechanism implements a global average pooling operation along the H and W directions, respectively, which generates the corresponding two 1D feature maps $\mathbf{z_h}$ and $\mathbf{z_w}$, as follows:

$$\mathbf{z_h}(h)=\frac{1}{W}\sum_{i=0}^{W}x_c(h,i)\ ,\ \mathbf{z_w}(w)=\frac{1}{H}\sum_{j=0}^{H}x_c(j,w) \tag{5}$$

where $\mathbf{z_h}$ and $\mathbf{z_w}$ denote the aggregation results of the feature map in the H and W directions, respectively. The $\mathbf{z_h}$ and $\mathbf{z_w}$ obtained above are spliced and downscaled by a 1×1 Conv H_1 operation to generate an intermediate feature map **Z** of size $C/r\times 1\times(W+H)$, where r is the channel reduction ratio. Next, batch normalization and nonlinear activation operations are applied to **Z** to obtain the processed feature map **f**, where δ denotes the nonlinear activation operation.

$$\mathbf{Z}=\mathrm{H}_1\left[\mathbf{z_h},\mathbf{z_w}\right] \tag{6}$$

$$\mathbf{f}=\delta\left[\mathrm{BN}(\mathbf{Z})\right] \tag{7}$$

Then, **f** is split into $\mathbf{f_h}$ and $\mathbf{f_w}$ in both H and W directions and the attention weights $\mathbf{g_h}$ and $\mathbf{g_w}$ are computed by 1×1 Conv H_h, H_w operations and Sigmoid activation function, respectively:

$$\mathbf{g_h}=\sigma\left[\mathrm{H_h}(\mathbf{f}_h)\right]\ ,\ \mathbf{g_w}=\sigma\left[\mathrm{H_w}(\mathbf{f_w})\right] \tag{8}$$

where σ denotes the Sigmoid function.

Finally, the input feature maps are multiplied with the generated attention weights $\mathbf{g_h}$ and $\mathbf{g_w}$ in the X and Y directions, respectively, to obtain the final feature maps after fusing the attentional information:

$$\mathbf{F}^{'}=\mathbf{F}\otimes\mathbf{g_h}\otimes\mathbf{g_w} \tag{9}$$

Integration of CFF module

In DeepLabV3+, the backbone network generates features at varying levels of abstraction during feature extraction. Shallow features are particularly effective for pixel-level localization tasks and ability to excel at capturing edge details and local information due to their high resolution. In contrast, deep features are progressively abstracted and aggregated through a multi-layer network, encapsulating rich global semantic information, which helps in understanding overall image context and category differentiation. Effective fusion of shallow and deep features is the key to achieving a more comprehensive and robust feature representation. However, the traditional DeepLabV3+ decoder employs a simple concatenation method for feature fusion, which inadequately addresses the distinct characteristics and synergistic relationships between shallow and deep features. This limitation can result in the loss of critical detail information, particularly in challenging scenarios such as segmenting bruised areas with blurred boundaries. Consequently, issues such as edge discontinuities and breaks may arise, significantly compromising segmentation accuracy.

To address these challenges, this study introduces the cascaded feature fusion (CFF) module to more effectively integrate shallow and deep features. The core processes of the CFF module are illustrated in Figure 6. First, the shallow feature map $\mathbf{F_1}$ undergoes a 3×3 dilated convolution with a dilation rate of 2 to adjust the channel dimensions. Simultaneously, the deep feature map $\mathbf{F_2}$ is upsampled by a factor of two using bilinear interpolation, followed by the same dilated convolution operation to align its dimensions and receptive field with $\mathbf{F_1}$. This approach not only ensures feature resolution consistency but also expands the receptive field, enabling the capture of richer contextual information while reducing computational overhead. Next, batch normalization is applied to both feature maps to standardize data distribution, which accelerates network training and enhances model convergence. Finally, the normalized $\mathbf{F_1}$ and $\mathbf{F_2}$ are element-wise added, and the result is processed through a ReLU activation function to produce the fused feature map, denoted as $\mathbf{F_{12}}$.

The integration of the CFF module enables a more effective combination of the edge detail information from shallow features and the global semantic information from deep features. This enhancement significantly improves segmentation accuracy for bruised areas while ensuring smoother and more continuous boundary delineation. Moreover, the optimized design of the CFF module reduces computational costs, thereby improving overall segmentation efficiency.

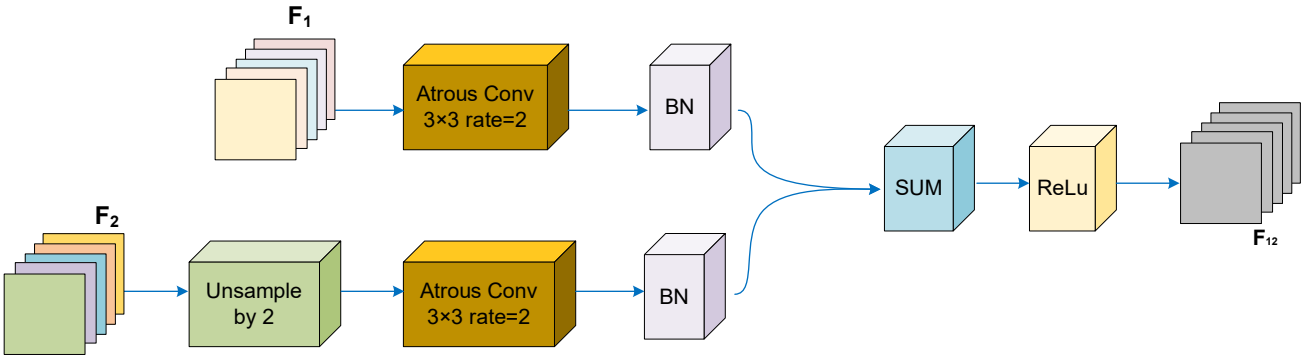


Fig. 6 - Structure of the cascaded feature fusion module

RESULTS AND ANALYSIS

Experimental setup and evaluation metrics

The experiment uses Windows11 operating system. The hardware platform employs Intel I7-11700K CPU and NVIDIA GeForce RTX 4080 GPU with 32G video memory. Network construction, debugging, training and testing are carried out under the Pytorch framework based on the Python language, in which the torch version is 1.13 and the CUDA version is 12.1. The size of the input image is set to 512×512, and the batch size and the number of iterations are set to 2 and 300, respectively, for the training process. In the experiment, the initial learning rate is set to 0.01, and the model is optimized by using SGD. The network is initialized by loading pre-training weights.

For quantitatively comparing the performance of semantic segmentation network models in detecting bruises on 'Huangguan' pears, four evaluation metrics were used: mean Intersection over Union (mIoU), mean Pixel Accuracy (mPA), mean Precison (mPrecision) and mean Recall (mRecall). mIoU is the ratio of the intersection to the union of the ground truth and predicted labels, averaged over all classes. mPA represents

the proportion of correctly classified samples, providing an intuitive measure of overall performance. mPrecision refers to the proportion of true position samples among those predicted as positive by the model, which serves as an evaluation of the model's accuracy in predicting positive cases. mRecall measures the model's ability to correctly identify positive samples, indicating the proportion of true positive samples correctly predicted from all actual positive samples. The four evaluation metrics are defined as follows:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN} \tag{10}$$

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$mPecision = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FP} \tag{12}$$

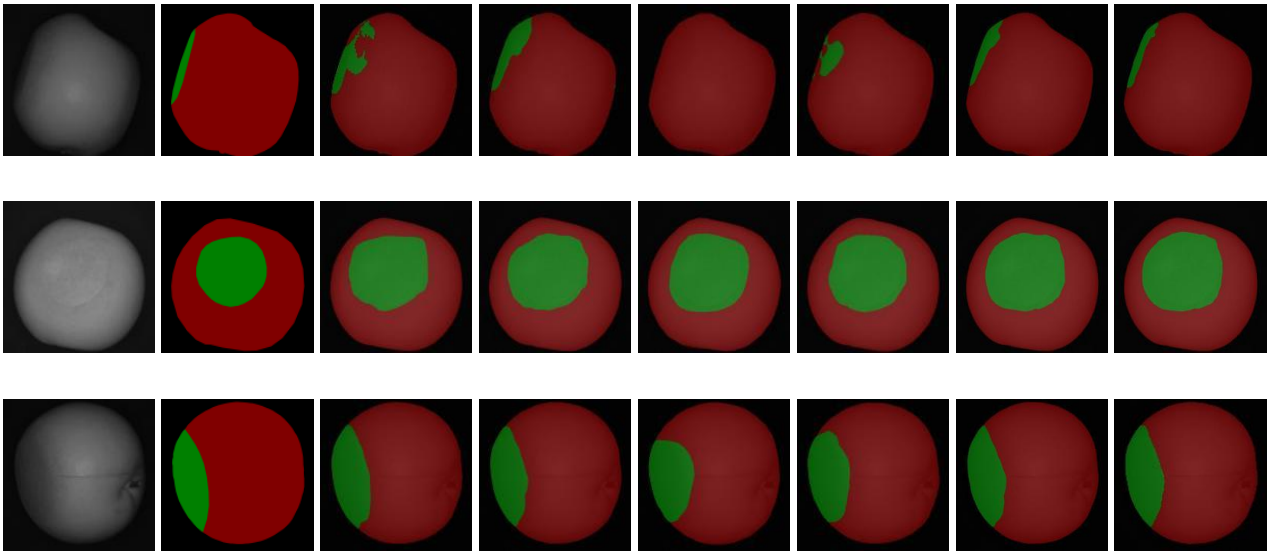
$$mRecall = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FN} \tag{13}$$

In the above formulas, TP (True Positive) refers to cases where the prediction is positive and the ground truth is also positive. FP (False Positive) denotes situations where the prediction is positive but the ground truth is negative. FN (False Negative) refers to instances where the prediction is negative but the ground truth is positive. TN (True Negative) indicates cases where both the prediction and the ground truth are negative.

Comparison experiment of different models

To verify the effectiveness of the proposed MCC-DeepLabV3+ model in early bruise segmentation of 'Huangguan' pears, it is compared with five classical segmentation models: UNet (*Ronneberger et al., 2015*), SegNet (*Badrinarayanan et al., 2016*), PSPNet (*Zhao et al., 2017*), HrNet (*Sun et al., 2019*), and DeepLabV3+. The experiments were conducted using the same parameter settings and test conditions, and the performance of each model in detecting bruises was visualized through segmentation results.

Figure 7 illustrates the segmentation performance of different models on randomly selected test data.



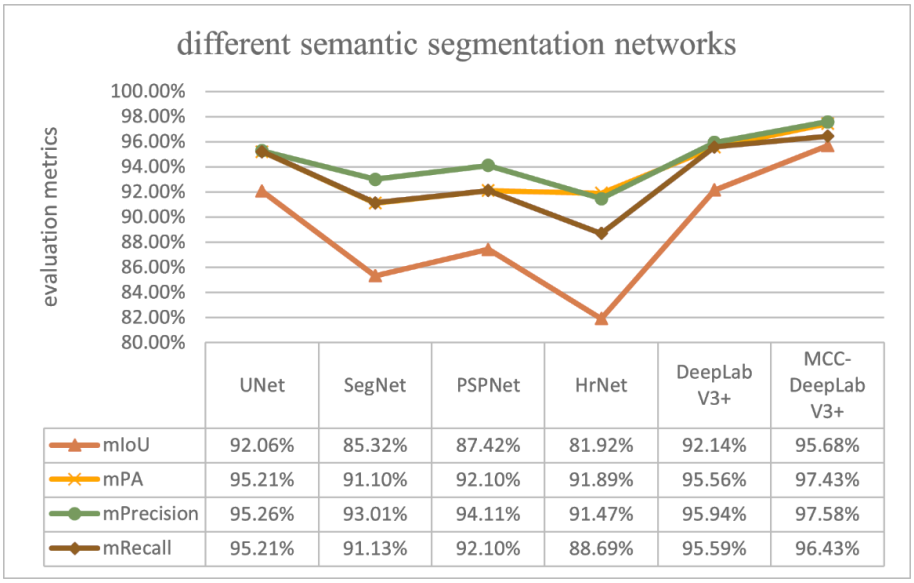
(a) Input *(b) Groud-truth* *(c) UNet* *(d) SegNet* *(e) PSPNet* *(f) HrNnet* *(g)DeepLab V3+* *(h) Ours*

Fig. 7 - Comparison of segmentation results of different networks

UNet performs poorly in locating bruise regions. In UNet, the upsampling process often introduces noise and skip connections that tend to mix high-level features leads to frequent misclassification of healthy tissue as bruise regions. SegNet suffers from significant loss of spatial information due to multiple pooling operations during the decoding stage, especially when precisely segmenting boundary regions. PSPNet is prone to missing bruise features, indicating insufficient sensitivity to the characteristics of bruise regions. HrNet retains high-resolution feature maps through its multi-resolution parallel network branches, capturing global context information.

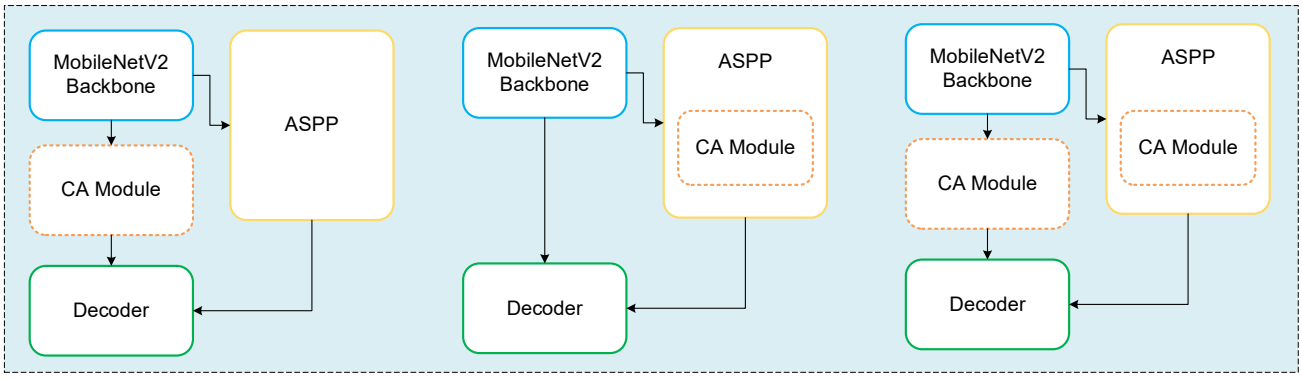
However, this design may overlook local details, especially when the contrast between the background and bruise areas is low, leading to misclassification of the background as bruise regions. Among the five models, DeepLabV3+ achieves relatively better performance. However, its spatial convolutions have limited capacity to preserve spatial location information, leading to some healthy regions being misclassified as bruise areas. In Figure 7(h), the proposed MCC-DeepLabV3+ model demonstrates more precise bruise segmentation compared to the other five models, significantly reducing misclassification and omission errors. It closely aligns with the ground truth shown in Figure 7(b), delivering more reliable and accurate segmentation results.

To further validate the performance of the proposed MCC-DeepLabV3+ model, mIoU, mPrecision, and mRecall metrics were evaluated on the 'Huangguan' pear bruise dataset, with results summarized in Figure 8. MCC-DeepLabV3+ achieved superior performance across all metrics, with mIoU, mPA, mPrecision, and mRecall reaching 95.68%, 97.43%, 97.58%, and 96.43%, respectively, outperforming all comparison models. For instance, in terms of mIoU, the proposed model demonstrated improvements of 3.93%, 12.14%, 9.45%, 16.80%, and 3.84% over UNet, SegNet, PSPNet, HrNet, and DeepLabV3+, respectively. These results show that the proposed model's robustness and effectiveness in achieving high segmentation accuracy and accurately detecting bruise regions. The above qualitative and quantitative comparison results show that MCC-DeepLabV3+ performs excellently in the task of early bruise segmentation in 'Huangguan' and pear. It effectively addresses the limitations of classical segmentation models, such as inaccurate bruise localization, insufficient boundary smoothing and incorrect segmentation. With higher segmentation accuracy and enhanced robustness, the model provides a practical and efficient solution for bruise detection in industrial applications.



	UNet	SegNet	PSPNet	HrNet	DeepLab V3+	MCC-DeepLab V3+
mIoU	92.06%	85.32%	87.42%	81.92%	92.14%	95.68%
mPA	95.21%	91.10%	92.10%	91.89%	95.56%	97.43%
mPrecision	95.26%	93.01%	94.11%	91.47%	95.94%	97.58%
mRecall	95.21%	91.13%	92.10%	88.69%	95.59%	96.43%

Fig. 6 - Comparison of segmentation performance of different semantic segmentation networks



Structure 1: CA after MobileNetV2

Structure 2: CA in ASPP

Structure 3: CA after MobileNetV2 and in ASPP

Fig. 7 - The structure of CA attention mechanisms in different positions

To determine the optimal application of the attention mechanism in DeepLabV3+, three different encoder architectures shown in Figure 9 were designed and evaluated on the custom 'Huangguan' Pear dataset. These architectures include: incorporating the attention mechanism only after the shallow feature output from MobileNetV2 (referred to as "CA after MobileNetV2"), integrating it solely within the ASPP module ("CA in ASPP"), and applying it to both MobileNetV2 output and the ASPP module ("CA after MobileNetV2 and in ASPP"). Figure 10 illustrates the segmentation results of the three structures, while Table 1 presents their performance metrics.

As shown in Table 1, Structure 3 achieved the highest performance across all metrics, with mIoU, mPA, mPrecision, and mRecall reaching 95.22%, 96.93%, 96.41%, and 95.57%, respectively. Specifically, as depicted in Figure 10, structure 1 enhances the extraction of local features but lacks reinforcement of deep semantic features, leading to inadequate global information capture. This limitation reduces segmentation accuracy, particularly for irregular bruise morphologies, which are prone to misclassification. Structure 2 (in ASPP) improves global contextual perception and optimizes the receptive field across multi-level features, enabling more accurate contour detection. However, it struggles with fine boundary delineation. In contrast, structure 3 combines the strengths of both approaches. The attention mechanism applied after MobileNetV2 refines edge features and improves detail segmentation, while the mechanism in the ASPP module strengthens global semantic feature representation and ensures effective spatial information capture. This dual-attention mechanism achieves a balanced integration of global and local features, significantly enhancing overall segmentation performance and edge detail accuracy. By combining subjective visual evaluations with objective performance metrics, it is evident that the dual-attention mechanism substantially improves the performance of DeepLabV3+ in 'Huangguan' pear bruise segmentation. This approach not only preserves the integrity of global semantic features but also restores fine edge details, thereby enhancing both segmentation accuracy and model robustness. Therefore, it was selected as the final design in this study.

Table 1

Comparison of network segmentation performance after adding attention mechanism at different locations

Network type	mIoU	mPA	mPrecision	mRecall
Structure 1	92.06%	95.21%	95.26%	93.21%
Structure 2	92.14%	95.56%	95.94%	94.59%
Structure 3	95.22%	96.93%	96.41%	95.57%

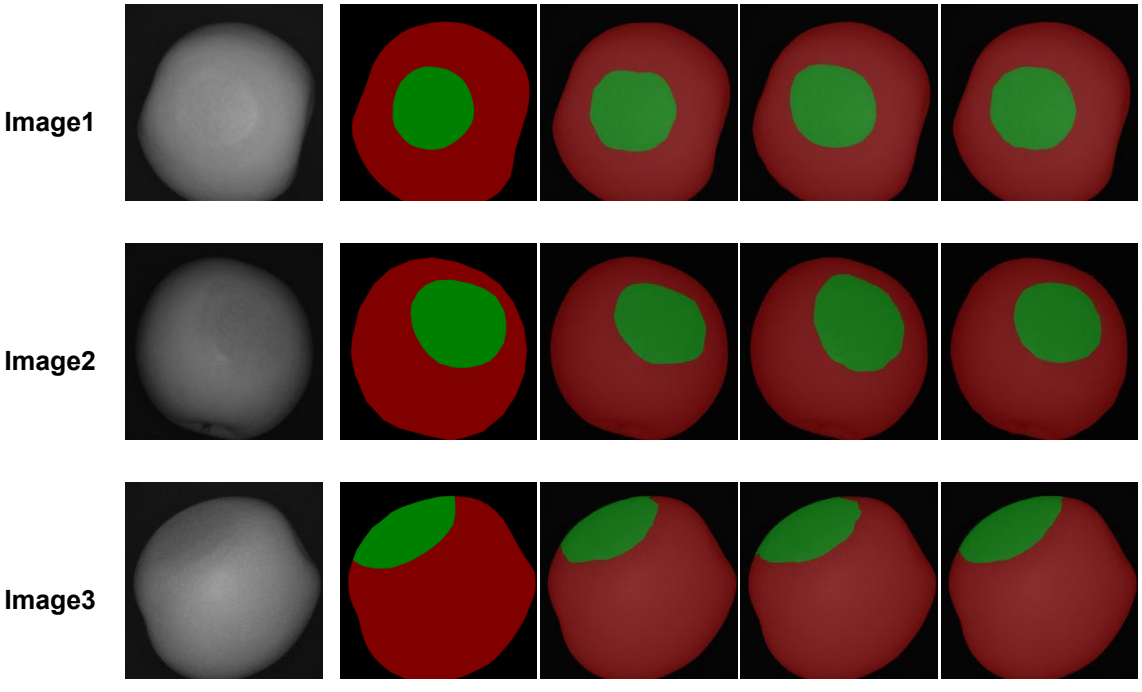


Fig. 8 - Comparison of segmentation results after adding attention mechanism at different locations

Ablation experiment

To verify the contribution of each module in 'Huangguan' pear bruise detection, this study conducts ablation experiments using a self-constructed dataset, with DeepLabV3+ as the base framework. By comparing the effects of embedding the MobileNetV2 backbone network, the Coordinate Attention (CA) mechanism, and the Cascaded Feature Fusion (CFF) module, the effectiveness and individual contributions of these modules are thoroughly evaluated. The test effect graphs and performance metrics for different module combinations are shown in Figure 11and Table 2, respectively.

Firstly, the backbone of DeepLabV3+ was replaced from Xception to MobileNetV2, reducing the model parameters from 54.71M to 5.82M, nearly a tenfold decrease. This number of parameters has decreased, but the mIoU, mPA, mPrecision, and mRecall metrics improved from 91.24% to 93.52%, 95.56% to 96.55%, 95.94% to 96.56%, and 95.59% to 96.55%, respectively, demonstrating the effectiveness of the lightweight design. The efficient depthwise separable convolutions and inverted residual structure of MobileNetV2 significantly reduce the parameter count while maintaining robust feature extraction capabilities, offering a practical solution for resource-constrained industrial applications. Next, the CA mechanism was embedded in both the output of MobileNetV2 and the ASPP module, keeping the model parameters unchanged at 5.82M. However, the performance improved further, with mIoU, mPA, mPrecision, and mRecall reaching 93.54%, 96.59%, 96.64%, and 96.57%, respectively. The CA mechanism effectively enhances the model's ability to capture global spatial information by combining spatial coordinates with channel attention. Specifically, it strengthens long-range dependencies through global average pooling in the horizontal and vertical directions, improving the model's capability to handle complex bruise shapes while significantly reducing omissions and misclassifications. This demonstrates the critical role of the CA mechanism in enhancing the spatial feature representation of the model. Finally, the introduction of the CFF module slightly increased the model parameters to 5.87M but improved performance. Compared to the initial network with Xception as the backbone, the mIoU, mPA, mPrecision, and mRecall improved by 1.65%, 1.10%, 0.77%, and 1.07%, respectively. The CFF module employs a cascaded feature fusion strategy, combining fine-grained details from shallow features with global semantics from deep features, enhancing the model's ability to capture multi-level features, improving segmentation accuracy.

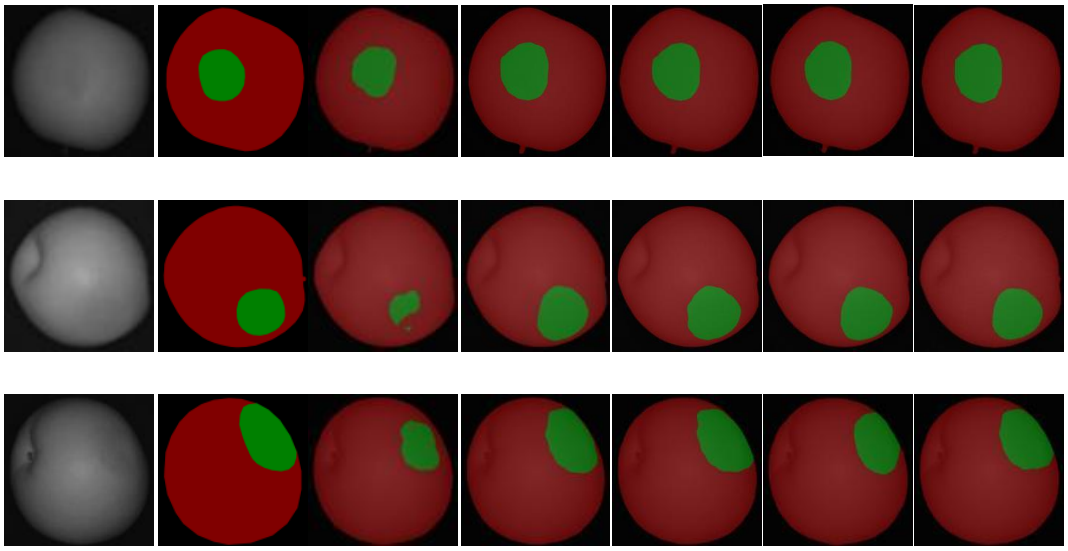
Table 2

The segmentation performance comparison of different modules

DeepLabV3+	Backbone		CA	CFF	Params (M)	mIoU (%)	mPA (%)	mPrecision (%)	mRecall (%)
	Xception	MobileNetV2							
√	√				54.71	92.14	95.56	95.94	95.59
√		√			5.82	93.52	96.55	96.56	96.55
√		√	√		5.82	93.54	96.59	96.64	96.57
√		√		√	5.87	93.66	96.61	96.68	96.61
√		√	√	√	6.10	95.68	97.43	97.58	97.43

Figure 11 illustrates the segmentation performance of the model on the test set after the progressive addition of different modules. Comparing Figure 11(c) and Figure 11(d), it can be observed that replacing the backbone network with MobileNetV2 resulted in clearer segmentation of bruise regions, enhancing the model's initial segmentation capability. Further comparison between Figure 11(d) and Figure 11(e) shows that the introduction of the CA attention mechanism significantly improved the model's feature extraction capability, leading to more precise segmentation and clearer delineation of bruise contours. Comparing Figure 11(e) and Figure 11(f), the inclusion of the CFF module enhanced the handling of edge details in early bruise region segmentation, effectively avoiding issues such as edge blurring or irregularities. Overall, the segmentation results of the proposed model (Figure 11(g)) are more closely aligned with the ground truth (Figure 11(b)), demonstrating the effectiveness of the combined modules in improving segmentation accuracy and detail preservation.

The above objective metrics and subjective segmentation results from the ablation experiments demonstrate that the MobileNetV2 backbone significantly reduced the model's parameters, improving segmentation efficiency. The CA mechanism enhanced the model's ability to represent spatial positional information, resulting in more accurate segmentation. Meanwhile, the CFF module, by integrating multi-level features, further improved the localization accuracy and detail preservation in bruise regions. Together, these three modules achieve a balance between lightweight design and high performance.



(a)Input	(b) Groud-truth	(c)Xception	(d) MobileNet V2	(e)MobileNetV2+CA	(f) MobileNet V2+CFF	(g) Ours

Fig. 9 - Comparison of segmentation results of different modules

CONCLUSIONS

This paper introduces a lightweight semantic segmentation network, MCC-DeepLabV3+, designed for early bruise detection in ‘Huangguan’ pears using near-infrared (NIR) imaging technology and deep learning techniques. A universal testing machine was employed to generate real early bruises, and a corresponding dataset of early bruising in ‘Huangguan’ pears was prepared. The proposed network adopts MobileNetV2 as the backbone, which effectively reduces the parameter size of the original Xception model, thereby addressing the issue of model complexity. Furthermore, a Coordinate Attention (CA) mechanism is integrated into both the shallow feature extraction stage and the Atrous Spatial Pyramid Pooling (ASPP) module, significantly enhancing the model’s segmentation performance and improving its ability to process edge details. The introduction of the Cascaded Feature Fusion (CFF) module further optimizes the integration of shallow detail features with deep semantic information, which is crucial for improving accuracy in detecting early bruises.

To assess the performance of the proposed model, a series of comparative experiments with classical segmentation networks such as UNet, SegNet, PSPNet, HrNet, and DeepLabV3+ was conducted. The evaluation focused on key metrics, including mIoU, mPA, mPrecision, and mRecall, as well as a detailed analysis of the segmentation results on the test set. Our findings reveal that MCC-DeepLabV3+ consistently outperforms the other models across all evaluated indices, demonstrating superior segmentation reliability and performance in detecting early bruises in ‘Huangguan’ pears.

In terms of attention mechanisms, the impact of adding CA attention mechanisms at various stages of the network, specifically at the shallow input of the MobileNetV2 backbone and within the ASPP module, was explored. Our results indicate that incorporating CA attention mechanisms at both locations simultaneously significantly improves model performance, compared to models with either individual or no attention mechanisms. This reinforces the importance of integrating attention mechanisms at multiple stages of the network to enhance segmentation results.

To further validate the effectiveness of the proposed enhancements, ablation experiments were conducted, which confirmed that MCC-DeepLabV3+ outperforms various combinations of individual improvements.

The model's segmentation results closely align with the ground truth labels, showcasing its ability to accurately detect early bruises in 'Huangguan' pears. This advancement provides a new, efficient, and precise approach for early bruise detection, which can be effectively applied in automated fruit quality detection and precision agriculture.

In conclusion, the enhanced MCC-DeepLabV3+ network introduced in this paper offers significant improvements in the accuracy and efficiency of early bruise detection in 'Huangguan' pears. While the proposed model demonstrates impressive performance, there remains potential for further optimization. Future research will focus on refining both the accuracy and speed of detection to further enhance the practicality and effectiveness of this approach for real-world applications in agriculture. Additionally, addressing challenges such as the detection of complete overlap between bruises in densely packed pears will be a key area for future improvement.

CREDIT AUTHOR STATEMENT

Congkuan Yan: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Haonan Zhao**: Formal Analysis, Data curation. **Dequan Zhu:** Supervision, Writing – review & editing. **Yuqing Yang**: Formal Analysis. **Ruixing Xing**: Data curation. **Qixing Tang**: Investigation, Validation. **Juan Liao**: Project administration, Funding acquisition, Writing – review & editing.

DECLARATION OF COMPETING INTEREST

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

ACKNOWLEDGEMENT

The research was funded by the National Natural Science Foundation of China (No.32201665) and the National Key Research and Development Program (2022YFD2001801-3).

REFERENCES

[1] Al-Dairi, M., Pathare, P. B., Al-Yahyai, R., Jayasuriya, H., & Al-Attabi, Z. (2024). Banana fruit bruise detection using fractal dimension based image processing. *Food chemistry*, *455*, 139812.

[2] Opara, U. L., & Pathare, P. B. (2014). Bruise damage measurement and analysis of fresh horticultural produce—A review. *Postharvest Biology and Technology*, *91*, 9-24.

[3] Patel, K. K., & Pathare, P. B. (2024). Principle and applications of near-infrared imaging for fruit quality assessment—An overview. *International Journal of Food Science and Technology*, *59*(5), 3436-3450.

[4] Li, Y., You, S., Wu, S., Wang, M., Song, J., Lan, W., ... & Pan, L. (2024). Exploring the limit of detection on early implicit bruised 'Korla' fragrant pears using hyperspectral imaging features and spectral variables. *Postharvest Biology and Technology*, *208*, 112668.

[5] Manavalan, R. (2020). Automatic identification of diseases in grains crops through computational approaches: A review. *Computers and Electronics in Agriculture*, *178*, 105802.

[6] Arun, R. A., & Umamaheswari, S. (2023). Effective multi-crop disease detection using pruned complete concatenated deep learning model. *Expert Systems with Applications*, *213*, 118905.

[7] Tian, M., Zhang, J., Yang, Z., Li, M., Li, J., & Zhao, L. (2024). Detection of early bruises on apples using near-infrared camera imaging technology combined with adaptive threshold segmentation algorithm. *Journal of Food Process Engineering*, *47*(1), e14500.

[8] Du, Z., Zeng, X., Li, X., Ding, X., Cao, J., & Jiang, W. (2020). Recent advances in imaging techniques for bruise detection in fruits and vegetables. *Trends in Food Science & Technology*, *99*, 133-141.

[9] Li, X., Liu, Y., Jiang, X., & Wang, G. (2021). Supervised classification of slightly bruised peaches with respect to the time after bruising by using hyperspectral imaging technology. *Infrared Physics & Technology*, *113*, 103557.

[10] Li, C., Li, L., Wu, Y., Lu, M., Yang, Y., & Li, L. (2018). Apple variety identification using near-infrared spectroscopy. *Journal of Spectroscopy*, *2018*(1), 6935197.

[11] Zhu, Q., Guan, J., Huang, M., Lu, R., & Mendoza, F. (2016). Predicting bruise susceptibility of 'Golden Delicious' apples using hyperspectral scattering technique. *Postharvest Biology and Technology*, *114*, 86-94.

[12] Wu, D., Wan, G., Jing, Y., Liu, G., He, J., Li, X., Shihu, Y., Ma, P., & Sun, Y. (2023). Hyperspectral imaging combined with deep learning for discrimination of Lingwu long jujube in terms of the time after bruising. *Microchemical Journal*, *194*, 109238.

[13] Liu, D., Lv, F., Wang, C., Wang, G., Zhang, H., & Guo, J. (2023). Classification of early mechanical damage over time in pears based on hyperspectral imaging and transfer learning. *Journal of Food Science*, *88*(7), 3022-3035.
[14] Mei, M., & Li, J. (2023). An overview on optical non-destructive detection of bruises in fruit: Technology, method, application, challenge and trend. *Computers and Electronics in Agriculture*, *213*, 108195.
[15] Ünal, Z., Kızıldeniz, T., Özden, M., Aktaş, H., & Karagöz, Ö. (2024). Detection of bruises on red apples using deep learning models. *Scientia Horticulturae*, *329*, 113021.
[16] Nandi, C. S., Tudu, B., & Koley, C. (2016). A machine vision technique for grading of harvested mangoes based on maturity and quality. *IEEE sensors Journal*, *16*(16), 6387-6396.
[17] Hu, Z., Tang, J., Zhang, P., & Patlolla, B. P. (2018). Identification of bruised apples using a 3-D multi-order local binary patterns based feature extraction algorithm. *IEEE Access*, *6*, 34846-34862.
[18] Li, X., Cai, C., Zheng, H., & Zhu, H. (2022). Recognizing strawberry appearance quality using different combinations of deep feature and classifiers. *Journal of Food Process Engineering*, *45*(3), e13982.
[19] Attri, I., Awasthi, L. K., Sharma, T. P., & Rathee, P. (2023). A review of deep learning techniques used in agriculture. *Ecological Informatics*, *77*, 102217.
[20] Shafik, W., Tufail, A., De Silva Liyanage, C., & Apong, R. A. A. H. M. (2024). Using transfer learning-based plant disease classification and detection for sustainable agriculture. *BMC Plant Biology*, *24*(1), 136. https://doi.org/10.1186/s12870-024-04825-y
[21] Barbosa, B. D. S., Costa, L., Ampatzidis, Y., Vijayakumar, V., & dos Santos, L. M. (2021). UAV-based coffee yield prediction utilizing feature selection and deep learning. *Smart Agricultural Technology*, *1*, 100010.
[22] Chen, M., Cui, Y., Wang, X., Xie, H., Liu, F., Luo, T., Zheng, S., Yufeng Luo, Y. (2021). A reinforcement learning approach to irrigation decision-making for rice using weather forecasts. *Agricultural Water Management*, 250: 106838.
[23] Zhong, L., Guo, X., Xu, Z., & Ding, M. (2021). Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma*, *402*, 115366.
[24] Wu, S. L., Tung, H. Y., & Hsu, Y. L. (2020, December). Deep learning for automatic quality grading of mangoes: methods and insights. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. pp.446-453. IEEE.
[25] Yuan, Y., Yang, Z., Liu, H., Wang, H., Li, J., & Zhao, L. (2022). Detection of early bruise in apple using near-infrared camera imaging technology combined with deep learning. *Infrared Physics & Technology*, *127*, 104442.
[26] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. pp. 801-818.
[27] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4510-4520.
[28] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp.13713-13722.
[29] Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*. 405-420.
[30] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. pp. 234-241. Springer international publishing.
[31] Badrinarayanan V., Kendall A., Cipolla R. (2016) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 卷 2017 卷 39 (12)*, 2481-2495.
[32] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 2881-2890.
[33] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* pp. 5693-5703.